Olle Selinus
Brian Alloway
José A. Centeno
Robert B. Finkelman
Ron Fuge
Ulf Lindh
Pauline Smedley
Editors

Essentials of Medical Geology

Revised Edition

INTERNATIONAL MEDICAL GEOLOGY ASSOCIATION
IMGA

Springer

Essentials of Medical Geology

Olle Selinus
Editor

Brian Alloway • José A. Centeno
Robert B. Finkelman • Ron Fuge • Ulf Lindh
Pauline Smedley
Associate Editors

Essentials of Medical Geology

Revised Edition

Springer

First edition has been previously published by Elsevier in 2005 (ISBN 0-12-636341-2)

ISBN 978-94-007-4374-8 ISBN 978-94-007-4375-5 (eBook)
DOI 10.1007/978-94-007-4375-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012952089

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

It is our pleasure to introduce an updated version of *Essentials to Medical Geology*. This award-winning book was first published in 2005 and helped stimulate the rapid development of medical geology. The book has been widely used as a textbook all over the world at universities, in short courses and as a resource for researchers. It has become the primary reference book in medical geology.

Essentials of Medical Geology has won three prestigious rewards. It was recognized as a "Highly Commended" title in the Public Health category by the British Medical Association. This is a very prestigious acknowledgment of the value of the book. The book was one of the best of all published books in Public Health in 2005. The British Medical Association bestow awards on publications "which are deemed to best fulfill the criteria of clinical accuracy and currency and which maintain a high standard of design and production".

It was also one of two winners in the "Geology/Geography" category of the 2005 Awards for Excellence in Professional and Scholarly Publishing of the Professional Scholar Publishing (PSP) Division of the Association of American Publishers. The PSP awards recognize both editorial standards and the design and production standards.

Essentials of Medical Geology also won a prestigious award in January 2006. Every year Choice (*Choice: Current Reviews for Academic Libraries* is the premier source for reviews of academic books, electronic media, and Internet resources of interest to those in higher education.) publishes a list of Outstanding Academic Titles that were reviewed during the previous calendar year. This prestigious list reflects the best in scholarly titles reviewed by Choice and brings with it extraordinary recognition from the very selective academic library community. Choice editors base their selections on the reviewer's evaluation of the work, the editor's knowledge of the field, and the reviewer's record. The list was known as Outstanding Academic Books until 2000. In awarding Outstanding Academic Titles, the editors apply several criteria to reviewed titles: overall excellence in presentation and scholarship; importance relative to other literature in the field; distinction as a first treatment of a given subject in book or electronic form, originality or uniqueness of treatment; value to undergraduate students, and; importance in building undergraduate library collections.

The book has thus been recognized in both communities for which it was intended, first by the British Medical Association, and then as a Geology resource!

In addition to its wide circulation in the English language *Essentials of Medical Geology* was also translated into Chinese in 2009.

Since 2005, many papers and book chapters on medical geology issues have been published. But also a book in medical geology was published by Springer: Selinus, O, Finkelman, B, Centeno, J., 2009, Editors. *Medical Geology – A Regional Synthesis*. Springer. 2010. This is a complement to *Essentials of Medical Geology*, presenting the medical geology issues observed in different countries and different regions all over the world.

Some of the chapters in the present book have not needed any updating, some are updated moderately and some have been updated considerably. We have also added two new chapters filling gaps, on public health, and agriculture and health. Those chapters are Cook, A. *Public Health and Geological Processes: An overview of a fundamental relationship* and Knez, M

and Graham, R. *The impact of micronutrient deficiencies in agricultural soils and crops on the nutritional health of humans*.

We are very pleased that Springer is prepared to print this new edition of *Essentials of Medical Geology*. This updated volume can now continue to be used as a textbook and reference book for all who are interested in this important topic and its impacts on the health and wellbeing of many millions of people all over the world.

Olle Selinus
Brian Alloway
José A. Centeno
Robert B. Finkelman
Ron Fuge
Ulf Lindh
Pauline Smedley

Preface

All substances are poisons; there is none which is not a poison. The right dose differentiates a poison and a remedy.

Paracelsus (1493–1541)

Medical geology is the science dealing with the relationship between natural geological factors and health in humans and animals and with understanding the influence of ordinary environmental factors on the geographical distribution of such health problems. It is a broad and complicated subject that requires interdisciplinary contributions from various scientific fields if its problems are to be understood, mitigated, or resolved. Medical geology, which focuses on the impacts of geologic materials and processes (i.e., the natural environment) on animal and human health, can be considered as complementary to environmental medicine. The field of medical geology brings together geoscientists and medical and public health researchers to address health problems caused or exacerbated by geologic materials such as rocks, minerals, and water and geologic processes such as volcanic eruptions, earthquakes, and dust.

Paracelsus defined a basic law of toxicology: Any increase in the amount or concentration of elements causes increasing negative biological effects, which may lead to inhibition of biological functions and, eventually, to death. However, despite the harmful effects of some elements, others are essential for life. Therefore, deleterious biological effects can result from either increasing or decreasing concentrations of various trace elements. Thus, as with many aspects of life, either too much or too little can be equally harmful. All of the elements that affect health are found in nature and form the basis for our existence as living creatures. The periodic table of elements, as an indicator of the roles played by the elements in the biosphere, is the basis for our understanding (Figure 1).

The writings of Hippocrates, a Greek physician of the Classical Period, demonstrate how far back our basic knowledge extends:

Whoever wishes to investigate medicine properly, should proceed thus....We must also consider the qualities of the waters, for as they differ from one another in taste and weight, so also do they differ much in their quality.

Hippocrates (460–377 BC)

Hippocrates held the belief that health and "place" are related to ancient origin. Knowledge of specific animal diseases also originated long ago. Even in Chinese medical texts of the third century BC, cause-and-effect relationships are found. Unfortunately, most such observations were lost because they were never written down. As the science grew, many previously unknown relationships began to be understood and a new scientific field evolved: medical geology. This book covers the essentials of our knowledge in this area.

Geology and Health

Geology may appear far removed from human health. However, rocks and minerals comprise the fundamental building blocks of the planet and contain the majority of naturally occurring

$_{1}$H																$_{2}$He
$_{3}$Li	$_{4}$Be											$_{5}$B	$_{6}$C	$_{7}$N	$_{8}$O	$_{9}$F	$_{10}$Ne
$_{11}$Na	$_{12}$Mg											$_{13}$Al	$_{14}$Si	$_{15}$P	$_{16}$S	$_{17}$Cl	$_{18}$Ar
$_{19}$K	$_{20}$Ca	$_{21}$Sc	$_{22}$Ti	$_{23}$V	$_{24}$Cr	$_{25}$Mn	$_{26}$Fe	$_{27}$Co	$_{28}$Ni	$_{29}$Cu	$_{30}$Zn	$_{32}$Ga	$_{32}$Ge	$_{33}$As	$_{34}$Se	$_{36}$Br	$_{36}$Kr
$_{37}$Rb	$_{38}$Sr	$_{39}$Y	$_{40}$Zr	$_{41}$Nb	$_{42}$Mo	$_{43}$Tc	$_{44}$Ru	$_{45}$Rh	$_{46}$Pd	$_{47}$Ag	$_{48}$Cd	$_{49}$In	$_{50}$Sn	$_{51}$Sb	$_{52}$Te	$_{53}$I	$_{54}$Xe
$_{55}$Cs	$_{56}$Ba	57-71	$_{72}$Hf	$_{73}$Ta	$_{74}$W	$_{75}$Re	$_{76}$Os	$_{77}$Ir	$_{78}$Pt	$_{79}$Au	$_{80}$Hg	$_{81}$Tl	$_{82}$Pb	$_{83}$Bi	$_{84}$Po	$_{85}$At	$_{86}$Rn
$_{87}$Fr	$_{88}$Ra	89-103	$_{104}$Db	$_{105}$Jo	$_{106}$Rf	$_{107}$Bh	$_{108}$Hn	$_{109}$Mt	110	111							
		$_{57}$La	$_{58}$Ce	$_{59}$Pr	$_{60}$Nd	$_{61}$Pm	$_{62}$Sm	$_{63}$Eu	$_{64}$Gd	$_{65}$Tb	$_{66}$Dy	$_{67}$Ho	$_{68}$Er	$_{69}$Tm	$_{70}$Yb	$_{71}$Lu	
		$_{89}$Ac	$_{90}$Th	$_{91}$Pa	$_{92}$U	$_{93}$Np	$_{94}$Pu	$_{95}$Am	$_{96}$Cm	$_{97}$Bk	$_{98}$Cf	$_{99}$Es	$_{100}$Fm	$_{101}$Md	$_{102}$No	$_{108}$Lr	

Figure 1 Periodic table illustrating major elements (*pink*), minor elements (*blue*), trace elements (*yellow*), and noble gases (*gray*) in the biosphere. Those in green are essential trace elements. Known established toxic elements are shown in *red*

chemical elements. Many elements are essential to plant, animal, and human health in small doses. Most of these elements are taken into the human body via food, water, and air. Rocks, through weathering processes, break down to form the soils on which crops and animals are raised. Drinking water travels through rocks and soils as part of the hydrological cycle and much of the dust and some of the gases contained in the atmosphere are of geological origin. Hence, through the food chain and through the inhalation of atmospheric dusts and gases, human health is directly linked to geology.

The volcanic eruption of Mount Pinatubo is a splendid example of the dramatic effects of geology. Volcanism and related activities are the principal processes that bring elements to the surface from deep within the Earth. During just two days in June 1991, Pinatubo ejected 10 billion metric tonnes of magma and 20 million tonnes of SO_2; the resulting aerosols influenced global climate for three years. This single event introduced an estimated 800,000 tonnes of zinc, 600,000 tonnes of copper, 550,000 tonnes of chromium, 100,000 tonnes of lead, 1000 tonnes of cadmium, 10,000 tonnes of arsenic, 800 tonnes of mercury, and 30,000 tonnes of nickel to the surface environment. (Garrett, R.G., 2000). Volcanic eruptions redistribute many harmful elements such as arsenic, beryllium, cadmium, mercury, lead, radon, and uranium. Many other redistributed elements have undetermined biological effects. At any given time, on average, 60 volcanoes are erupting on the land surface of the Earth, releasing metals into the environment. Submarine volcanism is even more significant than that at continental margins, and it has been conservatively estimated that at least 3000 vent fields are currently active along the mid-ocean ridges.

Goal and Approach

Because of the importance of geological factors on health and the widespread ignorance of the importance of geology in such relationships, in 1996 the International Union of Geological Sciences (IUGS) commission COGEOENVIRONMENT (Commission on Geological Sciences for Environmental Planning) established an International Working Group on Medical Geology with the primary goal of increasing awareness of this issue among scientists, medical specialists,

and the general public. In 2000 the United Nations Educational, Scientific, and Cultural Organization (UNESCO) became involved through a new International Geological Correlation Programme (IGCP) project 454 Medical Geology. Project 454 brings together, on a global scale, scientists working in this field in developing countries with their colleagues in other parts of the world and stresses the importance of geoscientific factors that affect the health of humans and animals. In 2002 the International Council for Science (ICSU) made it possible to put together international short courses on this subject, a cooperation involving the Geological Survey of Sweden, US Geological Survey, and the US Armed Forces Institute of Pathology in Washington, DC. The aim of these short courses, which are offered all over the world, is to share the most recent information on how metal ions and trace elements impact environmental and public health issues. The scientific topics of the courses include environmental toxicology; environmental pathology; geochemistry; geoenvironmental epidemiology; the extent, patterns, and consequences of exposures to metal ions; and analysis. Areas of interest include metal ions in the general environment, biological risk-assessment studies, modern trends in trace-element analysis, and updates on the geology, toxicology, and pathology of metal ion exposures.

Because of this increasing activity and interest in this field, we decided to write a book that could be used both as a reference and as a general textbook. Our goal is to emphasize the importance of geology in health and disease in humans and animals. The audience of the book consists of upper division undergraduates, graduate students, environmental geoscientists, epidemiologists, medics, and decision makers, but, we have also strived to make the book interesting and understandable to environmentally conscious members of the general public.

There are important relationships between our natural environment and human health. Our approach is to integrate these two fields to enable better understanding of these often complex relationships. All chapters have numerous cross-references not only among the other chapters but also to related reading.

Sectional Plan

Chapter 1 gives a brief history of medical geology. It is not intended to be an exhaustive overview; instead our overview highlights some important cases in the development of the science of medical geology.

The subsequent material is presented in four sections, each describing different aspects of the subject.

The first section covers environmental biology. Environmental biology may be characterized by interactions between geological and anthropogenic sources and the kingdoms of life. The geological sources provide life with essential major, minor, and trace elements. In addition, geology provides access to nonessential elements. To influence life, both beneficially and adversely, elements have to be in the environment as well as, in most cases, bioavailable. Therefore this section gives an introduction to the different aspects of environmental biology and provides a foundation for the following sections.

The second section, on pathways and exposures, covers many of the myriad different aspects of medical geology. It has long been said that "we are what we eat"; however, in terms of medical geology we are in fact what we eat, drink, and breathe. The major pathways into the human body of all elements and compounds, whether beneficial or harmful, derive from the food and drink we consume and the air we breathe. The twelve chapters of this section concentrate on the interrelationships among our natural environment, geology, and health. Numerous examples from all over the world are presented on topics ranging from element toxicities and deficiencies, to geophagia, to global airborne dust and give a clear view of the vast importance of the natural environment on our health. After reading these chapters, you should have no doubt that geology is one of the most important, although often neglected, factors in our well-being.

The third section, on environmental toxicology, pathology, and medical geology, covers the medical aspects of medical geology. In recent decades there has been an increasing awareness of the importance of the interaction of mammalian systems with their natural environment. The primary focus has been on understanding exposure to hazardous agents in the natural environment through air, water, and soil. Such appreciation has led to myriad investigations focused on identifying those natural (and sometimes anthropogenic) environmental risk factors that may be involved in the development of human and other animal diseases. These five chapters describe the different effects of elements in our bodies, how geology affects us, and how we can recognize these effects.

The fourth section, on techniques and tools, brings together in a very practical way our knowledge of the different relevant disciplines. Geoscientists and medical researchers bring to medical geology an arsenal of valuable techniques and tools that can be applied to health problems caused by geologic materials and processes. Although some of these tools may be common to both disciplines, practitioners of these disciplines commonly apply them in novel ways or with unique perspectives. In this part we look at some of these tools and techniques.

Finally, we have included three appendices. Appendix A covers international and some national reference values for water and soils. Appendix B lists Web links from Chapter 20. Appendix C is a large glossary to be used whenever you need a term explained. We have tried to make this glossary as comprehensive as possible but there will of course be some shortcomings. The glossary can also be found and downloaded from the Internet (www.medicalgeology.org); go to Medical Geology.

Reference

Garrett, R.G., 2000. Natural sources of Metals in the Environment. *Human and Ecological Risk Assessment*, Vol. 6, No. 6, pp 954–963.

Acknowledgments

A volume like this does not come into being without the efforts of a great number of dedicated people. We express our appreciation to the 60 authors who wrote chapters. In addition to writing chapters, the authors have carried out revisions. To ensure the quality and accuracy of each contribution, at least two independent reviewers scrutinized each chapter from a scientific point of view. However, we have gone even one step further. An interdisciplinary book like this must be written at a level that makes it accessible to workers in many different professions and also to members of the general public interested in environmental sciences. Therefore each chapter has also been read by additional reviewers. The geoscientific chapters have been read by those from the medical profession and the medical chapters have been read by geoscientists. We wish to thank all these people for making this book possible.

We are indebted to the following reviewers: Gustav Åkerblom, James Albright, Neil Ampel, Arthur C. Aufderheide, Stefan Claesson, Rick Cothern, Gerald Feder, Peter Frisk, Arne Fröslie, Mark Gettings, Larry Gough, Steve Guptill, Michalan Harthill, Anders Hemmingsson, Brenda Houser, John Hughes, John M. Hunter, Nils Gunnar Ihlbäck, G. V. Iyengar, Erland Johansson, Karen Johnson, Chris Johnson, Andre Kajdacsy-Balla, David G. Kinniburgh, Allan Kolker, Orville Levander, Jan Luthman, Bo L. Lönnerdal, Jörg Matschullat, Chester G. Moore, Maria Nikkarinen, Björn Öhlander, James Oldfield, Rolf Tore Ottesen, Joseph P. Pestaner, Geoffrey Plumlee, Clemens Reimann, Reijo Salminen, Ulrich Siewers, Håkan Sjöström, David Smith, Barry Smith, Alex Stewart, David Templeton, and Paul Younger.

In addition to these, all editors and several of the authors have also acted as reviewers for different chapters.

We also thank the dedicated team of Academic Press—our executive editor Frank Cynar, project manager Jennifer Helé, and production editor Daniel Stone—for their outstanding work and for encouraging us and helping us when needed.

Lastly, we also want to remember Prof. Valentin K. Lukashev, of Minsk, Belarus, our good colleague and friend who attended the first planning meeting of this book. He died among us of a heart attack in Uppsala, Sweden, on June 8, 1998, shortly after having given a presentation on medical geology in Belarus and the former Soviet Union and after having contributed valuable suggestions for the contents of this book at the first planning discussions. The logotype of medical geology will always serve as a fond remembrance of him, since he had suggested it just before he passed away.

Olle Selinus
Brian Alloway
José A. Centeno
Robert B. Finkelman
Ron Fuge
Ulf Lindh
Pauline Smedley

Contents

Contributors

Peter W. Abrahams Aberystwyth University, Aberystwyth, UK

Brian J. Alloway Soil Research Centre, Department of Geography and Environmental Science, University of Reading, Whiteknights Reading, RG6 6DW, UK

J. Donald Appleton British Geological Survey, Nottingham, UK

Charlotte Bowman University of Victoria, Victoria, BC, Canada

Mark W. Bultman Geology, Minerals, Energy and Geophysics Science Center, U.S. Geological Survey, Tucson, AZ, USA

Joseph E. Bunnell United States Department of the Interior, United States Geological Survey, Reston, VA, USA

Allen P. Burke Joint Pathology Center, Silver Spring, MD, USA

Sergio Caroli Community Reference Laboratory, Instituto Superiore di Sanita, Rome, Italy

José A. Centeno Joint Pathology Center, Biophysical Toxicology, Malcolm Grow Medical Clinic, Joint Base Andrews Air Naval Facility, Washington, DC, USA

M. George Cherian Department of Pathology, University of Western Ontario, London, ON, Canada

Gerald F. Combs Jr. Grand Forks Human Nutrition Center, USDA Agricultural Research Service, Grand Forks, ND, USA

Angus Cook School of Population Health, University of Western Australia, Crawley, Australia

Brian E. Davies 3 Old Hall Barns, Markfield Road, Groby, Leicestershire LE6 0FL, England

Theo C. Davies University of Venda, Thohoyandou, Limpopo, South Africa

Edward Derbyshire Royal Holloway, University of London, Egham, Surrey, UK

W. Mike Edmunds School of Geography and the Environment, Oxford University, Oxford, OX1 3QY UK

Robert B. Finkelman University of Texas at Dallas, Richardson, TX, USA

Frederick S. Fisher Department of Geosciences, University of Arizona, Tucson, AZ, USA

Fiona M. Fordyce British Geological Survey, Edinburgh, UK

Jefferson Fowles Tox-Logic Consulting, LLC Petaluma, CA, USA

Teri J. Franks Joint Pathology Center, Silver Spring, MD, USA

Ron Fuge Aberystwyth University, Aberystwyth, UK

Robert G. Garrett Geological Survey of Canada, Natural Resources Canada, Ottawa, ON, Canada

Pierre D. Glynn United States Department of the Interior, United States Geological Survey, Reston, VA, USA

Robin D. Graham School of Biological Sciences, Flinders University of South Australia, Adelaide, Australia

Tee L. Guidotti Medical Advisory Services, Rockville, MD, USA

Stephen C. Guptill Geographic Information Systems Spatial Epidemiology, 3431 Lyrac Street, Oakton, USA

Kevin M. Hiscock School of Environmental Sciences, University of East Anglia, Norwich, UK

Claire J. Horwell Institute of Hazard, Risk and Resilience, Department of Earth Sciences, Durham University, Durham, UK

Kamal G. Ishak (deceased) Department of Hepatic Pathology, Armed Forces Institute of Pathology, Washington, DC, USA

Tina Kold Jensen Research Unit of Environmental Medicine, Institute of Public Health, University of Southern Denmark, Odense, Denmark

Wayne B. Jonas East Coast U.S. Office, Samueli Institute for Information Biology, Alexandria, VA, USA

Bernt Jones Department of Clinical Science, Faculty of Veterinary Medicine and Animal Science, Swedish, University of Agricultural Sciences, Uppsala, Sweden

Alexander W. Karlsen United States Department of the Interior, United States Geological Survey, Reston, VA, USA

David G. Kinniburgh British Geological Survey, Oxfordshire, UK

Marija Knez School of Biological Sciences, Flinders University of South Australia, Adelaide, Australia

Leonard F. Konikow United States Department of the Interior, United States Geological Survey, Reston, VA, USA

Michael N. Koss Medical Laboratories, University of Southern California Medical Center, Los Angeles, CA, USA

Gerald Lalor International Centre for Environmental and Nuclear Sciences, University of the West Indies, Kingston, Jamaica

Ulf Lindh Biology Education Centre, Research in Metal Biology, Uppsala University, Uppsala, Sweden

Stephen Macko Department of Environmental Sciences, University of Virginia, Charlottesville, VA, USA

Bernhard Michalke Helmholtz Center Munich, German Research Center for Environmental Health, Munich, Germany

Chester G. Moore Department of Microbiology, Immunology and Pathology, School of Veterinary Medicine and Biomedical Sciences, Colorado State University, Fort Collins, CO, USA

Florabel G. Mullick Joint Pathology Center, Silver Spring, MD, USA

Jesper B. Nielsen Institute of Public Health, University of Southern Denmark, Odense, Denmark

Monica Nordberg Institute Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Demosthenes Pappagianis School of Medicine, Medical Microbiology & Immunology, University of California-Davis, Davis, USA

Daniel P. Perl Uniformed Services University of the Health Sciences, Bethesda, MD, USA

Joseph P. Pestaner Medical Examiner Office, Washington, DC, USA

Eva Rubenowitz-Lundin Department of Environmental Medicine, Göteborg University, Göteborg, Sweden

Olle Selinus Linneaus University, Kalmar, Sweden

Timothy M. Shields Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD, USA

H. Catherine W. Skinner Yale University, New Haven, CT, USA

Pauline L. Smedley British Geological Survey, Oxfordshire, UK

Paul B. Tchounwou College of Science, Engineering and Technology, Jackson State University, Jackson, MS, USA

Todor Todorov Crustal Geophysics and Geochemistry Science Center, United States Geological Survey, Denver, CO, USA

Chian-Hsiao Tseng Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

Mitko Vutchkov International Centre for Environmental and Nuclear Sciences, University of the West Indies, Kingston, Jamaica

Philip Weinstein University of South Australia, Adelaide, Australia

Robert J.P. Williams Inorganic Chemistry Laboratory, University of Oxford, Oxford, UK

Medical Geology: Perspectives and Prospects

1

Brian E. Davies, Charlotte Bowman, Theo C. Davies, and Olle Selinus

Contents

This chapter is a brief history of medical geology—the study of health problems related to "place." This overview is not exhaustive; instead, it highlights some important cases that have arisen during the development of the science of medical geology. An excess, deficiency or imbalance of inorganic elements originating from geological sources can affect human and animal well-being either directly (e.g., a lack of dietary iodine leading to goiter) or indirectly (e.g., effect on metabolic processes such as the supposed protective effect of selenium in cardiovascular disease). Such links have long been known but were unexplained until alchemy evolved into chemistry in the seventeenth century, when medicine ceased to be the art of monks versed in homeopathic remedies and experimental explanations of disease was sought rather than relying on the writings of the Classical Greek philosophers, and modern geology was forged by Lyell and Hutton. In addition, the exploitation of mineral resources gathered pace in the seventeenth century and brought in its train the widespread release of toxic elements to the environment. New sciences of public health and industrial hygiene emerged and their studies have helped inform our understanding of the health implications of the natural occurrence of these elements.

1.1 The Foundations of Medical Geology

1.1.1 Ancient Reports

Many ancient cultures made reference to the relationship between environment and health. Often health problems were linked to occupational environments but close links to the natural environment were also noted. Chinese medical texts dating back to the third century BC contain several references to relationships between environment and health. During both the Song Dynasty (1000 BC) and the Ming Dynasty (Fourteenth to Seventeenth century AD), lung problems related to rock crushing and symptoms of occupational lead poisoning were recognized. Similarly, the Tang

B.E. Davies (✉)
3 Old Hall Barns, Markfield Road, Groby, Leicestershire LE6 0FL, England
e-mail: ewartdavies@gmail.com

C. Bowman
University of Victoria, Victoria, BC, Canada

T.C. Davies
University of Venda, Thohoyandou, Limpopo, South Africa

O. Selinus
Linneaus University, Kalmar, Sweden
e-mail: olle.selinus@gmail.com

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_1,

Dynasty alchemist Chen Shao-Wei stated that lead, silver, copper, antimony, gold and iron were poisonous (see discussion in Nriagu, 1983).

Contemporary archaeologists and medical historians have provided us with evidence that the poor health often revealed by the tissues of prehistoric cadavers and mummies can commonly be linked to detrimental environmental conditions of the time. Goiter, for example, which is the result of severe iodine deficiency, was widely prevalent in ancient China, Greece, and Egypt, as well as in the Inca state of Peru. The fact that this condition was often treated with seaweed, a good source of iodine, indicates that these ancient civilizations had some degree of knowledge with regard to the treatment of dietary deficiencies with natural supplements.

As early as 1,500 years ago, certain relationships between water quality and health were known:

> Whoever wishes to investigate medicine properly, should proceed thus....We must also consider the qualities of the waters, for as they differ from one another in taste and weight, so also do they differ much in their quality (Hippocrates 460–377 BC).

Hippocrates, a Greek physician of the Classical period, recognized that health and place are causally related and that environmental factors affected the distribution of disease (Låg 1990; Foster 2002). Hippocrates noted in his treatise *On Airs, Waters, and Places* (Part 7) that, under certain circumstances, water "comes from soil which produces thermal waters, such as those having iron, copper, silver, gold, sulphur, alum, bitumen, or nitre," and such water is "bad for every purpose." Vitruvius, a Roman architect in the last century BC, noted the potential health dangers related to mining when he observed that water and pollution near mines posed health threats (cited in Nriagu 1983). Later, in the first century AD, the Greek physician Galen reaffirmed the potential danger of mining activities when he noticed that acid mists were often associated with the extraction of copper (cited in Lindberg 1992).

An early description linking geology and health is recounted in the travels of Marco Polo and his Uncle Niccoló. Journeying from Italy to the court of the Great Khan in China in the 1270s they passed to the south and east of the Great Desert of Lop:

> At the end of the ten days he reaches a province called Su-chau....Travelers passing this way do not venture to go among these mountains with any beast except those of the country, because a poisonous herb grows here, which makes beasts that feed on it lose their hoofs; but beasts born in the country recognize this herb and avoid it (Latham 1958).

The animal pathology observed by Marco Polo that resulted from horses eating certain plants was similar to a condition that today we know is caused by the consumption of plants in which selenium has accumulated; this explorer's account may be the earliest report of selenium toxicity. Marco Polo also described goiter in the area around the oasis city of Yarkand (Shache) and ascribed it to a peculiarity of the local water. Earlier, near Kerman on the Iranian eastern frontier, he commented on a lack of bellicosity in the tribesmen that he attributed to the nature of the soil. In what could be considered the first public health experiment, Marco Polo imported soil to place around the tribe's tents in an effort to restore their bellicosity. His approach proved to be effective (see also Chap. 16).

Health problems resulting from the production of metal have been identified in many parts of the world. The common use of metals in ancient societies suggested their toxicity. Although the relationship between lead and a variety of health risks is now well documented in modern society, the relationship was less well known in the past. Lead has been exploited for over six millennia, with significant production beginning about 5,000 years ago, increasing proportionately through the Copper, Bronze, and Iron Ages, and finally peaking about 2,000 years ago (Hong et al. 1994; Nriagu 1998). Several descriptions of lead poisoning found in texts from past civilizations further corroborate the heavy uses of lead. Clay tablets from the middle and late Assyrian periods (1550–600 BC) provide accounts of lead-poisoning symptoms, as do ancient Egyptian medical papyri and Sanskrit texts dating from over 3,000 years ago (Nriagu 1983). About 24% of discovered lead reserves were mined in ancient times (Nriagu 1998).

It has been estimated that during the time of the Roman Empire the annual production of lead approached 80,000 tonnes (Hong et al. 1994; Nriagu 1998), and copper, zinc and mercury were also mined extensively (Nriagu 1998). Lead usage exceeded 550 g per person per year, with the primary applications being plumbing, architecture, and shipbuilding. Lead salts were used to preserve fruits and vegetables, and lead was also added to wine to stop further fermentation and to add color or bouquet (Nriagu 1983). The use of large amounts of lead in the daily life of Roman aristocracy had a significant impact on their health, including epidemics of plumbism, high incidence of sterility and stillbirths and mental incompetence. Physiological profiles of Roman emperors who lived between 50 and 250 BC suggest that the majority of these individuals suffered from lead poisoning (Nriagu 1983)., It has been claimed that a contributing factor to the fall of the Roman Empire, in AD 476, may have been the excessive use of lead in pottery, water pipes and other sources (Hernberg 2000).

Mercury was used during the Roman Empire to ease the pain of teething infants, as well as to aid in the recovery of gold and silver. Such applications were also widely found in Egypt in the twelfth century and in Central and South America in the sixteenth century (Eaton and Robertson 1994; Silver and Rothman 1995). Mercury was used to treat syphilis during the sixteenth century and in the felting process in the 1800s (Fergusson 1990).

Copper was first used in its native form approximately 7,000 years ago, with significant production beginning some

2,000 years later and eventually peaking at a production rate of about 15,000 tonnes annually during the Roman Empire, when it was used for both military and civilian purposes, especially coinage. A significant drop in the production of copper followed the fall of the Roman Empire, and production remained low until about 900 years ago when a dramatic increase in production occurred in China, reaching a maximum of 13,000 tonnes annually and causing a number health problems (Hong et al. 1994).

Arsenic was used for therapeutic purposes by the ancient Greeks, Romans, Arabs, and Peruvians, because small doses were thought to improve the complexion; however, it has also long been used as a poison (Fergusson 1990). In the sixteenth century, Georgius Agricola (Agricola 1556) described the symptoms of "Schneeberger" disease among miners working in the Erzgebirge of Germany to mine silver in association with uranium. That disease has since been identified as lung cancer deriving from metal dust and radon inhalation.

1.1.2 More Recent Reports

The industrial revolution in Europe and North America encouraged people to quit the poverty of subsistence agriculture in the countryside to live in increasingly crowded cities where they found work in factories, chemical plants and foundries; however, such occupations exposed the workers to higher levels of chemical elements and compounds that, as rural dwellers, they would rarely have encountered. Friedrich Engels wrote graphic descriptions of the ill health of the new English proletariat in his politically seminal book, *The Conditions of the Working Class in England*, published in 1845. He described the plight of children forced to work in the potteries of Staffordshire: "By far the most injurious is the work of those who dip. . .into a fluid containing great quantities of lead, and often of arsenic.... The consequence is violent pain, and serious diseases of the stomach and intestines. . .partial paralysis of the hand muscles. . .convulsions" (Engels 1845). Engels further characterized the conditions of workers in mid-nineteenth century industrial England as "want and disease, permanent or temporary."

The sciences of toxicology and industrial medicine arose in response to the health problems caused by unregulated industrialization. These sciences have provided the clinical data that allow us to understand the consequences of excess exposure to elements in the natural environment, whether it be due to simple exposure to particular rocks or the exploitation of mineral resources. The emergence of modern geological sciences coupled with increasingly powerful analytical techniques laid the foundation for determining the nature and occurrence of trace elements in rocks and sediments. Scientific agriculture has focused attention on inorganic element deficiencies in plants and animals. Modern medicine has provided reliable descriptions of diseases and more accurate diagnoses through internationally recognized nomenclatures.

Rural people have always recognized that the health of domesticated animals is influenced by drinking water or diet and, therefore, soil properties. These observations could not be explained until the advent of scientific agriculture in the nineteenth century, when it required only a small step to suggest that humans may also be caught up in similar relationships. Diseases now known to be caused by a lack or excess of elements in soil and plants were given names that reflected where they occurred, such as Derbyshire neck in the iodine-deficient areas of the English Midlands or Bodmin Moor sickness over the granites of southwest England where cobalt deficiency is endemic in sheep unless treated. It is interesting to note that in Japan, before the 1868 Meiji Restoration, meat was rarely eaten so there was no tradition of animal husbandry. Japanese authors have suggested that this lack of animal indicators largely contributed to the failure to recognize the significance of metal pollution until it became catastrophic.

Archaeologists have also noted links between health and environmental factors. Analysis of bone material has provided an excellent tool for studying the diet and nutritional status of prehistoric humans and animals (Krzysztof and Glab 2001). For example, the transition from a hunter–gatherer society to an agriculturally based economy resulted in a major dietary change and an accompanying iron deficiency. Iron in plants is more difficult to absorb than iron from a meat source; hence, it has been proposed that this new reliance on a crop diet may have resulted in iron deficiency and anemia among the general population (Roberts and Manchester 2007).

Skeletal remains found in Kentucky have provided prime examples of the relationship between geology and ancient human health. Native Americans established permanent settlements in the area and began normal crop cultivation practices. As a result of soil micronutrient deficiencies, their maize contained extremely low levels of zinc and manganese. These deficiencies led to a range of diet-related health effects that have been clearly documented through the study of dental and skeletal pathology in human remains (Moynahan 1979).

Several landmark discoveries in medical geology have been made in Norway. For a long time, Norwegian farmers have been aware of the unusually frequent occurrence of osteomalacia among domestic animals in certain districts, and to combat the disease they initiated the practice of adding crushed bones to the feed of the animals. Some farmers suspected that a particular pasture plant caused osteomalacia, and a Norwegian official named Jens Bjelke (1580–1659), who had an interest in botany and a knowledge

of foreign languages, gave the suspected plant the Latin name *Gramen ossifragum* ("the grass that breaks bones"). The name has also been *written Gramen Norwagicum ossifragum.*

One hundred years ago, the geologist J. H. L. Vogt learned of the practice of adding crushed bones to the diets of farm animals and investigated a region where osteomalacia was common. When he found very small amounts of the mineral apatite in the rocks, he drew the logical and correct conclusion that a deficiency of phosphorus was the cause of the osteomalacia. Another Norwegian geologist (Esmark 1823) had previously pointed out that vegetation was extraordinarily sparse over the bedrock which was found by Vogt to be very poor in apatite. Once the cause of the osteomalacia was determined, it became a relatively simple matter to prevent the damage by adding phosphorus fertilizer to the soil (Låg 1990) (see also Chap. 15, this volume).

A significant publication was André Voisin's book, *Soil, Grass and Cancer* (1959), especially in light of today's interest in the dangers of free radicals in cells and the protective effects of antioxidant substances and enzymes. Over 40 years ago, Voisin stressed the protective role of catalase and observed that copper deficiency was accompanied by low cytochrome oxidase activity.

Oddur Eiriksson and Benedikt Pjetursson provided detailed descriptions of the damage to teeth of domestic animals that resulted from the eruption of the Icelandic volcano Hekla in 1693. At that time it was not known that the cause was fluorosis. The relationship between the incidence of fluorine deficiency and dental caries has been carefully studied in Scandinavia since World War II, with attention being particularly centered around the need for fluoridation of water. Analyses of the fluoride content of natural waters from various sources and their relationships to the frequency of caries have been reported from several districts (see also Chap. 10, this volume).

1.2 Geochemical Classification of the Elements

The principles of geochemistry and, hence, medical geology were established at a time when modern analytical techniques were in their infancy and most scientists relied on the very laborious classical chemical approaches. Despite the limitations imposed by a lack of rapid analysis of rocks and soils, the basic principles of geochemistry were known by the start of the twentieth century. In 1908, Frank W. Clarke, of the U.S. Geological Survey, published the original edition of *The Data of Geochemistry*, in which he adopted a systems approach to present his information. Clarke's book was the forerunner of several texts published during the first half of the twentieth century that have helped us understand how geochemistry is linked to health. Arguably the most important text of the period was V. Goldschmidt's *Geochemistry* (1954), which was based on work by Linus Pauling; it was completed by Alex Muir in Scotland and published after Goldschmidt's death in 1947. Two of Goldschmidt's ideas are of special relevance to medical geology: his geochemical classification of the elements and his recognition of the importance of ionic radii in explaining "impurities" in natural crystals.

Table 1.1 Geochemical classification of elements

Group	Elements
Siderophile	Fe, Co, Ni, Pt, Au, Mo, Ge, Sn, C, P
Atmophile	H, N, O
Chalcophile	Cu, Ag, Zn, Cd, Hg, Pb, As, S, Te
Lithophile	Li, Na, K, Rb, Cs, Mg, Ca, Sr, Ba, Al, rare earths (REE)

Goldschmidt's geochemical classification groups elements into four empirical categories (Table 1.1). The *siderophilic* elements are those primarily associated with the iron–nickel (Fe–Ni) core of the Earth; these elements may be found elsewhere to some extent, but this classification explains why, for example, platinum and associated metals are normally rare and dispersed in crustal rocks. This fundamental geochemical observation allowed Alvarez et al. (1980) to recognize the significance of the high iridium contents of clays found at the Cretaceous/Tertiary (K/T) boundary. They proposed the persuasive idea that the impact of an asteroid (Fe–Ni type) on the surface of the Earth could explain the massive species extinctions that define the K/T boundary, including the demise of the dinosaurs. Was this an example of medical geology on a global scale?

The *atmophilic* elements are those dominating the air around us, and *lithophilic* elements are common in crustal silicates (Alvarez et al. 1980). Of special interest are the *chalcophilic* elements, which derive their name from a geochemical grouping of these elements with copper (Greek χαλκός). These elements are encountered locally in high concentrations where recent or ancient reducing conditions (and hydrothermal conditions) have led to the reduction of sulfate to sulfide, resulting in the formation of sulfide minerals such as pyrite (FeS_2) and the ores of lead (galena, PbS) or zinc (sphalerite, ZnS). This same thiophilic tendency underlies the toxicity of lead, mercury, and cadmium because they readily link to the —SH groups of enzymes and thereby deactivate them. Goldschmidt's empirical classification of chalcophilic elements is now reinterpreted in terms of hard and soft acids and bases; soft bases (e.g., R—SH or $R—S^-$) preferentially bind to soft acids (e.g., Cd^{2+} or Hg^{2+}).

Goldschmidt's (and Pauling's) second important concept was the importance of ionic size in explaining both the three-dimensional structures of silicate crystals and how other elements can become incorporated in them. The rules are now generally known as *Goldschmidt's rules of substitution*:

1. The ions of one element can replace another in ionic crystals if their radii differ by less than about 15%.
2. Ions whose charges differ by one unit can substitute provided electrical neutrality of the crystal is maintained.
3. For two competing ions in substitution, the one with the higher ionic potential (charge/radius ratio) is preferred.
4. Substitution is limited when competing ions differ in electronegativity and form bonds of different ionic character.

These rules of substitution and the geochemical classification of elements are fundamental to our growing understanding of medical geology, for they explain many environmental occurrences of toxic elements and allow scientists to predict where such occurrences might be found.

1.3 Contributions to Medical Geology from Public Health and Environmental Medicine

Although most public health problems involve diseases caused by pathogens, inorganic elements and their compounds can also affect public health; among these elements are arsenic, cadmium, and mercury. The effects of mercury on human health can be traced back several centuries. For example, in the sixteenth century and later, mercury and its compounds were widely used to treat syphilis despite its known toxicity (D'itri and D'itri 1977), and mercuric nitrate solution was used to soften fur for hat making. Long-term exposure caused neurological damage in workers handling mercury and gave rise to expressions such as "mad as a hatter" or the "Danbury shakes." In Birmingham, England, buttons were gilded by exposing them to a gold–mercury (Au–Hg) amalgam followed by vaporization of the mercury. By 1891, many tons of mercury had been dissipated around Birmingham, to the great detriment of that city's inhabitants, many of whom suffered from "Gilder's palsy". Neurological damage due to exposure to inorganic mercury compounds was well understood by the end of the nineteenth century. In recent decades there has been concern about environmental levels of mercury in Amazonia, where the amalgamation of gold by mercury in small-scale mining operations has caused widespread mercury pollution

Modern concerns are focussed on methyl mercury, a lipid-soluble organic compound that concentrates up the food chain. Recognition of such a problem resulted from the outbreak of methylmercury poisoning in 1956 in Minamata city in Japan, thus the name used today—Minamata disease (Harada 1995). Subsequently, methylmercury poisoning has been observed in, for example, Niigata (Japan), Sweden, Iraq, and the United States. In the USA mercury emissions from coal burning have led to restrictions on lake and river fishing (EPA 2011).

Concern about environmental cadmium can be traced back to the outbreak of itai itai disease in Japan earlier in the twentieth century (Chaney et al. 1998). The disease resulted in severe bone malformations in elderly women, and a zinc mine in the upper reaches of the Jintsu river was found to be the source of the cadmium that caused the disease. Later, cadmium was found to be linked to kidney damage, and the element was found to build up in soil following the application of some sewage sludges. Many countries now control the land application of sludge and have set limits in terms of permissible cadmium additions (Friberg et al. 1974).

The coloured compounds of arsenic were used as pigments as early as the Bronze Age, and knowledge of its toxicity is just as old. Of concern today are the skin lesions and cancers observed among the millions of people drinking arsenic-rich well water, especially in West Bengal and Bangladesh. As with mercury, links between arsenic and certain cancers were identified early on. Fowler's solution, which contained potassium arsenite, was widely prescribed as a tonic. Patients who believed that if a little of something (a few drops) would do them good then a lot of it must do them a lot of good and tended to overdose on the solution. By the late eighteenth century, it was recognized that injudicious use of Fowler's solution led first to peripheral neuritis, which was followed by skin lesions and cancer (see also Chaps. 12 and 25, this volume).

Coal is a sedimentary rock formed by the diagenesis of buried peats, which, in turn, form from organic debris under wet, reducing conditions. This process favors the precipitation of the sulfides of chalcophilic metals (especially pyrite, FeS_2). Pyrite can contain significant concentrations of arsenic as well as mercury, thallium, selenium, nickel, lead, and cobalt. Incineration of coal releases mercury to the atmosphere; sulfur gases, which cause acid precipitation, and arsenic compounds may also be released or remain in the ash.

In the autumn of 1900, an epidemic of arsenic poisoning occurred among beer drinkers in Manchester, Salford, and Liverpool in England. The poisoning was first traced to the use of sulfuric acid to make the glucose required for the brewing process; apparently, the breweries had unknowingly switched from de-arsenicated acid (sulfuric acid is a valuable by-product of smelting industries, including those dealing with arsenic ores). Additionally, however, malted barley

was dried over coal fires, which contributed to the problem. Even moderate beer drinkers suffered from peripheral neuritis and shingles (herpes zoster), which can be induced by arsenic exposure. Arsenic poisoning has recently emerged again in China, where severe arsenic poisoning has been reported in recent years as a result of consumption of vegetables dried over coal fires (Finkelman et al. 1999).

1.4 Development of Medical Geology

1.4.1 The Knowledge Gained from Single-Element Studies

Over the course of the twentieth century, geoscientists and epidemiologists gained a greater understanding of the many ways in which the environment of Earth can affect the health of its inhabitants. Incidents of metal poisoning and the identification of specific relationships between dietary constituents and health became representative examples of more general human reactions to exposures to the geochemical environment. The clearest example of the relationship between geology and health is when the presence of too much or too little of a single element in the environment is found to cause or influence disease as a result of being transferred into the body through dust in the soil or air or via water or food.

Iodine remains the classic success story in medical geology as far as human health is concerned. The most common health effect associated with an iodine deficiency is goiter, a swelling of the thyroid gland. Late in the nineteenth century, it was determined that iodine concentrates in the thyroid gland, but the iodine concentrations were reduced in the thyroids of patients from endemic goitrous areas. Iodine deficiency disorders (IDDs) remain a major threat to the health and development of populations the world over. Clinical treatment of IDDs is, of course, the prerogative of medical doctors; nonetheless, a greater understanding of the conditions leading to IDDs has resulted from the work of geoscientists. (Iodine is described in detail in Chap. 17.)

The study of arsenic remained the province of toxicology and forensic medicine until the middle twentieth century. A paper on arsenic in well water in the Canadian Province of Ontario stated: "The occurrence of arsenic in well water is sufficiently rare to merit description" (Wyllie 1937). Pictures accompanying the text illustrate keratosis on the feet and the palm of a hand. It was concluded in the article that the occurrence of arsenic poisoning from well water was infrequent. Less than 40 years later, however, the scientific world learned of "blackfoot disease" in the Republic of China (Taiwan), and skin disorders and cancer due to arsenic-polluted well water have been described in Chile, Mexico, and Argentina. Serious problems are currently being reported in West Bengal and Bangladesh. In all cases, the geological link is clear (described in detail in Chaps. 12 and 25).

Cobalt deficiency provides a good example of the relationship between animal health and the geological environment. In New Zealand, cobalt deficiency was known as "bush sickness" or Morton Mains disease; in Kenya, as *nakuruitis*; in England, as pining; in Denmark, as *vosk* or *voskhed*; and in Germany, as *hinsch*. The underlying cause was discovered by Dr. Eric Underwood, an early expert in the medical geology field (Underwood and Filmer 1935). His discovery in 1935 of the essentiality of cobalt is an example of triumph over analytical difficulty. Underwood and Filmer showed that "enzootic marasmus" could be cured by treatment with an acid extract of the iron oxide limonite, from which all but negligible quantities of iron had been removed using the laborious methods of classical qualitative analysis. In all cases, the problem can be traced back to a low cobalt content of the soil parent material. Inadequate cobalt is passed up the food chain for microflora in the gut of herbivores to use in the synthesis of the essential cobalt-containing cobalamin or vitamin B_{12}. Only one case of human cobalt deficiency appears to have been published (Shuttleworth et al. 1961). A 16 month-old girl on an isolated Welsh hill farm was a persistent dirt eater and suffered from anemia and behavioral problems. The cattle on the farm were being treated for cobalt deficiency, and the child recovered her health after oral administration of cobaltous chloride.

Lead poisoning has dominated the environmental agenda for several decades. It is interesting to note that geologists were aware of the potential health problems associated with lead when medical opinion on the subject was still mixed. In mid-nineteenth century Britain, residents expressed growing concern about the unregulated disposal of mine and industrial wastes in rivers. In west Wales, farmers complained that lead mining was ruining their fields as a result of the deposition of polluted sediment when rivers flooded. A Royal Commission in 1874 evaluated their complaints, and legislation soon followed (River Pollution Prevention Act, 1878); however, it was too late. Well into the twentieth century, cattle poisoning in the Ystwyth valley of west Wales continued to occur due to the earlier contamination by mines in the previous century. As late as 1938, the recovery of these rivers was monitored, and even in the 1970s evidence of past pollution was still evident (Davies and Lewin 1974). It was the late Professor Harry Warren in Vancouver, Canada, who first recognized the important implications of high levels of environmental lead. He devoted the last 30 years of his professional life to arguing

for the significance of lead uptake by garden vegetables and its possible role in the etiology of multiple sclerosis. Warren had pioneered the use of tree twigs in prospecting for mineral ores in British Columbia, Canada, and he was surprised to observe that lead contents were often higher in forests bordering roads and concluded that "industrial salting" was a widespread and serious problem. Nonetheless, until the 1960s, environmental lead remained a mere curiosity. Health problems were thought to occur only from industrial exposure or due to domestic poisoning from lead dissolved by soft water from lead pipes.

Over the past 20 years, the removal of lead from gasoline, food cans, and paint has reduced US population blood lead levels by over 80%. Milestones along the way included evidence that dust on hands and direct soil consumption (pica) by children represented a major pathway of lead exposure (Gallacher et al. 1984). The phasing out of lead in gasoline in the United States was accompanied by a general reduction in blood lead levels (Mahaffey et al. 1982). Adding to the debate was the contention that even relatively low levels of lead exposure could harm the development of a child's brain (Davies and Thornton 1989; Nriagu 1983; Ratcliffe 1981; Warren and Delavault 1971).

The medical geology of selenium provides a good example of the interaction between geology and medicine. In the late 1960s, selenium was shown to be essential for animals and to be an integral part of glutathione oxidase, an enzyme that catalyzes the breakdown of hydrogen peroxide in cells (Prasad 1978). In sheep and cattle, a deficiency in selenium accounted for "white muscle disease" (especially degeneration of the heart muscle), and glutathione peroxidase activity was found to be a good measure of selenium status. The problem was particularly widespread among farm animals in Great Britain (Anderson et al. 1979). Humans have also been shown to suffer from selenium deficiency, and in China this condition is referred to as Keshan disease (Rosenfeld and Beath 1964; Frankenberger and Benson 1994; Frankenberger and Engberg 1998). The disease has occurred in those areas of China where dietary intakes of selenium are less than 0.03 mg/day because the selenium content of the soils is low. The condition is characterized by heart enlargement and congestive heart failure. The disease has been primarily seen in rural areas and predominantly among peasants and their families. Those most susceptible have been children from 2 to 15 years of age and women of child-bearing age (Chen et al. 1980; Jianan 1985). Also, it has been suggested that adequate selenium intake may be protective for cancers (Diplock 1984), and self-medication with selenium supplements has become widespread with the belief that a lack of selenium is a risk factor in heart diseases. (Selenium is described in greater detail in Chap. 16.)

1.4.2 The Importance of Element Interactions is Recognized

The number of productive single-element studies has obscured two fundamental geochemical principles: First, from a geochemistry perspective, elements tend to group together, and, second, the study of physiology recognizes that elements can be synergistic or antagonistic. Cadmium is a good example of both principles. In some environments, soil cadmium levels are high because of rock type (such as black shales) or from mining contamination. A highly publicized polluted environment is that of the village of Shipham, which in the eighteenth century was a thriving zinc mining village in the west of England.

A study in 1979 suggested that 22 out of 31 residents showed signs of ill health that could be traced to cadmium (Carruthers and Smith 1979). As a result, the health of over 500 residents was subsequently assessed and compared with that of a matching control population from a nearby non-mining village, but "there was no evidence of adverse health effects in the members of the population studied in Shipham" (Morgan and Simms 1988, Thornton et al. 1980). Chaney et al. (1998) have commented on the disparity between the reports of ill health in Japan and no-effect observations from other parts of the world: "research has shown that Cd transfer in the subsistence-rice food-chain is unique, and that other food-chains do not comprise such high risk per unit soil Cd" and "Evidence indicates that combined Fe and Zn deficiencies can increase Cd retention by 15-fold compared to Fe and Zn adequate diets. . .it is now understood that rice grain is seriously deficient in Fe, Zn, and Ca for human needs".

Copper and molybdenum taken individually and together demonstrate the importance of not relying upon simple single-cause relationships. In Somerset (England) there is an area in which pasture causes scouring in cattle. The land is known locally as "teart" and was first reported in the scientific literature in 1862 (Gimingham 1914), but the cause of the disorder (molybdenum) was not ascertained until 1943 (Ferguson et al. 1943), when it was shown that the grass contained 20–200 mg molybdenum per kg (d.m.) and that the disorder could be cured by adding cupric sulfate to the feed. The origin of the excess molybdenum was the local black shales (Lower Lias) (Lewis 1943). Over 20 years later, geochemical reconnaissance of the Lower Lias throughout the British Isles showed that elevated molybdenum contents in soils and herbage were a widespread problem over black shale, regardless of geological age, and that this excess molybdenum was the cause of bovine hypocuprosis (Thornton et al. 1966, 1969; Thomson et al. 1972). A moose disease in Sweden provides another example of the effects of molybdenum, in this case resulting from the interaction of molybdenum with copper.

This disease is covered in detail in Chap. 21 (see also Kabata-Pendias and Pendias 1992; Kabata-Pendias 2001; Adriano 2001).

1.4.3 Mapping Diseases as a Tool in Medical Geology

Medical geology benefits from the work of medical geographers who have mapped diseases in different countries. For some important groups of diseases (e.g., cancers, diseases of the central nervous system, and cardiovascular disease), the causes are by and large uncertain. When the incidence or prevalence of these diseases has been mapped, especially in countries of western Europe, significant differences from place to place have been reported that are not easily explained by genetic traits or social or dietary differences. Howe (1963) pioneered the use of standardized epidemiological data in his 'National Atlas of Disease Mortality in the United Kingdom'. His stomach cancer maps clearly identified very high rates in Wales. Environmental influences appear to be involved in the etiologies, and a role for geology has been suggested by many authors (see, for example, Chap. 14). An early study of gastrointestinal cancer in north Montgomeryshire, Wales (Millar 1961) seemed to show an association with environmental radioactivity because local black shales were rich in uranium. There was no direct evidence to support the hypothesis, and the study was marked by a problem of earlier work—namely, an indiscriminate use of statistics. Work in 1960 in the Tamar valley of the west of England appeared to show that mortality from cancer was unusually low in certain villages and unusually high in others (Davies 1971). Within the village of Horrabridge, mortality was linked to the origin of different water supplies: The lowest mortality was associated with reservoir water from Dartmoor, whereas the highest mortality was associated with well or spring water derived from mineralized rock strata. Although this study was again statistically suspect, it stimulated a resurgence of interest in the link between cancer and the environment.

Stocks and Davies (1964) sought direct associations between garden soil composition and the frequency of stomach cancer in north Wales, Cheshire, and two localities in Devon. Soil organic matter, zinc, and cobalt were related positively with stomach cancer incidence but not with other intestinal cancer. Chromium was connected with the incidence of both. The average logarithm of the ratio of zinc/copper in garden soils was always higher where a person had just died of stomach cancer after 10 or more years of residence than it was at houses where a person had died similarly of a nonmalignant cause. The effect was more pronounced and consistent in soils taken from vegetable gardens, and it was not found where the duration of residence was less than 10 years.

Association is not necessarily evidence for cause and effect. For mapping approaches to be reliable, two conditions must be satisfied. First, it is essential to be able to show a clear pathway from source (e.g., soil) to exposure (e.g., dirt on hands) to assimilation (e.g., gastric absorption) to a target organ or physiological mechanism (e.g., enzyme system). The second condition, rarely satisfied, is that the hypothetical association must be predictive: If the association is positive in one area, then it should also be positive in a geologically similar area; if not, why not? This condition is well illustrated by fluoride and dental caries—environments where fluoride is naturally higher in drinking water have consistently proved to have lower caries rates.

A possible link between the quality of water supply, especially its hardness, was the focus of much research in the 1970s and 1980s. This was noticed, for example, in Japan in 1957. A statistical relationship was found between deaths from cerebral hemorrhage and the sulfate/carbonate ratio in river water which, in turn, reflected the geochemical nature of the catchment area. In Britain, calcium in water was found to correlate inversely with cardiovascular disease, but the presence of magnesium did not; thus, hard water may exercise some protective effect. Attention has also been paid to a possible role for magnesium, because diseased heart muscle tissue is seen to contain less magnesium than healthy tissue. Still, it has to be pointed out that hard waters do not necessarily contain raised concentrations of magnesium; this occurs only when the limestones through which aquifer water passes are dolomitized, and most English limestones are not. More details can be found in Chap. 14.

Mapping diseases has also been a valuable tool for a long time in China, where pioneering work has been done by Tan Jianan (1989). Modern mapping techniques are now widely used in medical geology; mapping and analytical approaches to epidemiological data are covered in Cliff and Haggett (1988), while discussions on using GIS and remote sensing, as well as several examples, are offered in Chaps. 28 and 29.

1.4.4 Dental Health Provides an Example of the Significance of Element Substitutions in Crystals

Dental epidemiology has provided some of the most convincing evidence that trace elements can affect the health of communities (Davies and Anderson 1987). Dental caries is endemic and epidemic in many countries, so a large population is always available for study. Because diagnosis relies upon a noninvasive visual inspection that minimizes ethical restrictions, a high proportion of a target population can be surveyed. Where the survey population is comprised

of children (typically 12 year olds), the time interval between supposed cause and effect is short, and it is possible to make direct associations between environmental quality and disease prevalence. In the case of fluoride, a direct link was established over 50 years ago that led to the successful fluoridation of public water supplies. This is an example of medical geology influencing public health policy. The relationship between dental caries and environmental fluoride, especially in drinking water, is probably one of the best known examples of medical geology. So strong is the relationship that the addition of 1 mg of fluoride per liter to public water supplies has been undertaken regularly by many water utilities as a public health measure.

The history of the fluoride connection is worth recounting. In 1901, Dr. Frederick McKay opened a dental practice in Colorado Springs, Colorado, and encountered a mottling and staining of teeth that was known locally as "Colorado stain." The condition was so prevalent that it was regarded as commonplace but no reference to it could be found in the available literature. A survey of schoolchildren in 1909 revealed that 87.5% of those born and reared locally had mottled teeth. Inquiries established that an identical pattern of mottling in teeth had been observed in some other American areas and among immigrants coming from the volcanic areas of Naples, Italy. Field work in South Dakota and reports from Italy and the Bahamas convinced McKay that the quality of the water supply was somehow involved in the etiology of the condition. He found direct evidence for this in Oakley, Idaho, where, in 1908, a new piped water supply was installed from a nearby thermal spring and, within a few years, it was noticed that the teeth of local children were becoming mottled. In 1925, McKay persuaded his local community to change their water supply to a different spring, after which stained teeth became rare.

A second similar case was identified in Bauxite, Arkansas, where the water supply was analyzed for trace constituents, as were samples from other areas. The results revealed that all the waters associated with mottled teeth had in common a high fluoride content 2–13 mg of fluoride per liter. In the 1930s, it was suggested that the possibility of controlling dental caries through the domestic water supply warranted thorough epidemiological-chemical investigation. The U.S. Public Health Service concluded that a concentration of 1 mg fluoride per liter drinking water would be beneficial for dental health but would not be in any way injurious to general health. Fluoride was first added to public water supplies in 1945 in Grand Rapids, Michigan. Fluoridation schemes were subsequently introduced in Brantford, Ontario (1945); Tiel, The Netherlands (1953); Hastings, New Zealand (1954); and Watford, Anglesey, and Kilmarnock in Great Britain (1955). There is no doubt that whenever fluorides have been used a reduction in the prevalence of dental caries follows (Davies and Anderson 1987; Leverett 1982) (see also Chap. 13, this volume).

1.5 An Emerging Profession

The field of medical geology (or geomedicine) has developed around the world over the last few decades. The development of activities and the organizational structure of medical geology in a number of regions will be discussed in this section, including the United States, Great Britain, Scandinavia, some African countries, and China.

As research interest in medical geology grew during the 1960s, the desire emerged for conference sessions or even entire conferences dedicated to the subject. The late Dr. Delbert D. Hemphill of the University of Missouri organized the first Annual Conference on Trace Substances in Environmental Health in 1967, and these meetings continued for a quarter of a century. Early in the 1970s, several countries took the initiative to organize activities within the field of medical geology, and a symposium was held in Heidelberg, West Germany, in October 1972. In the United States, Canada, and Great Britain, research on relationships between geochemistry and health were carried out, and the Society for Environmental Geochemistry and Health (SEGH) was established. Geochemistry has for a long time maintained a strong position in the former Soviet Union, and basic knowledge of this science is routinely applied to medical investigations. Medical geology has a long tradition in northern Europe, and the development of this emerging discipline in Scandinavia has been strong. In Norway, too, geochemical research has been regarded as important for quite some time.

In North America in the 1960s and 1970s, a number of researchers made important contributions to our understanding of the role of trace elements in the environment and their health effects; among these are Helen Cannon and Howard Hopps (1972), H. T. Shacklette et al. (1972), and Harry V. Warren (1964). A meeting on environmental geochemistry and health was held and sponsored by the British Royal Society in 1978 (Bowie and Webb 1980). Another landmark date was 1979, when the Council of the Royal Society (London) appointed a working party to investigate the role in national policy for studies linking environmental geochemistry to health. This was chaired by Professor S. H. U. Bowie of the British Geological Survey (Bowie and Thornton 1985). In 1985, the International Association of Geochemistry and Cosmochemistry (IAGC) co-sponsored with the Society for Environmental Geochemistry and Health (SEGH) and Imperial College, London, the first International Symposium on Geochemistry and Health (Thornton 1985). In 1985 Professor B E Davies became editor of a journal then titled 'Minerals and the Environment', rebranded it as 'Environmental Geochemistry and Health' (Davies 1985) and formally linked it with SEGH. The journal is now published by Springer under the editorship of Professor Wong Ming Hung and is in its 34th volume.

In 1987, a meeting on geochemistry and health was held at the Royal Society in London, and in 1993 a meeting on environmental geochemistry and health in developing countries was conducted at the Geological Society in London (Appleton et al. 1996).

Traditionally, the terms *geomedicine* and *environmental geochemistry and health* have been used. Formal recognition of the field of geomedicine is attributed to Ziess, who first introduced the term in 1931 and at the time considered it synonymous with *geographic medicine*, which was defined as "a branch of medicine where geographical and cartographical methods are used to present medical research results." Little changed until the 1970s, when Dr. J. Låg, of Norway, redefined the term as the "science dealing with the influence of ordinary environmental factors on the geographic distribution of health problems in man and animals" (Låg 1990).

The Norwegian Academy of Science and Letters has been very active in the field of medical geology and has arranged many medical geology symposia, some of them in cooperation with other organizations. The proceedings of 13 of these symposia have been published. Since 1986, these symposia have been arranged in collaboration with the working group Soil Science and Geomedicine of the International Union of Soil Science. The initiator of this series of meetings was the late Dr. Låg, who was Professor of Soil Science at the Agricultural University of Norway from 1949 to 1985 and who was among the most prominent soil scientists of his generation, having made significant contributions to several scientific disciplines. During his later years, much of Dr. Låg's work was devoted to medical geology, which he promoted internationally through his book (Låg 1990).

The countries of Africa have also experienced growth in the field of medical geology. The relationships between the geological environment and regional and local variations in diseases such as IDDs, fluorosis, and various human cancers have been observed for many years in Africa. Such research grew rapidly from the late 1960s, at about the same time that the principles of geochemical exploration began to be incorporated in mineral exploration programs on the continent. In Africa, evidence suggesting associations between the geological environment and the occurrence of disease continues to accumulate (see, for example, Davies 2003, 2008; Davies and Mundalamo 2010), but in many cases the real significance of these findings remains to be fully appreciated. The reasons are threefold: (1) the paucity of reliable epidemiological data regarding incidence, prevalence, and trends in disease occurrence; (2) the lack of geochemists on teams investigating disease epidemiology and etiology; and (3) a shortage of analytical facilities for measuring the contents of nutritional and toxic elements at very low concentration levels in environmental samples (Davies 1996). Confronting these challenges, however, could prove to be exceedingly rewarding, for it is thought that the strongest potential significance of such correlations exists in Africa and other developing regions of the world. Unlike the developed world, where most people no longer eat food grown only in their own area, most of the people in Africa live close to the land and are exposed in their daily lives, through food and water intake, to whatever trace elements have become concentrated (or depleted) in crops from their farms (Appleton et al. 1996; Davies 2000).

The first real attempt to coordinate research aimed at clarifying these relationships took place in Nairobi in 1999, when the first East and Southern Africa Regional Workshop was convened, bringing together over 60 interdisciplinary scientists from the region (Davies and Schlüter 2002). One outcome of this workshop was the constitution of the East and Southern Africa Association of Medical Geology (ESAAMEG), establishing it as a chapter of the International Medical Geology Association (IMGA). The Geomed 2001 workshop held in Zambia testified to the burst of interest and research activities generated by that first workshop (Ceruti et al. 2001). As a result of this increasing awareness of medical geology problems around the continent, membership and activities of the ESAAMEG have continued to grow. This is a welcome sign on both sides of what has hitherto been an unbridged chasm between geology and health in Africa.

China has a long history of medical geology. Chinese medical texts dating back to the third century BC contain several references to relationships between geology and health. During both the Song Dynasty (1000 BC) and the Ming Dynasty (fourteenth to seventeenth century), lung ailments related to rock crushing and symptoms of occupational lead poisoning were recognized. Similarly, as noted earlier, the Tang Dynasty alchemist Chen Shao-Wei stated that lead, silver, copper, antimony, gold, and iron were poisonous.

In the twentieth century, much research has been carried out in China (for example, on the selenium-responsive Keshan and Kashin Beck diseases) that has resulted in clarification of the causes of a number of diseases, including endemic goiter and endemic fluorosis. One of the centers for this research has been the Department of Chemical Geography at the Chinese Academy of Sciences. At this institute, several publications have been produced, such as *The Atlas of Endemic Diseases and Their Environments in the People's Republic of China* (Jianan 1985). Also the Institute of Geochemistry in Guiyang in Southern China is known for its studies in the field that is now referred to as medical geology.

International Medical Geology Association, IMGA, (Fig. 1.1) (www.medicalgeology.com) in its present form was founded in January 2006, but began as an idea 10 years before in 1996 when a working group on Medical Geology

Fig. 1.1 International Medical Geology Association, IMGA

was established by International Union of Geological Sciences (Skinner and Berger 2003). Its primary aim was to increase awareness among scientists, medical specialists and the general public on the importance of geological factors for health and wellbeing. It was recognised that the limited extent of cooperation and communication among these groups restricted the ability of scientists and public health workers to solve a range of complex environmental health problems. The term "Medical Geology" was adopted in 1997.

A first short course on Medical Geology was carried out in Lusaka, Zambia and at the University of Zambia. It was decided that the short course would be brought to developing countries which faced critical Medical Geology problems. This proposal was supported by the International Commission on Scientific Unions (ICSU) and later on UNESCO in support of short courses in Medical Geology to be held in 2002–2003. Since 2001 short courses in medical geology have been held in more than 50 countries.

A first Hemispheric Conference on Medical Geology was organized in Puerto Rico in 2005. The International Medical Geology Association, IMGA, was established 2006. This year also a special symposium on Medical Geology was held at the Royal Academy of Sciences in Stockholm. The 2nd Hemispheric Conference on Medical Geology was held in Atibaya, Brazil in 2007. United Nations announced in 2008 Medical Geology as one of the themes of the International Year of Planet Earth. IMGA also had several sessions and a short course at the 33 International Geological Conference in Oslo with 7,000 participants. In 2009 IMGA was involved in 'Mapping GeoUnions to the ICSU Framework for Sustainable Health and Wellbeing' as full members. The 3rd Hemispheric Conference on Medical Geology was held in Montevideo, Uruguay and in 2011 the 4th International conference in medical geology was organised in Italy with more than 400 participants.

The development of medical geology has been tremendous and education has started at several universities and medical geology is on the agenda all over the world with many active local chapters spread in all continents (Selinus et al. 2010).

1.6 Prospects

As we progress into the early years of the twenty-first century, it can be safely claimed that medical geology has emerged as a serious professional discipline. If respect for medical geology as a discipline is to continue to grow, then future studies must go well beyond simplistic comparisons of geochemical and epidemiological data. Dietary or other pathways must be traced and quantified and causative roles must be identified with regard to target organs or body processes. Moreover, studies must become predictive. Occasionally, simple direct links between geochemistry and health may be identified, but even in these instances confounding factors may be present (for example, the possible role of humic acids in arsenic exposure or the established role of goitrogenic substance in goiter). Ordinarily, geochemistry will provide at best only a risk factor: Unusual exposures, trace element deficiencies, or elemental imbalances will contribute toward the disturbance of cellular processes or activation of genes that will result in clinical disease. The problem of geographical variability in disease incidence will remain.

Rapid growth in the field of medical geology is predicted, as it is a discipline that will continue to make valuable contributions to the study of epidemiology and public health, providing hyperbole is avoided and a dialogue is maintained among geochemists, epidemiologists, clinicians, and veterinarians.

The structure of all living organisms, including humans and animals, is based on major, minor, and trace elements—given by nature and supplied by geology. The occurrence of these gifts in nature, however, is distributed unevenly. The type and quantity of elements vary from location to location—sometimes too much, sometimes too little. It is our privilege and duty to study and gain knowledge about natural conditions (e.g., the bioavailability of elements essential to a healthy life), and the field of medical geology offers us the potential to reveal the secrets of nature.

See Also the Following Chapters. Chapter 10 (Volcanic Emissions and Health) • Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 13 (Fluoride in Natural Waters) • Chapter 14 (Water Hardness and Health Effects)

• Chapter 15 (Bioavailability of Elements in Soil) • Chapter 16 (Selenium Deficiency and Toxicity in the Environment) • Chapter 17 (Soils and Iodine Deficiency) • Chapter 21 (Animals and Medical Geology) • Chapter 25 (Environmental Pathology) • Chapter 28 (GIS in Human Health Studies) • Chapter 29 (Investigating Vector-Borne and Zoonotic Diseases with Remote Sensing and GIS).

Further Reading

Adriano DC (2001) Trace elements in terrestrial environments: biogeochemistry, bioavailability and risk of metals. Springer, Berlin

Agricola G (1950) De Re Metallica (1556) (trans: Hoover H and Hoover LH). Dover Publications, New York

Alvarez LW, Alvarez W, Asaro F, Michel H (1980) Extraterrestrial cause for cretaceous-tertiary extinction. Science 208:1095–1108

Anderson PH, Berrett S, Patterson DSP (1979) The biological selenium status of livestock in Britain as indicated by sheep erythrocyte glutathione peroxidase activity. Vet Rec 104(March 17):235–238

Appleton JD, Fuge R, McCall GJH (eds) (1996) Environmental Geochemistry and Health with Special Reference to Developing Countries. Geological Society Special Publication no. 113. Geological Society, London

Bowie SHU, Thornton I (eds) (1985) Environmental geochemistry and health. D. Reidel Publishing, Dordrecht

Bowie SHU, Webb JS (1980) Environmental geochemistry and health. Royal Society, London

Cannon, Helen L, Hopps, Howard C (1972) Geochemistry and the Environment: Volume I: The relation of selected trace elements to health and disease. A Report of the Workshop at the Asilomar Conference Grounds; February 7–12, 1972 U.S. National Committee for Geochemistry Division of Earth Sciences National Research Council National Academy of Sciences Washington, D.C. 1974; p. 133

Carruthers M, Smith B (1979) Evidence of cadmium toxicity in a population living in a zinc mining area-Pilot survey of Shipham residents. Lancet 1(8121):845–847

Ceruti PO, Davies TC, Selinus O (2001) Geomed 2001—medical geology, the African perspective. Episodes 24(4):268–270

Chaney RL, Ryan JA, Li Y-M, Brown SI (1998) Soil cadmium as a threat to human health. In: McLaughlin MJ, Singh BR (eds) Cadmium in soils, plants and the food chain. Kluwer, Dordrecht, pp 219–256

Chen X, Yang G, Chen J, Chen X, Wen Z, Ge K (1980) Studies on the relations of selenium and Keshan disease. Biol Trace Element Res 2:91–107

Cliff DC, Haggett P (1988) Atlas of disease distributions. Blackwell Reference, Oxford

D'itri PA, D'itri FM (1977) Mercury contamination: a human tragedy. Wiley, New York

Davies BE (1971) Trace element content of soils affected by base metal mining in the west of England. Oikos 22:366–372

Davies BE (1985) Airs, waters and places: editorial. Environ Geochem Health 7:1

Davies TC (1996) Geomedicine in Kenya. In: Davies TC (ed) Environmental geology of Kenya. Special issue, J Afr Earth Sci 23 (4):577–591

Davies TC (2000) Editorial. Environ Geochem Health 24(2):97 (Special issue)

Davies TC (2003) Historical development of medical geography in Africa. Moi University, Kenya (Unpublished report)

Davies TC (2008) Environmental health impacts of east African rift volcanism. Environ Geochem Health 30(4):325–338

Davies BE, Anderson RJ (1987) The epidemiology of dental caries in relation to environmental trace elements. Experientia 43:87–92

Davies BE, Lewin J (1974) Chronosequences in alluvial soils with special reference to historic lead pollution in Cardiganshire, Wales. Environ Pollut 6:49–57

Davies TC, Mundalamo HR (2010) Environmental health impacts of dispersed mineralisation in South Africa. J Afr Earth Sci 58 (4):652–666

Davies TC, Schlüter T (2002) Current status of research in geomedicine in East and Southern Africa. Environ Geochem Health 24(2):99–102

Davies BE, Thornton I (1989) Environmental pathways of lead into food: a review. The International Lead Zinc Research Organization, Research Triangle

Diplock AT (1984) Biological effects of selenium and relationships with carcinogenesis. Toxicol Environ Chem 8:305–311

Eaton DL, Robertson WO (1994) Toxicology. In: Rosenstick L, Cullen WR (eds) Textbook of clinical, occupational and environmental medicine. W.B. Saunders, Philadelphia, pp 116–117

Engels F (1845) The condition of the working class in England. In: Kernan V (ed) Penguin classics. Penguin Books, London

Esmark J (1823) Om noritformation. Magazin for Naturvideskaberne, 1:205

EPA (2011) Revised technical support document: national-scale assessment of mercury risk to populations with high consumption of self-caught freshwater fish. (U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Health and Environmental Impacts Division), Research Triangle Park, NC, EPA-452/R-11-009

Ferguson WS, Lewis AH, Watson SJ (1943) The teart pastures of Somerset. I. The cause and cure of teartness. J Agric Sci (Cambridge) 33:44–51

Fergusson JE (1990) The heavy elements: chemistry, environmental impact and health effects. Pergamon Press, New York

Finkelman RB, Belkin HE, Zheng B (1999) Health impacts of domestic coal use in China. Proc Natl Acad Sci 96:3427–3431

Foster HD (2002) The geography of disease family trees: the case of selenium. In: Balkema AA, Bobrowsky P (eds) Geoenvironmental mapping: methods, theory and practice. A. A. Balkema Publishers, Lisse, pp 497–529

Frankenberger WT, Benson S (eds) (1994) Selenium in the environment. Marcel Dekker, New York

Frankenberger WT, Engberg RA (eds) (1998) Environmental chemistry of selenium. Marcel Dekker, New York

Friberg L, Piscator M, Nordberg G, Kjellstrom T (1974) Cadmium in the environment, 2nd edn. CRC Press, Boca Raton

Gallacher J, Elwood MD, Phillips KM, Davies BE, Jones DT (1984) Relation between pica and blood lead in areas of differing lead exposure. Arch Dis Child 59:40–44

Gimingham CT (1914) The scouring lands of Somerset and Warwickshire. J Agric Sci (Cambridge) 6:328–336

Goldschmidt VM (1954) Geochemistry. Clarendon Press, London

Hallberg R (ed) (2007) Special issue on medical geology. AMBIO, Roy Swed Acad Sci 37(1)

Harada M (1995) Minamata disease: methylmercury poisoning in Japan caused by environmental pollution. Crit Rev Toxicol 25:1–24

Hernberg S (2000) Lead poisoning in a historical perspective. Am J Ind Med 38:244–254

Hong S, Candelone JP, Patterson CC, Boutron CF (1994) Greenland ice evidence of hemisphere lead pollution two millennia ago by Greek and Roman civilizations. Science 265:1841–1843

Howe GM (1963) National atlas of disease mortality in the United Kingdom. Thomas Nelson and Sons Ltd., London, 111 pp

Jianan T (1985) The atlas of endemic diseases and their environments in the People's Republic of China. Science Press, Beijing

Kabata-Pendias A (2001) Trace elements in soils and plants, 3rd edn. CRC Press, Boca Raton

Kabata-Pendias A, Pendias H (1992) Trace elementsin soils and plants, 2nd edn. CRC Press, Boca Raton
Krzysztof K, Glab H (2001) Trace element concentrations in human teeth from Neolithic common grave at Nakonowo (Central Poland). Variabil Evol 9:51–59
Låg J (1990) General survey of geomedicine. In: Låg J (ed) Geomedicine. CRC Press, Boca Raton, pp 1–24
Latham R (1958) The travels of Marco Polo. Penguin, London
Leverett DH (1982) Fluorides and the changing prevalence of dental caries. Science 217:26–30
Lewis AH (1943) The teart pastures of Somerset. II. Relation between soil and teartness. J Agric Sci (Cambridge) 33:52–57
Lindberg DC (1992) The beginnings of western science. University of Chicago Press, Chicago
Mahaffey KR, Annest JL, Roberts J, Murphy RS (1982) National estimates of blood lead levels: United States, 1976–1980. N Engl J Med 307:573–579
Millar IB (1961) Gastrointestinal cancer and geochemistry in North Montgomeryshire. Br J Cancer 15(2):176–199
Morgan H, Simms DL (1988) The Shipham report: discussion and conclusions. Sci Total Environ 75:135–143
Moynahan EJ (1979) Trace elements in man. Phil Trans R Soc Lond 288:65–79
Nriagu JO (1983) Lead exposure and lead poisoning. In: Lead and lead poisoning in antiquity. Wiley, New York, pp 309–424
Nriagu JO (1998) Tales told in lead. Science 281:1622–1623
Prasad AS (1978) Trace elements and iron in human metabolism. Wiley, Chichester
Ratcliffe JM (1981) Lead in man and the environment. Ellis Horwood, New York
Roberts C, Manchester K (2007) The archaeology of disease. Cornell University Press, Ithaca, 338 pp
Rosenfeld I, Beath OA (1964) Selenium: geobotany, biochemistry, toxicity, and nutrition. Academic, New York
Selinus O, Finkelman B, Centeno J (eds) (2010) Medical geology—a regional synthesis. Springer, Dordrecht
Shacklette HT, Sauer HI, Miesch AT (1972) Distribution of trace elements in the environment and the occurrence of heart disease in Georgia. Bull Geol Soc Am 83:1077–1082
Shuttleworth VS, Cameron RS, Alderman G, Davies HT (1961) A case of cobalt deficiency in a child presenting as "earth eating". Practitioner 186:760–766
Silver CS, Rothman DS (1995) Toxics and health: the potential long-term effects of industrial activity. World Resources Institute, Washington, DC
Skinner HCW, Berger A (eds) (2003) Geology and health: closing the gap. Oxford University Press, New York
Stocks P, Davies RI (1964) Zinc and copper content of soils associated with the incidence of cancer of the stomach and other organs. Br J Cancer 18:14–24
Thomson I, Thornton I, Webb JS (1972) Molybdenum in black shales and the incidence of bovine hypocuprosis. J Sci Food Agric 23:879–891
Thornton I (ed) (1985) Proceedings of the first international symposium on geochemistry and health. Science Reviews, Ltd., St. Albans, UK
Thornton I, Atkinson WJ, Webb JS, Poole DBR (1966) Geochemical reconnaissance and bovine hypocuprosis in Co. Limerick, Ireland. Irish J Agric Res 5(2):280–283
Thornton I, Moon RNB, Webb JS (1969) Geochemical reconnaissance of the lower lias. Nature (London) 221:457–459
Thornton I, John S, Moorcroft S, Watt J (1980) Cadmium at Shipham—a unique example of environmental geochemistry and health. Trace Substances Environ Health Symp 14:27–37
Underwood EJ, Filmer JF (1935) The determination of the biologically potent element (cobalt) in limonite. Aust Vet J 11:84–92
Voisin A (1959) Soil, grass and cancer (trans: Herriot CTM, Kennedy H). Crosby Lockwood & Son, London
Warren HV (1964) Geology, trace elements and epidemiology. Geogr J 130(Pt. 4):525–528
Warren HV, Delavault RE (1971) Variations inthe copper, zinc, lead, and molybdenum contents of some vegetables and their supporting soils. Geol Soc Am Mem 123:97–108
Wyllie J (1937) An investigation of the source of arsenic in a well water. Can Publ Health J 28:128–135
The Norwegian Academy of Science and Letters has published several proceedings from symposia on medical geology with Professor Jul Låg as editor. These include:
Geomedical aspects in present and future research (1980)
Geomedical research in relation to geochemical registrations (1984)
Geomedical consequences of chemical composition of freshwater (1987)
Commercial fertilizers and geomedical problems (1987)
Health problems in connection with radiation from radioactive matter in fertilizers, soils, and rocks (1988)
Excess and deficiency of trace elements in relation to human and animal health in Arctic and Subarctic regions (1990)
Human and animal health in relation to circulation processes of selenium and cadmium (1991)
Chemical climatology and geomedical problems (1992)
Geomedical problems related to aluminum, iron, and manganese (1994)
Chemical data as a basis of geomedical investigations (1996)
Some geomedical consequences of nitrogen circulation processes (1997)
Geomedical problems in developing countries (2000)
Bölviken B (ed) (2001) Natural ionizing radiation and health. Trondheim, Norway
"Environmental Geochemistry and Health", published in six issues annually by Springer, Dordrecht, The Netherlands, is an international journal focussed on the interaction of the geochemical environment and health and is the official Journal of the Society for Environmental Geochemistry and Health

Public Health and Geological Processes: An Overview of a Fundamental Relationship

2

Angus Cook

Contents

2.1 Public Health Definitions and Approaches

The health of human populations is highly dependent on a range of environmental determinants, including those relating to geologic materials and processes on our planet (Fig. 2.1). There is a clear continuum of geochemical cycles and pathways for uptake of elements into plants, animals and humans. Indeed, the optimal functioning of the human body itself is reliant upon the maintenance of an appropriate balance of elements and minerals. The discipline of medical geology can assist in the elucidation and management of many public health issues, ranging from the effects of exposure to or deficiencies of trace elements and minerals in the diet, to the transportation, modification and concentration of natural compounds in the soil, atmosphere and water sources.

The activity known as *public health* has been subject to a range of interpretations. In fundamental terms, however, its focus is to provide *population-based solutions to collectively-defined health problems*. The emphasis on the incidence, distribution, and control of diseases within a population is often contrasted with the focus of clinical medicine, which is directed towards how illness affects an individual patient.

One widely cited definition is from the *Acheson Report on Public Health* (1988), in which public health is described as:

> The science and art of preventing disease, prolonging life, and promoting health through organised efforts of society.
>
> (Great Britain. Dept. of Health and Social Security. Committee of Inquiry into the Future Development of the Public Health Function. and Great Britain. Dept. of Health and Social Security. 1988)

Another, more extensive definition, was provided by John Last in the *Dictionary of Epidemiology* (1995):

> Public health is one of the efforts organised by society to protect, promote, and restore the people's health. It is the combination of sciences, skills, and beliefs that is directed to the maintenance

A. Cook (✉)
School of Population Health, The University of Western Australia, Crawley, Australia
e-mail: angus.cook@uwa.edu.au

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_2,

Fig. 2.1 A community exposed to occupational and environmental geogenic material: an iron-ore mining town located in the arid Pilbara region, Western Australia (Picture courtesy of Fiona Maley)

> and improvement of the health of all the people through collective or social actions. The programs, services, and institutions involved emphasize the prevention of disease and the health needs of the population as a whole. Public health activities change with changing technology and social values, but the goals remain the same: to reduce the amount of disease, premature death, and disease-produced discomfort and disability in the population. Public health is thus a social institution, a discipline, and a practice.
>
> (Last and International Epidemiological Association 1995)

There are two core components needed for the operation of a successful public health system: an "*analysis/investigation" function* and *an "action/intervention" function*. The first set of activities includes monitoring and surveillance of communicable (infectious) and non-communicable – often more chronic – diseases. In response to the information gathered from these sources, an intervention – or group of interventions – is then implemented. These may range across a wide spectrum of activities, including health legislation (such as those relating to tobacco control or driving under the influence of alcohol or drugs), regulations to control occupational or environmental hazards, the provision of specific public health services (such as screening programs), or the implementation of health promotion measures (such support of health lifestyles and behaviours) (World Health Organization 1998). The public health measures may be directed towards the entire population or they may, depending on the health issue of interest, target vulnerable and high-risk populations.

This chapter will explore a number of these public health approaches in greater detail, where possible using examples to illustrate how medical geology can and has contributed to the elucidation and mitigation of public health problems.

2.2 Historical Background

Although many of the features of modern public health systems arose in the nineteenth and twentieth centuries, there were a number of precedents in earlier periods. The Classical Romans constructed systems to supply clean water

to their cities – including impressive networks of aqueducts, cisterns and domestic pipes – although these were often lined with a malleable metal (*plumbum*, or lead) whose toxic effects were not fully appreciated. The public health implications of fuel use became apparent in the early 1300s, with the banning of coal burning in London because it lead "to the annoyance of the magistrates, citizens, and others there dwelling and to the injury of their bodily health" (Nef 1966). In the expanding cities of late medieval Europe, there were attempts to regulate urban waste and to limit the movements of infected people (such as those with leprosy). The Italian city-states were notable in their achievements, with the foundation of civic hospitals (including specialised plague hospitals), establishment of public Boards of Health (such as the Milan *Sanità*, 1424), and the strict enforcement of quarantine regulations. In the 1600s, the field of medical demography and analysis of population data advanced with the publication of John Graunt's *Natural and Political Observations … upon the Bills of Mortality* (1662), in which it was noted that more boys were born than girls and that urban mortality exceeds that of rural areas (Porter 1998).

However, it was the Industrial Revolution and subsequent societal shifts in the nineteenth century that created a new series of health challenges. Increasing numbers of people migrated from the countryside to the cities, where many were forced to take on dangerous, arduous work for near-subsistence wages. The population was growing rapidly without a concurrent expansion in new housing, and overcrowding contributed to the relatively fast spread of disease in many urban centres.

One of the leading reformers of this period was Sir Edwin Chadwick (1800–1890), a lawyer from Manchester who took an active part in the modifications to the Poor Law and in factory legislation. Public health became a political issue in 1842 with the publication of Chadwick's *Report on the Sanitary Condition of the Labouring Population of Great Britain* (Chadwick and Great Britain. Poor Law Commissioners. 1842). Chadwick had collated vast quantities of vital statistics, information from town maps, and descriptions of dwellings and problems with living conditions (such as inadequate drainage and foul odours). Chadwick viewed the problems of destitution, slums, smoke, water supply and sewerage as matters for *public action,* as opposed to simply being the responsibility of individuals or local authorities (Porter 1998). Chadwick's advocacy and influence contributed to the first *Public Health Act* in 1848, which established a General Board of Health. This central authority was empowered to set up local boards with the task of ensuring that new homes had proper drainage and that local water supplies were dependable. Furthermore, the first Medical Officer of Health for Great Britain was appointed. By 1872, local boards of health were compelled to employ medical officers of health who were responsible for sanitary inspection and improvement, and isolation and tracing of persons with infections. Other public health legislation, covering factory management, child welfare, care of the elderly and those with physical or mental disabilities, and many other initiatives in social reform were introduced.

In the United States, boards of health were created in the eastern seaboard cities in the 1830s, but they were largely concerned with quarantine regulations. A national awareness of epidemic disease arose from the trauma of the Civil War (1861–1865), in which a large proportion of deaths arose from infections (especially dysentery). Louisiana formed the first state board of health in 1855 and most other states followed by the 1870s. Social conditions deteriorated in the large cities between 1860 and 1890 as a result of increased industrialisation and immigration, leading to a range of social reforms, including those relating to public health. In 1872, the American Public Health Association was formed as a multidisciplinary consortium of engineers, physicians and public-spirited citizens, especially members of the women's movement who spoke of sanitary reform as 'municipal housekeeping' (Rosen 1993).

New discoveries in the field of microbiology also fired a series of new public initiatives and a greater degree of community awareness about infectious disease. Improved methods for water and food sanitation and safety emerged. Rodent and pest control grew in importance. Individual hygiene practices were encouraged. In 1920, Charles Winslow, an American authority on public health and professor at Yale University, emphasised the importance of *"organised community efforts for the sanitation of the environment", "the control of community infections"* and, more broadly, *"the development of the social machinery which will ensure to every individual in the community a standard of living adequate for the maintenance of health"* (Winslow 1920).

Despite these improvements in public health in the early years of the twentieth century, there were a number of unwanted effects of increasing industrialisation, population growth and urbanisation. The environmental and health impacts of widespread mining, metallurgical activities (such as smelting and refining) and fuel extraction (such as coal) also became increasingly apparent (Krech et al. 2004). High levels of pollutants were released into the atmosphere, waterways and soils. The Industrial Revolution was also associated with a growing incidence of occupational diseases, such as lung disorders (examples include coal workers pneumoconiosis, silicosis, and asbestosis), occupational cancers, and poisoning from various toxic compounds (including lead, white phosphorous, mercury, cadmium, arsenic and radioactive agents).

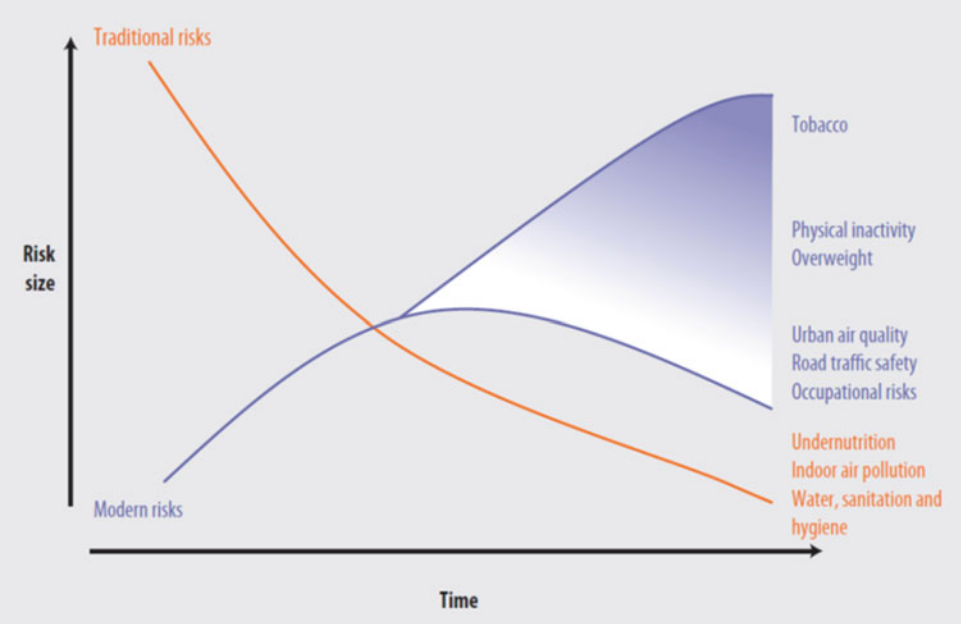


Fig. 2.2 Risk transition in public health. The diagram indicates the magnitude of the shift from "traditional" risks (such as from inadequate nutrition and poor water supply) to "modern" risks over time. The scale of the time variable varies depending on the particular community (Source: World Health Organisation 2009: http://www.who.int/healthinfo/global_burden_disease/global_health_risks/en/index.html)

2.3 Current and Future Trends and Priorities in Public Health

In general terms, industrialised, high-income nations have undergone a significant shift in disease patterns in the past century (Mathers et al. 2009). The increasing average wealth of a nation broadly corresponds to improvements in hygiene and nutrition, substantial declines in the frequency of infectious diseases, and improved population health indicators (such as life expectancy). In many respects, high-income nations have many of the previous "external" or "environmental" causes of disease under control through prevention or management of infectious disease, removal of many water and food contaminants, and improved safety standards and regulations (Fig. 2.2). However, affluence has also been accompanied by increased levels of sedentary behaviour and overnutrition, contributing to such problems as obesity, metabolic disorders and cardiovascular disease. The causes of death in industrialised society primarily relate to chronic diseases – such as cancer, cardiovascular disease, some respiratory diseases – and to a lesser degree, injury and accidents (e.g. in workplaces or from motor vehicle accidents). Many of these current health problems relate – at least partly – to attitudes and behaviours relating to smoking, diet, exercise levels and alcohol intake. As leading researchers have noted with respect to one of the major "lifestyle" diseases, type II diabetes mellitus: "overall, a healthy diet, together with regular physical activity, maintenance of a healthy weight, moderate alcohol consumption, and avoidance of sedentary activities and smoking, could nearly eliminate [this disorder]" (Schulze and Hu 2005).

There remains a significant contrast in health status between those in lower- and higher-income countries. A comparison of causes of death (in millions) for low-, middle- and high-income countries (2004) is provided in Table 2.1. Traditionally, as we noted above, affluence has been linked to those diseases arising from tobacco use, consumption of alcohol and energy-rich foods, and a sedentary lifestyle. In contrast, those living in poverty have usually been perceived as being at risk of inadequate health infrastructure and limited access to care, with the prospect of famine, recurrent infections and a limited life span.

However, this overly simplified description does not capture the current and evolving profile of global health (Lopez and Disease Control Priorities Project 2006). In recent decades, there have been profound changes in many "developing" countries (especially across Asia and Latin America) with accelerating industrialisation and economic development, urbanisation of populations, and globalisation of food systems. Life expectancy is increasing in many parts of Latin America, North Africa, the Middle East and Central and East Asia, although there remain significant variations within and between communities. This rise in average life expectancy primarily relates to a decrease in infant/childhood deaths.

An additional shift has been the "nutrition transition": that is, a shift away from traditional diets to "Westernised" diets (highly processed, high energy, low fibre) and towards lower levels of physical activity. This has led to the co-existence of under- and over-nutrition in many middle-income countries. Rates of cardiovascular disease and diabetes are now dramatically escalating (Yach et al. 2006).

At the same time, major health problems in lower-income countries continue to be associated with poor sanitation,

Table 2.1 Comparison of causes of death (in millions) between low-, middle- and high-income countries (2004)

Disease or injury	Deaths (millions)	Percent of total deaths	Disease or injury	Deaths (millions)	Percent of total deaths
World			***Low-income countries***		
1 Ischaemic heart disease	7.2	12.2	1 Lower respiratory infections	2.9	11.2
2 Cerebrovascular disease	5.7	9.7	2 Ischaemic heart disease	2.5	9.4
3 Lower respiratory infections	4.2	7.1	3 Diarrhoeal diseases	1.8	6.9
4 COPD	3.0	5.1	4 HIV/AIDS	1.5	5.7
5 Diarrhoeal diseases	2.2	3.7	5 Cerebrovascular disease	1.5	5.6
6 HIV/AIDS	2	3.5	6 COPD	0.9	3.6
7 Tuberculosis	1.5	2.5	7 Tuberculosis	0.9	3.5
8 Trachea, bronchus, lung cancers	1.3	2.3	8 Neonatal infections	0.9	3.4
9 Road traffic accidents	1.3	2.2	9 Malaria	0.9	3.3
10 Prematurity and low birth weight	1.2	2	10 Prematurity and low birth weight	0.8	3.2
Middle-income countries			***High-income countries***		
1 Cerebrovascular disease	3.5	14.2	1 Ischaemic heart disease	1.3	16.3
2 Ischaemic heart disease	3.4	13.9	2 Cerebrovascular disease	0.8	9.3
3 COPD	1.8	7.4	3 Trachea, bronchus, lung cancers	0.5	5.9
4 Lower respiratory infections	0.9	3.8	4 Lower respiratory infections	0.3	3.8
5 Trachea, bronchus, lung cancels	0.7	2.9	5 COPD	0.3	3.5
6 Road traffic accidents	0.7	2.8	6 Alzheimer and other dementias	0.3	3.4
7 Hypertensive heart disease	0.6	2.5	7 Colon and rectum cancers	0.3	3.3
8 Stomach cancer	0.5	2.2	8 Diabetes mellitus	0.2	2.8
9 Tuberculosis	0.5	2.2	9 Breast cancer	0.2	2
10 Diabetes mellitus	0.5	2.1	10 Stomach cancer	0.1	1.8

Source: http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/index.html
COPD chronic obstructive pulmonary disease

inadequate hygiene standards, low vaccination coverage, poor access to health services, and malnutrition. Most preventable deaths in these nations occur in children: globally, around 11 million per year die before the age of 5 years, and it is estimated that half of these could be averted through simple, low-cost solutions (Black et al. 2006). The main causes of premature mortality in lower income nations are infections i.e. respiratory tract infections (such as pneumonia), vaccine-preventable disease (including measles, whooping cough, poliomyelitis, tetanus) and diarrhoeal illnesses, many of which are superimposed on a background of poverty and malnutrition. Malaria and HIV/AIDS are also important contributors to mortality (World Health Organization 2003). These health concerns have prompted a range of responses at the local and international level, including the development of the United Nations Millennium Development Goals (MDGs), which seek to alleviate extreme poverty, hunger, illiteracy and disease (UN Millennium Project et al. 2005; World Health Organization 2005).

It is critical to consider the wider social, economic and political context when addressing public health issues. There is ample historical evidence that public health efforts have been most successful when the socio-cultural context has been changed. For example, the separation of sewage from drinking water is arguably the greatest of all public health achievements, although it remains to be achieved in many societies.

Social deprivation and inequality are strong predictors of health status within and between populations, and of trends over time, and consequently much public health effort is directed towards the alleviation of these imbalances (e.g., the World Health Organisation's *Health for All by the Year 2000* (World Health Organization 1981)). Social inequalities lead to differentials across a range of health measures, including: (i) overall mortality rates and life expectancies; (ii) morbidity and disability from a range of specific conditions (e.g., infectious diseases, cardiovascular diseases, adverse perinatal outcomes, injury and poisoning); (iii) risk behaviours/factors (e.g., smoking, physical inactivity; lipid/cholesterol profiles; higher consumption of refined sugars and fats; less fruit and vegetable intake); (iv) physical environmental factors (e.g., residence in flood-prone regions; proximity to contaminated sites, such as those with elevated lead levels from old housing, leaded fuels, and industry);

(v) utilisation of health services (disadvantaged groups tend to have reduced utilisation of preventive services, such as screening). The patterns and underlying causes of such health disparities have been explored in detail in a number of reports and publications (Berkman and Kawachi 2000; Great Britain Working Group on Inequalities in Health and Great Britain Department of Health and Social Security 1980; Marmot and Feeney 1997; Marmot 2005). However the relationship today between health and average income in developing countries is not a simple correlation. Regions with significant improvements in health despite low per person wealth are Sri Lanka, Costa Rica, Kerala in India and parts of China. Mass education and political commitment to primary care are considered to have played an important part in such improvements (Black et al. 2006).

A related issue in public health is the nature and structure of the health system itself. This encompasses not only the specific health facilities (such as clinics and hospitals), but the entire set of related social structures and processes that impinge on the well-being of the community (Scutchfield and Keck 2009). Attributes of the health system that affect health status and the success of disease control include: (i) overall allocation of resources at the governmental level, such as how much of the national budget is to be spent on health as opposed to other areas (such as defence); (ii) the range and effectiveness of preventive, treatment, rehabilitative and continuing care programs available; (iii) access to services (e.g. universality of health care; systems of pricing health services); (iv) the provision of special services to disadvantaged groups.

In 1981, the World Health Assembly endorsed a resolution that health was to be one of the major social goals of government, thereby initiating the "*Health for All by the Year 2000*" (HFA2000) movement (World Health Organization 1981). In this period, health policy shifted progressively towards emphasising *community-based interventions* as the main vehicle for improving health. This has been accompanied by a general shift away from an overly "medicalised" model of health care. This transition was taken further by the *Ottawa Charter* (1986), which advanced the view that it is *people themselves* that hold power in health matters, not just institutions, officials, professionals or those who develop and control technology (Green and Kreuter 1990; Green and Tones 1999). The Charter stressed the importance of *health promotion*, defined as follows:

> Health promotion is the process of enabling people to increase control over, and to improve, their health. . . Health is a positive concept emphasizing social and personal resources, as well as physical capacities. Therefore, health promotion is not just the responsibility of the health sector, but goes beyond healthy lifestyles to well-being.
>
> *Ottawa Charter* (1986) (World Health Organization. Division of Health Promotion. 1986)

A number of key areas of activity were identified to achieve the health promotion model. These include the *creation of supportive environments*, with provision of living and working conditions that are safe, stimulating, satisfying, enjoyable and conducive to health-enhancing behaviours. This emphasises the need to identify and manage environmental barriers that might prohibit or inhibit optimal health. Another element of promoting health is through *strengthening of community actions*, with an emphasis on using and supporting community resources to promote health. This approach encourages community "ownership" of health and empowerment, in which individuals act collectively to manage the determinants of their health and the quality of life in their community.

2.4 Measuring Health and Disease in Populations

Health and health status is a complex and abstract concept: there is no single satisfactory definition which encompasses all aspects of health. Traditionally, health has been viewed as the absence of disease. However, the World Health Organization defined health much more broadly as "*a state of physical, emotional and social well-being, and not merely the absence of disease or infirmity.*" Although there has been considerable debate about the merits of this definition, it has provided a focus for the measurement of a broad range of indicators of health beyond just the presence or absence of some clinical or pathological entity.

In attempting to capture a population's overall "state of health," there is a wide array of measures from which to choose. However, most of the commonly used measures have a shared goal: to meaningfully quantify and summarise some dimension of health or disease in a population. An example is the *mortality rate* (or *death rate).* These take different forms: the *crude* mortality rate is the number of deaths in an entire population over a set period (usually a year), expressed per unit of population at risk of dying, whereas *specific* mortality rates are the number of deaths occurring within a subgroup of the population, such as by gender or across particular age strata or ethnic groups. A common measure used for international comparisons is the *infant mortality rate* (IMR): the ratio of the number of deaths of children under 1 year of age (in a given year) divided by the number of live births (in that year).

A related population-based measure is *life expectancy*: the average number of years yet to be lived for individuals at a specified age (commonly taken from birth). It is based on a set of age-specific death rates, and usually draws on the mortality conditions prevailing at the time of the estimate. In low-income countries with high infant mortality, most

of the gains in life expectancy occur by improving the likelihood of survival of infants. In contrast, for high-income countries with low infant mortality, most of the gains in life expectancy occur in the oldest members of the population.

However, measures such as mortality rates and life expectancy only capture part of a community's experience of health and disease. For example, mental illness, visual loss and osteoarthritis are all major sources of ill-health in the population, but the extent of such conditions is not reliably captured by examining death rates. Examples from the field of medical geology include the iodine deficiency disorders (IDDs), which are estimated to affect hundreds of millions of people worldwide. A considerable community burden arises from serious neurological deficits (including cretinism and hearing impairments) arising from inadequate iodine during foetal and early childhood development (De Benoist et al. 2004; Mason et al. 2001). However, IDDs in children and adults are not commonly a direct cause of death, and therefore will not usually be detected solely by examining mortality records.

Other measures are required to capture the frequency of a condition or disease in a population, also known as *morbidity*. One of the most important measures of disease frequency is the *incidence*. Incidence reflects the frequency of *new* health- or disease- related events (such as the first onset of a particular disease). These may be expressed as rates, such as new cases per person-year, per person-month, or per person-day. These values may be converted to a rate per 1,000 or 100,000 person-years (for example, cancers are often reported in terms of an incidence rate of *X* cases "per 100,000 person-years".) *Prevalence* is another common measure of morbidity. This encompasses the frequency of *existing* (whether new or pre-existing) *health- or disease-related events*. In other words, prevalence focuses on *disease status,* or the *current "burden of disease"*, in a community, as opposed to the number of new events (which is captured using incidence). Common sources of morbidity data include: disease registries (e.g. for cancers); notification systems for infectious diseases, industrial diseases and accidents, discharge information from hospitals and registers in general practice (family medicine clinics).

In terms of assessing the frequency of events, morbidity differs from mortality in some crucial respects. Death is a well-documented and final state, and is relatively easily enumerated. When counting and summarising other (non-fatal) "disease states" across the community, however, there are a far wider range of measures available. At the biological level, morbidity may for example be assessed using *biochemical markers* (e.g. blood glucose), *physiological markers* (e.g. blood pressure), and *pathological markers* (e.g. tumour size; evidence of dental or skeletal fluorosis). Alternatively, functional measures of morbidity reflect how a disease or condition results in *impairment* (reduction in physical or mental capacities at an organ or bodily system level), *disability* (restriction in a person's ability to perform a particular task or function, such as walking) and *handicap* (limitations in a person's ability to fulfil a normal social role, such as their usual job).

Morbidity may also be assessed using *self-reported measures*, which are used to assess factors associated with quality of life, such as "wellness," psychological and emotional wellbeing, or social functioning. These aspects of poor health and disability are of great importance to patients but are often difficult to measure.

'Composite' measures of health status combine the aspects of those described above, such as life expectancy and morbidity indicators. One example is the *Disability Adjusted Life Year, or DALY*. This metric takes into account both the fatal and non-fatal outcomes of disease. One DALY can be thought of as one lost year of "healthy" life due to death, disease or disability. It quantifies the health burden in terms of years of life lost (due to premature death) and years "lost" to disability, with different "weights" assigned to medical conditions depending on their severity. One of the most important functions of DALYs is to assess the national (or global) burdens of disease and injury associated with certain diseases or groups of diseases. In higher-income countries, significant DALYs are associated with cardiovascular diseases, mental illness, dementia, cancers, diabetes and injury (e.g. road traffic accidents). In middle- and low-income countries, cardiovascular diseases, and mental illnesses are also important, but infectious diseases and perinatal conditions also account for a significant burden (Lopez and Disease Control Priorities Project 2006). The use of DALYs can be extended to assess the global burden of disease and injury associated with particular risk factors, such as DALYs "lost" from malnutrition (including iron and/or iodine deficiencies), unsafe water supplies (including those containing excess fluoride or arsenic), tobacco and alcohol use.

In general, the uses of these various sources of health data are diverse. Information on the frequency of diseases causing death and disability provide a general framework for health policy. Morbidity statistics are of particular relevance to health services planning. Statistics on utilisation patterns and the diseases and illnesses presenting to hospitals and other health care institutions are essential to a rational policy for the organisation of health care.

2.5 Assessing Risk in the Community

The term "risk" has multiple interpretations, and is used in different contexts in the field of public health. A *risk factor*, for example, denotes a factor or characteristic associated with (but not necessarily always the direct cause of) a

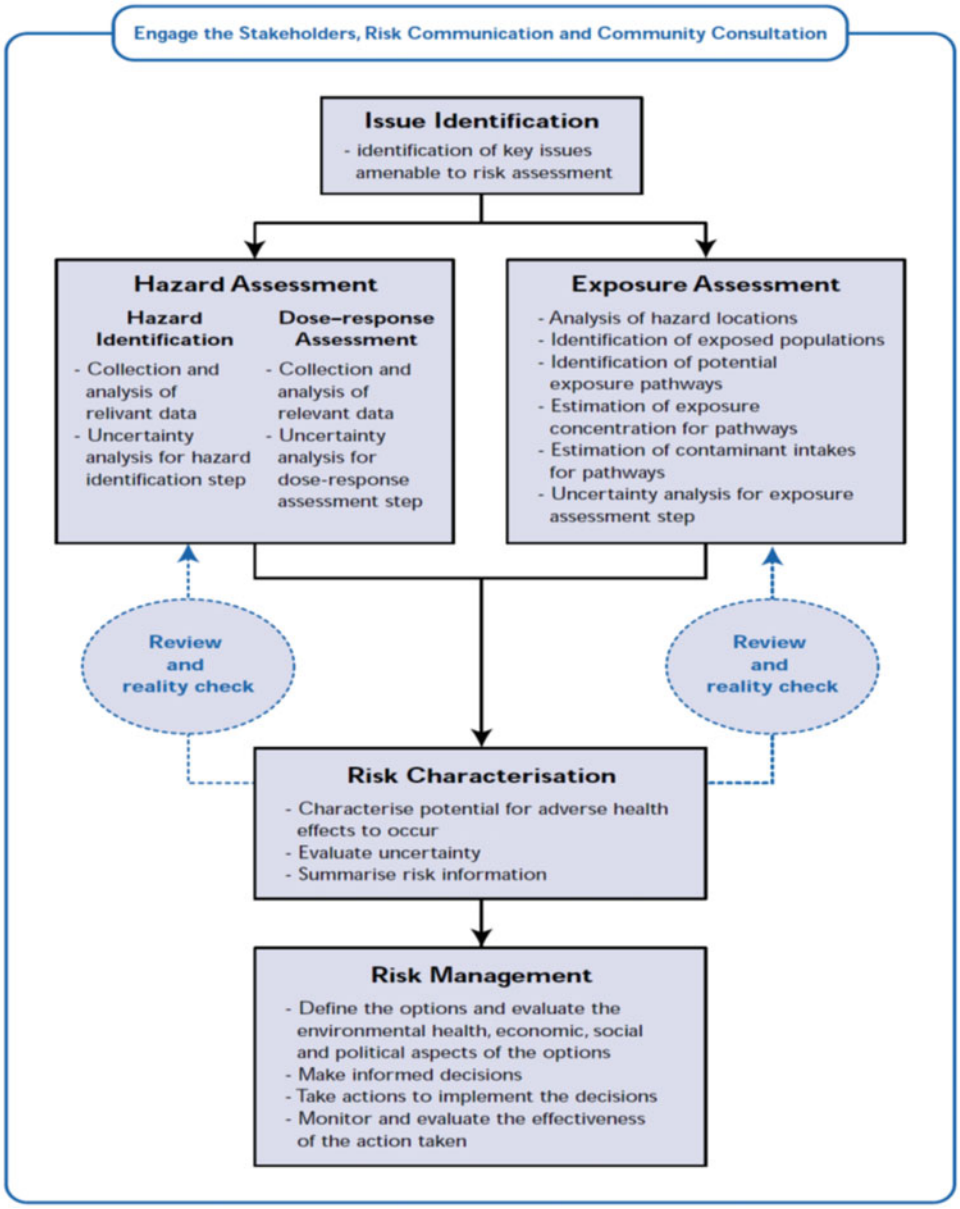


Fig. 2.3 Stages of risk assessment for environmental hazards (from Environmental health risk assessment: guidelines for assessing human health risks from environmental hazards, used by permission of the Australian Government) (Source: Department of Health and Ageing, Canberra, Australia/ enHealth Council 2002: http://www.nphp.gov.au/enhealth/council/pubs/pdf/envhazards.pdf)

particular disease or outcome. A risk factor can be an aspect of behaviour or lifestyle, an environmental exposure, or an inborn or inherited characteristic, that is associated with an increased occurrence of disease or other health-related event or condition.

Another usage of the term "risk" is within the framework of a *risk assessment*. In general, this process involves estimating the potential impact of chemical, biological or physical agents on a specified human population over a given time. In a medical geology setting, such assessments can be considered in terms of: (i) *defining exposure pathways*: such as the nature of the geological hazard, including its physicochemical properties and bioavailability; whether there are multiple and/or interacting hazards (e.g. from various sources); the routes of exposure; and projected patterns of exposure in at-risk populations; coupled with: (ii) *defining the progression to health end-points:* that is, the probability that the geological processes or agents are likely to produce any adverse health effects. This may involve defining the relationship between toxicant "dose" and occurrence of a particular disease, and calculation of the potential rates of disease in the given population.

This process of evaluation has been separated into a number of discrete stages (Brownson and Petitti 1998; National Research Council (U.S.). Committee on the Institutional Means for Assessment of Risks to Public Health 1983). These are summarised in Fig. 2.3 (enHealth Council and Australian Department of Health and Ageing 2002).

The first component of the risk assessment is **Issue identification**. This involves identifying and confirming a role of

the risk assessment approach for a particular issue: that is, Why is this issue of particular concern? Is the issue urgent? How do all the stakeholders (often including the community) perceive the issue? At this early stage, it is essential to consider whether the issue is amendable to the risk assessment approach and to determine the social context (for many environmental issues, there may be a high level of anger, anxiety, and impatience).

Evaluation of risk from complex industrial processes or contaminated sites is often a challenging task because of the diversity of compounds produced and multiplicity of potential exposure contexts and health impacts. In mine tailings, for example, a mixture of pollutants may be present, including: trace elements such as arsenic, mercury, lead, and cadmium; radioactive materials; acid-producing sulphide compounds; fine-grained minerals such as asbestos and quartz; hydrocarbons, often introduced by mining and processing equipment; as well as additives such as cyanide (including leaching agents, sodium cyanide and hydrogen cyanide) (Nriagu 2011).

The next stage is **Hazard assessment**, which – as applied to a medical geology context – often involves investigating a hazardous process or the inherent properties of hazardous agents. This process of assessment requires two major activities: (i) *Hazard identification: Are the agents of concern presumed or known to cause adverse health effects?* The disciplines of toxicology (including animal or *in vitro* studies) and epidemiology (studies of human populations) are used to help address this question. For hazardous chemicals, it is often necessary to collect and compare relevant data on physical state, volatility and mobility as well as potential for degradation, bioaccumulation and toxicity; (ii) *Dose-response assessment: What is the relationship between the dose and occurrence of the adverse health effects? At what dose does this health effect occur?* Estimations of human health risks from exposure to specific chemicals are generally based on extrapolations of the results of toxicological experiments on animals. These extrapolations provide standard human "dose-response" relationships for the chemicals. The validity of the data and the weight-of-evidence of various toxicity data must be assessed.

For example, the International Agency for Research on Cancer (IARC) grades hazards according to whether they are likely to be carcinogenic. One common contaminant, arsenic, is classed as a Group 1 carcinogen based on evidence of its association with a spectrum of cancers (including those of the skin and liver, lung, kidney, and bladder) arising in human populations exposed to this contaminant (such as smelter workers, arsenical pesticide workers, patients treated with arsenic-containing medicinals, and communities who have ingested arsenic in drinking water across many geographical regions (Tchounwou et al. 2003)). However, arsenic is a widely distributed metalloid and occurs in different oxidation states and forms, including elemental arsenic, arsenides, sulphides, oxides, arsenates and arsenites (Centeno et al. 2006; Tchounwou et al. 2004). Therefore, *speciation* of arsenic (and many other compounds) is critical for accurate toxicological evaluation. The particular chemical form (i.e. oxidation/valency states) and physical attributes (morphological state) of the element often determine its toxicity, mobility of an element in the environment, and bioavailability (Refer to Chaps. 27 and 12)

The next component of the risk assessment process is **Exposure assessment,** which seeks to define the exposure pathways that are currently experienced or anticipated under different conditions: that is, *Under what circumstances or conditions could people be exposed to the hazard? What forms of assessment or estimation will be used to determine whether exposure had occurred?* This requires assessment of the following parameters: the frequency, extent, and duration of the exposure; the locations in which the exposure is likely to occur; the exposed populations; the pathways of exposure and actual or predicted intake of contaminants. Depending on the agent and exposure context, intake usually occurs through ingestion (such as incidental ingestion of contaminated soil or dust, or via contaminated food or drinking water), inhalation, or dermal absorption. It must be noted, however, that compounds (such as metals) detected in the environment are not necessarily biologically available to humans because of large particle size, low solubility, limited release from soil, or entrainment in surrounding rocks and soils. Exposure can be estimated *directly* (such as with biological testing or personal monitoring of each individual, including urine, blood, hair and nail samples) and/or *indirectly* through environmental monitoring, questionnaires and diaries.

Risk characterisation [also called **risk estimation**] integrates the preceding analyses of hazard and exposure measurements in order to decide: *What is the estimated probability that adverse outcomes will occur in this particular population at the specific level of exposure?* This assessment on the likelihood of harm is then used to guide the process of decision-making (Stern et al. 1996).

Throughout the risk assessment process, it is important to conduct ongoing **evaluation of limitations and uncertainties.** There are always likely to be elements of uncertainty in the analyses and these need to be explicitly acknowledged. These include: gaps in information about the profile of the hazards at a particular site; inadequate exposure information, including problems defining the population affected or the geographical area involved; limits in the availability and consistency of toxicology data, including extrapolations of findings (e.g. from animal studies) to human populations; limitations in the use of point estimates in the present day when trying to infer past exposures over a long duration. For many toxic agents, there may be

no specific models to define the relationship between contaminant levels or durations of exposure and the probability of disease outcomes.

The conclusions from the risk assessment process are then used to inform the final stage: **Risk management**. This involves evaluating possible actions and alternatives to minimise risk, taking into account all the relevant factors: the risk assessment, as well as the practicality (cost and technological constraints), social and political implications of the available options. A "zero risk" option is usually not achievable, given that almost any action potentially entails some degree of risk. The strategies are then implemented and monitored to ensure that they have been effective. Risk management also involves deciding upon the process and form of engagement and communication with relevant stakeholders, including the community. This process is often influenced by the degree of outrage experienced by public, the intensity of media attention, and attribution of blame (e.g. whether the problem was caused by a natural event or by human error) (Covello et al. 1989).

2.6 Monitoring and Surveillance of Disease

Many communities and countries have a formal set of arrangements to prevent, monitor and respond to communicable and non-communicable diseases. These "surveillance systems" aim to provide *"a continuous and systematic process of collection, analysis, interpretation, and dissemination of descriptive information for monitoring health problems*" (Rothman et al. 2008). Usually the surveillance systems are operated by public health officials to assist in disease prevention and to guide control.

In *active* surveillance, the health agency actively initiates information gathering, such as by regular calls or visits to doctors or hospitals. In contrast, for *passive* surveillance, the health agency does not actively contact reporters – it leaves the contact to others (such as treating doctors). Often both active and passive options exist in the same system. For example, the surveillance organisation may actively contact large representative hospitals while leaving smaller centres to passively participate.

In the process of *notifiable disease reporting*, health workers are (often legally) required to contact a central public health agency when a particular disease is identified. Examples include infectious diseases with serious health implications, such as whooping cough, cholera or rabies. The time requirements vary: some diseases require rapid notification; with others, there is less urgency. Other systems for obtaining health data include: *laboratory-based surveillance,* in which summaries (e.g. of microbiological samples) are provided by a laboratory; *registries*, which are designed to capture all occurrences of a disease or group of diseases (such as cancers) in a specified area; and *surveys*, which involves direct contact with a community, or a sample population, to define an outcome of interest, such as the presence of disease, levels of risk behaviour, or use of health services (Silva and International Agency for Research on Cancer 1999). Diagnostic tests form an important basis for identifying those with the disease or outcome of interest. It is imperative to consider the accuracy and availability of the tests, how they are used, and whether they are being reliably interpreted.

The populations to which surveillance is applied may be defined in narrow terms (such as a hospital) or in broad terms (e.g. the global population, as conducted by the World Health Organisation in tracking the emergence of new influenza strains). Surveillance systems need to maximise confidentiality: it is an ethical requirement and is required to engender community trust in the system. Usually security measures are put in place to ensure that no violations of privacy can occur.

Surveillance systems have different objectives. They may be to provide rapid feedback (e.g. in infectious disease outbreaks) or may be used for longer term health care planning or monitoring of prevention programs. For example, surveillance data can be used to identify whether changes in disease incidence are occurring, such as by comparing them to historical records. Declining trends can follow the pattern of disease or the effectiveness of control measures (e.g. infection incidence after the introduction of an immunisation campaign).

Examples of surveillance systems with particular applications to medical geology include those designed to assess the health consequences of lead exposure. The presence of lead in the environment was, and in many locations remains, a major environmental hazard, with sources including leaded gasoline, lead-based paint, other household items containing lead (e.g. ceramics, toys), and contamination from mining, smelting and other industrial processes (e.g. battery manufacture). The presence of lead and its capacity for ingestion by children – such as in paint, dust or contaminated soils – has major implications for health, particularly in relation to neurodevelopmental effects. Elevated lead has been linked to behavioural disturbances, delayed learning, and diminished intellectual capacity in children. Other effects of lead exposure in children and adults include renal damage, anaemia and a range of other toxic effects on the reproductive, neurological and cardiovascular systems.

In the United States, population-based programs have been established to evaluate young children for signs of lead poisoning, primarily through blood tests and clinical evaluations (United States. Dept. of Health and Human Services. Committee to Coordinate Environmental Health

and Related Programs. Subcommittee on Risk Management and Centers for Disease Control (U.S.) 1991). This public health issue has been addressed through various forms of surveillance, including: (i) at the State level: reviews of blood lead level (BLL) analyses from laboratories, which may also incorporate information gathered directly by public health officials or clinicians (e.g. evidence of lead toxicity – such as developmental and behavioural disorders – and/or risk factors for lead exposure in screened children); (ii) at a national level: the Centres for Disease Control's National Center for Environmental Health is responsible for developing and maintaining the national surveillance system for childhood lead levels. In practical terms, this task involves centralising the data (mainly laboratory blood lead tests, with personal details removed to ensure confidentiality) from participating agencies across the United States. This aggregated information is then collated and analysed to identify those regions and communities at highest risk, to target interventions, and to track the progress of programs that aim to reduce lead exposure (Centers for Disease Control and Prevention CDC's National Surveillance Data (1997–2008)).

2.7 Disease Prevention and Control

Preventive health care is a term used to describe a range of both *technical* and *educative strategies* applied by doctors, nurses, allied health professionals and public health workers in community and clinical settings. These interventions are designed to prevent the onset of disease or to slow or stop the progress of illness, and may be applied at different stages in the natural history of disease.

These strategies are commonly categorised in terms of *primary, secondary* and *tertiary prevention*. *Primary prevention* seeks to prevent the occurrence of disease altogether (that is, it aims to reduce disease *incidence*). It often focuses on strategies to control hazards (e.g. infectious agents; chemicals) and to modify risk factors in the population through health promotion and health education (such as by reducing smoking or encouraging greater participation in aerobic exercise).

With respect to minimising exposure to hazardous chemicals, primary prevention usually involves a combination of education and legislation. For example, exposure to asbestos and natural asbestiform compounds has major implications for health, including asbestosis (diffuse fibrosis of the lung), pleural lesions and various cancers. Contact with asbestos and asbestiform compounds may be controlled through the following strategies (United States Environmental Protection Agency 2012):

- Government legislation to prevent extraction and use of asbestos in building or for other industrial purposes
- Public and worker education
- Establishment of exposure limits (e.g. maximum allowable concentrations)
- Strict controls on removal and replacement of existing asbestos materials (e.g. insulation)
- Use of protective equipment and warning signs

Many other geological materials are also monitored and regulated in order to limit the degree of public and occupational exposure. These include metals and metalloids (such as arsenic, lead, mercury, cadmium, chromium, and beryllium), silica dusts, and radionuclides.

Secondary prevention is designed to intervene early in the course of a disease to halt or slow its progress, and thereby stop or reduce its clinical manifestations. Secondary prevention of both communicable and non-communicable diseases is closely related to the concept of disease *screening*, which is the identification of preclinical disease (usually by a relatively simple test). Screening is the evaluation of people who are apparently healthy (that is, without clinically overt symptoms) "to detect unrecognized disease [or its early biological manifestations] in order that measures can be taken that will prevent or delay the development of disease or improve the prognosis" (Last and International Epidemiological Association 1995). In practice, secondary prevention involves the *interaction of community health measures* (mass screening campaigns, central registers e.g. for cytology, mammography) and *clinical medicine*, such as through family doctors who provide clinical screening services (such as cervical smears).

In general, the screening procedure is not *in itself* designed to simply diagnose the end-stage illness. It is assumed that screening will detect a disease at an earlier stage than would have occurred otherwise, and will thereby offer the potential for improved prognosis (a greater chance of survival). However, it is also important to emphasise that screening alone (that is, simply achieving early detection) is not enough to constitute prevention. It is the combination of screening and the subsequent application of an effective early intervention that comprises secondary prevention.

Tertiary prevention differs from the primary and secondary approaches in that it is applied after the diagnosis of the disease. It aims to intervene later in the course of disease so as to reduce the number and impact of complications which add to the patient's disability and suffering. Tertiary prevention may be difficult to distinguish from treatment because it forms an integral part of the clinical care plan for patients with established disease. The basis for the distinction is that a *preventive intervention* is one that is applied before a potential problem occurs, such as debilitating complications (e.g. eye or kidney damage from diabetes mellitus), whereas *treatment* is applied to alleviate a problem that has actually occurred. The different stages of prevention are summarised in Fig. 2.4a**,** with an example provided using exposure to asbestos-induced cancer (Fig. 2.4b) (Das et al. 2007; Tiitola et al. 2002; Wagner 1997).

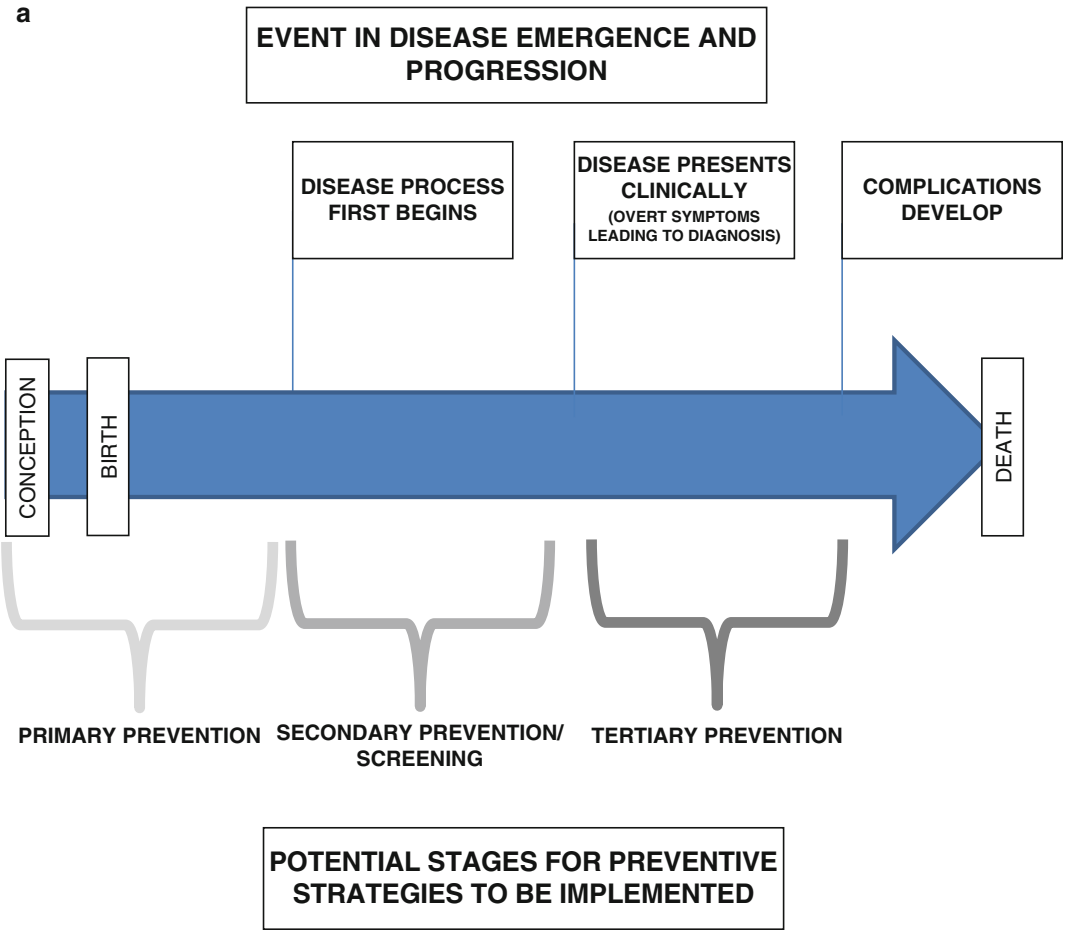


Fig. 2.4 (**a**) Overview of preventive strategies in relation to disease progression. (**b**) Illustration of preventive strategies as applied to an asbestos-induced cancer

In the field of preventive medicine, a distinction is often possible between two competing strategies for prevention: those oriented towards the *population* versus those oriented towards *high-risk groups or individuals*. In other words, there is debate over whether such initiatives should seek to address the needs of "sick individuals" or "sick populations" (Rose 1985)? In population strategies, efforts at prevention are directed *en masse* at an entire community (or significant portions of the community). Examples include the provision of general nutritional advice to consume more fruit and vegetables; screening of biochemical disorders at birth; health checks in women, men or older age groups. In contrast, a high risk strategy targets specific diseases or predisposing conditions in individuals known to be at higher risk of developing the condition. For example, lung function tests and imaging may be used to detect lung disease (such as pneumoconioses) in those who have worked in the industries with high levels of mineral dusts.

The distinction between population and high risk strategies is illustrated by considering possible public health strategies for the control of elevated blood lipids ("high cholesterol") (Rose 1993). A general population approach would tend to use regulatory, educative and structural strategies across the community (such as through advice on nutrition and exercise through the media) to reduce the overall levels of blood lipids in the entire population. In comparison, a "high-risk" approach might involve screening for high levels of blood lipids (such as through the individual's family doctor) followed by specific dietary or medication regimes in those identified as being at high-risk. Depending upon the nature of the risk factor and the disease, a combination of both population-based and high-risk strategies may be implemented.

In practical terms, a large number of people exposed to a small degree of risk (e.g. moderately elevated cholesterol or moderately excessive bodyweight) may generate many more cases than a small number of people with high risk (Rose 1993). Therefore it may be necessary to focus on modifying behaviour in the large mass of people with slightly elevated risk in order to have a major impact on the overall rate of

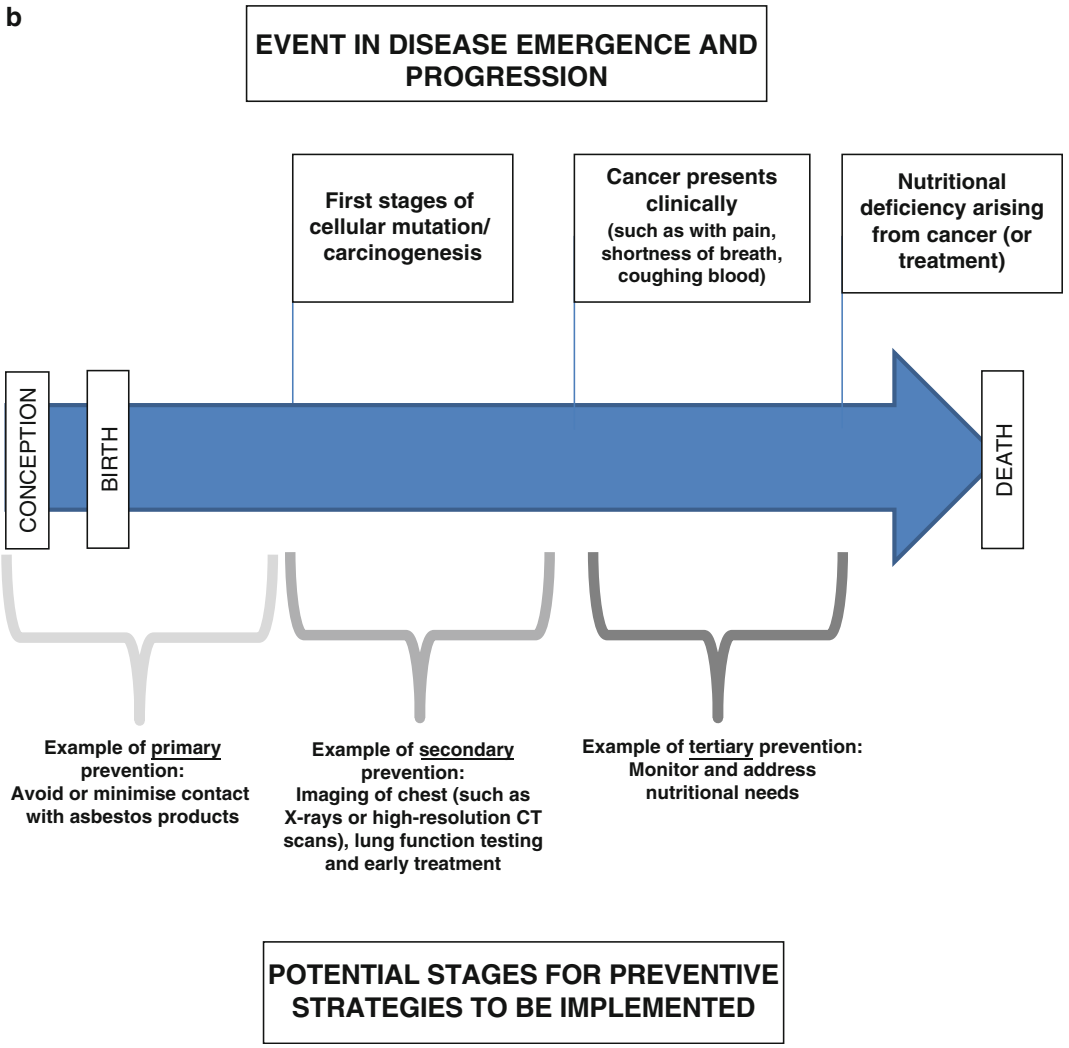


Fig. 2.4 (continued)

disease. In general, any preventive strategy should seek to provide the greatest benefit to the largest number of people, while also attempting to minimise the chance of causing inadvertent harm to the same population.

2.8 Research Methods in Public Health

Research into public health issues often employs the approaches and techniques of *epidemiology*. Epidemiology is broadly defined as the study of disease patterns in populations. Epidemiological analyses often seek to discover the causes, determinants or risk factors of a disease, usually in order to make prevention possible. The factors that may impact on disease include: hazardous agents such as micro-organisms or toxic chemicals, trace elements and minerals; lifestyle factors; and genetic influences. The term *environmental epidemiology* is often used to apply factors in the environment that impact on disease, ranging from pollutants to infectious agents. Many of the principles of environmental epidemiology also have implications for studies used in the field of medical geology, and are described in Chap. 23.

There are three underlying questions common to all epidemiological studies. The first question is: *Who is to be sampled?* The study population must contain a proportion of people who either have the disease of interest or are potentially at risk of developing the disease, and should be representative of any broader populations to which the study results will be applied. The study population must be of sufficient size – and, by implication, a sufficient number of the participants must experience the exposure and disease events of interest – to achieve the required statistical power in the analysis. For example, an analysis of the relationship between inhalation of beryllium dust and lung cancer that drew all its participants from the general population would

be inefficient, because high levels of exposure to this compound in the community is relatively uncommon. Other practical issues that must be considered include the likely degree of cooperation from the study population, accessibility of the participants for enrolment, maintenance of confidentiality of the data, and the overall cost of recruitment, evaluation and follow-up of the participants.

The next question is: *How is data collection to proceed?* This involves considering the optimal approach to capturing information on the *risk factor or exposure of interest* and the *subsequent disease events*. Many methods may be used to estimate risk factors and exposure, including the use of interviews, questionnaires or diaries for each individual (Nieuwenhuijsen 2003; White et al. 2008). In some situations, physical or chemical measurements may be used at the point of contact, such as with personal monitoring (two well-known examples include dust monitors used by mining personnel and personal dosimeters used to record ionising radiation exposure). Bioindicators or biomarkers are measurements from body tissue, fluid or excretion products to obtain data on past exposure to chemicals or other agents. Exposure may also be inferred through use of physical, chemical and biological measurements of the environment, including soil, air, food, water and indicator organisms. For disease events, the process of ascertainment may include self-reports of illness, medical records, laboratory results, or the study itself may involve testing procedures in the protocol (Silva and International Agency for Research on Cancer 1999).

Lastly, *What sort of analysis and evaluation is planned*? This involves a thorough interrogation of the data: Were the sampling procedures conducted successfully? How strong is the evidence of an association between the exposure and the disease? How much of the relationship between exposure and disease might be influenced by random variation, bias (systematic distortion of results) or confounding (mixing of effects)? How generalisable are our data to other populations?

Epidemiological results are often presented in terms of *measures of effect*. In general, these provide an estimate of the magnitude of association between the *risk factor or exposure* and *the disease* (or outcome of interest). A commonly used measure of effect is the *relative risk (RR)*, which estimates how many times more likely it is for "exposed" persons to develop the disease relative to "non-exposed" persons. An example of how these relative measures are applied is provided by Baris and Grandjean (2006) in their prospective study of mesothelioma mortality in Turkish villages exposed to fibrous zeolite. (In certain villages on the Anatolian plateau in Turkey, inhabitants are exposed to erionite, a form of fibrous zeolite, which is present in the volcanic tuffs that are used as building stone. Mesothelioma is an aggressive form of cancer linked to exposure to certain fibrous minerals.) In this study, the residents from a selection of villages in the area were followed up for the period 1979–2003, and the mortality rates from mesothelioma were calculated. For one of the analyses, adults from villages with high levels of erionite exposure were compared with a non-exposed "general population": in this case, the population of another country (Denmark) was used as the referent. The authors reported that the mortality rate from pleural mesothelioma (which affect the membranous linings the lungs and chest cavity) of highest-risk villages *relative to* the general population was 485 (a measure called the *standardised mortality ratio* for pleural mesothelioma). In other words, the "exposed" population (that is, residents from the high-risk villages) were 485 times more likely to develop pleural mesotheliomas than a "non-exposed" ("general") population.

Epidemiological studies are often classified into different groups, depending upon their purpose and design. These include *descriptive studies*, *analytic studies*, and *intervention/experimental studies*. A brief outline only is provided here, and readers seeking more detail are encouraged to refer to Chap. 23 or a general textbook on epidemiological principles (such as (Farmer and Lawrenson 2004; Gordis 2009; Rothman et al. 2008; Szklo and Nieto 2007)).

Descriptive studies often focus on the "time, place and person" component of epidemiology. In other words, such studies seek to determine: *Who* is at particular risk of this disease? *When* did they get the disease? *Where* are they located?

The "person" component of the analysis identifies the characteristics of the group at risk of the disease, such as their age, gender, ethnicity, occupation, personal habits and presence of co-existing disease. The "place" component determines the spatial dimension of the disease events. Diseases may vary with the biological environment (e.g. local ecology may influence the presence of disease vectors, such as mosquitoes carrying malaria), the physical environment (e.g. climate; geology) and the sociocultural environment (e.g. sanitation and hygiene practices; cultural practices; nature of social interactions). The importance of place and of mapping diseases has been part of epidemiology since the discipline formally began: John Snow famously mapped cholera cases around the Broad Street water pump in London in 1854, thereby helping to confirm that the disease was water-borne (and not spread by miasmatic vapours, as had been widely assumed). The spatial patterns of disease can often provide supporting evidence for the underlying disease process, such as relating cases of thyroid cancer to the Chernobyl nuclear accident or the increased risk of kidney damage in areas with cadmium contamination in Japan.

The "time" component of descriptive studies defines how the frequency of the disease varies over a defined interval (or may seek to determine whether the disease frequency is different now than in the past). *Time series*

studies refer to analyses that follow the rates of disease in a given community or region through time. Time patterns in disease may be: *short-term*: temporary aberrations in the incidence of disease, such as outbreaks of gastroenteritis; *periodic*/*cyclical*: diseases with a recurring temporal pattern, such as the seasonal fluctuations in respiratory diseases, which are often more common in the winter months, or; *long-term*/*secular:* changes in the incidence of disease over a number of years or decades, such as the rise in HIV/AIDS incidence since the 1980s.

Different terms are used to capture variations in the disease pattern. An *epidemic* refers to a general rise in case numbers above the background rate of the disease. An *outbreak* usually refers to a localised epidemic, and is often applied to infectious diseases. Although a *cluster* also refers to a close grouping of disease (or related events) in space and/or time, it is usually applied to uncommon or specific diseases, such as birth defects or cancer.

An example of how this descriptive approach may be applied is provided by the outbreak of coccidioidomycosis in California in 1994 (Pappagianis et al. 1994). This infectious disease usually results from inhalation of spores of the dimorphic fungus *Coccidioides immitis*, which grows in topsoil. In their report, the study authors define the "time, place, person" components as follows: "[f]rom January 24 through March 15, 1994, 170 persons with laboratory evidence of acute coccidioidomycosis were identified in Ventura County, California." This number "substantially" exceeded the usual number of coccidioidomycosis cases seen in Ventura County. To account for this change in incidence, the authors noted that the "increase in cases follows the January 17 earthquake centered in Northridge (in adjacent Los Angeles County), which may have exposed Ventura County residents to increased levels of airborne dust" (Pappagianis et al. 1994). The temporal and geographical pattern were supported by subsequent environmental data indicating that significant volumes of dust been generated by landslides in the wake of the earthquake and aftershocks, and that this dust had been dispersed into nearby valleys by northeast winds (Schneider et al. 1997).

Another group of epidemiological studies are termed *analytic*. Traditionally, these address specific hypotheses using more formalised designs than the descriptive studies, and incorporate a well-defined comparison ("control") group. For example, in *case-control studies*, the population is first defined with reference to the presence or absence of disease (such as lung cancer). The individuals with the disease, known as the "cases", are compared to a non-diseased group (known as the "controls") with respect to some exposure history (such as their past employment in the asbestos industry). In contrast, for *cohort studies*, a study population is first defined with reference to their degree of exposure or presence of the risk factor (such as the number of years they have been employed in the asbestos industry). This population is "disease-free" at the start of the study period. The participants (known collectively as a "cohort") are then followed over time to determine who experiences the disease (such as lung cancer) and who does not.

This analytical approach is demonstrated in the prospective study of arsenic exposure from drinking water and skin lesions (skin thickening or changes in pigmentation) in Bangladesh by Argos et al. (2011). (The skin lesions are often precursors to later skin cancers.) This cohort study involved identifying a group of around 10,000 individuals who were free of skin lesions at the start of the study. The exposure variable was defined as differential use of wells in the study area as a source of drinking water. There were a total of 5,966 wells, encompassing a range of arsenic levels from the detection limit (0.1 μg/L) to >200 μg/L. The study participants were followed for the years 2000–2009 (inclusive) and assessed clinically to determine who developed skin lesions. The authors reported that for every quintile increase in the concentration of arsenic measured in the well water, there was a 31 % increase in the risk of developing skin lesions.

In environmental epidemiology, one of the main challenges in conducting analytic studies relates to the reliable estimation of exposure. Although a range of environmental sampling techniques are available, it may be difficult to infer personal exposure based on such measurements. In practical terms, samples are often collected over short periods of time that may not correspond with the – often prolonged – process of disease emergence in a population. The disease may only occur after a significant delay in time, often decades in the case of cancer. Often, little may be known about the concentrations of contaminants which produce epidemiological effects in human populations, or the time period between the exposure and an expected effect. Reliance on, or inferences based on, past measurements may also be problematic because of inaccuracies in the historical data.

The last group of epidemiological studies use *intervention-based (or experimental) designs*. Such studies compare a group of subjects who receive an intervention (the "treatment group") versus a group who receive another (or no) intervention (the "control group"). Often there is a process of random assignment of participants into these two groups. For example, a new cholesterol-lowering tablet may be compared to an alternative form of medication to determine whether there is significant difference on the rates of a subsequent cardiovascular event (such as a heart attack). Environmental epidemiological studies often do not use experimental interventions because it is difficult to manipulate environmental variables on a large scale, and the interventions may be unethical if the safety or wellbeing of either the "treatment" or "control" group is jeopardised.

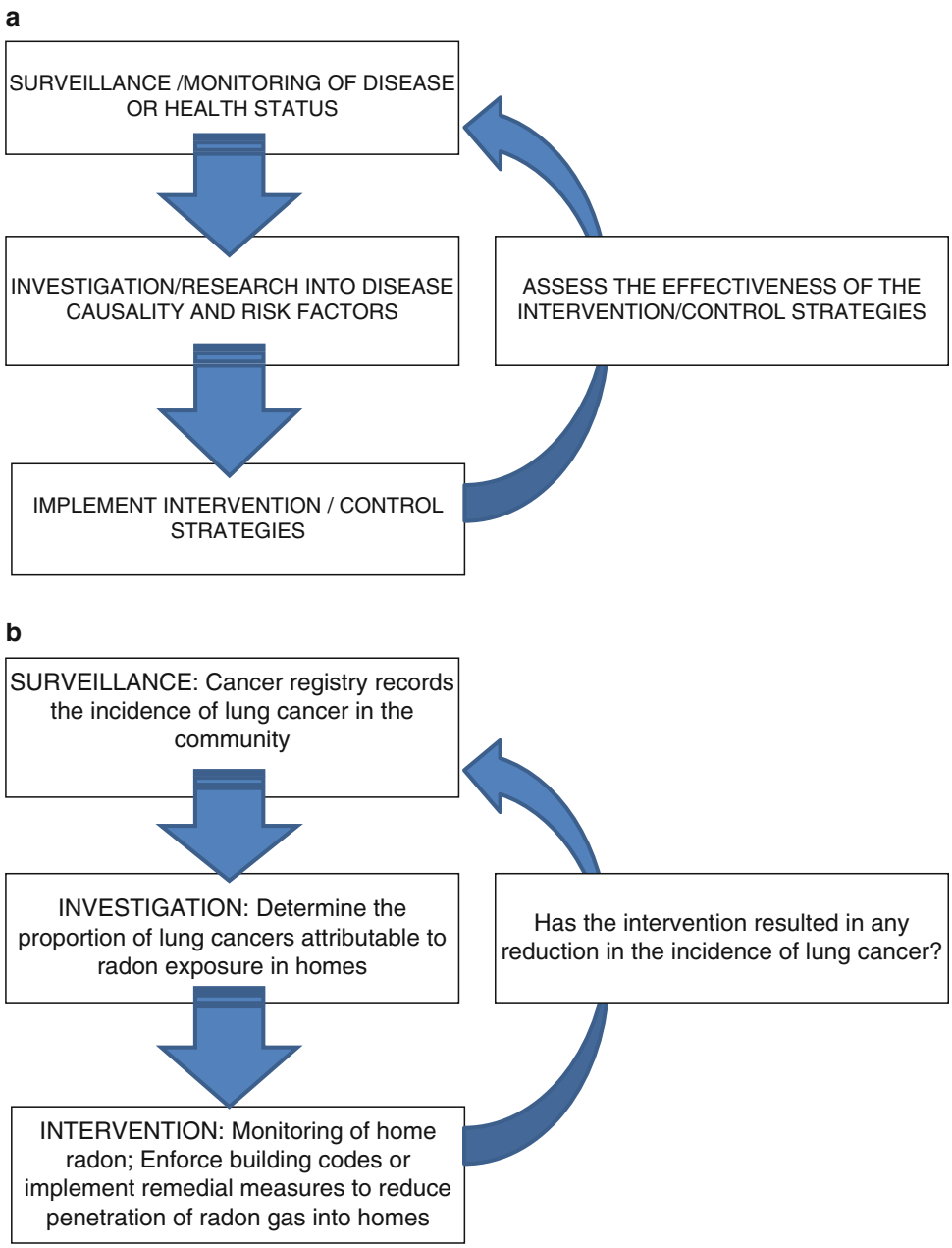


Fig. 2.5 (**a**) Schema of the main elements of a public health system. (**b**) Public health in a medical geology context: Radon exposure in the home and lung cancer

2.9 Integration and Application

In summary, geological and geochemical dynamics are integral to most of Earth's systems, and should be incorporated into existing descriptions and models of public health. The discipline of medical geology emphasises the fundamental degree to which geological processes are interconnected with the wellbeing of human communities. In many situations, patterns of disease emergence must directly or indirectly take account of geologically-driven determinants. Maintenance of wellbeing in human populations relies upon effective monitoring and management of the geosphere, especially given that anthropogenic interventions can accelerate the nature and pace of change. An integrative framework is required to evaluate, investigate and manage potential risks to communities, and must incorporate knowledge from both earth sciences and health sciences.

The core elements of the public health system can be conceptualised in Fig. 2.5a.

As noted, there are many applications of medical geology that have direct relevance to the field of public health. An integrative example will be provided here, emphasising the major themes and inter-disciplinary links covered in this chapter.

A major risk to human populations is the radioactive gas, radon, which has been linked to lung cancer. Radon-222 is a naturally occurring decay product of uranium-238 which is

commonly found in soils and rocks. Radon-222 progeny, particularly polonium-218, lead-214, and bismuth-214, are of health importance because they can be inspired and retained in the lung. Public health agencies rank residential radon-222 exposure as one of the leading causes of lung cancer after tobacco smoking. An example of how the main components of a public health system relate to the issue of radon exposure in the home and lung cancer is provided in Fig. 2.5b.

See Also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements) • Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 19 (Natural Aerosolic Mineral Dusts and Human Health) • Chapter 24 (Environmental Medicine) • Chapter 25 (Environmental Pathology).

References

Argos M et al (2011) A prospective study of arsenic exposure from drinking water and incidence of skin lesions in Bangladesh. Am J Epidemiol 174(2):185–194

Baris YI, Grandjean P (2006) Prospective study of mesothelioma mortality in Turkish villages with exposure to fibrous zeolite. J Natl Cancer Inst 98(6):414–417

Berkman LF, Kawachi Ī (2000) Social epidemiology. Oxford University Press, New York

Black RE, Mills A, Merson MH (2006) International public health: diseases, programs, systems, and policies, 2nd edn. Jones and Bartlett, Sudbury

Brownson RC, Petitti DB (1998) Applied epidemiology: theory to practice. Oxford University Press, New York

Centeno JA (2006) Environmental pathology and health effects of arsenic poisoning: a critical review. In: Ravi N, Euan S, Gary O, Prosun B, Peter N (eds) Managing arsenic in the environment: from soil to human health. CSIRO Publishing: Collingwood, Victoria, Australia, pp 311–327

Centers for Disease Control and Prevention CDC's National Surveillance Data (1997–2008) Available from: http://www.cdc.gov/nceh/lead/data/national.htm. Accessed 17 Jan 2012

Chadwick E, Great Britain. Poor Law Commissioners (1842) Report to Her Majesty's Principal Secretary of State for the Home Department from the Poor Law Commissioners on an Inquiry into the sanitary condition of the labouring population of Great Britain, with appendices. H.M.S.O, London

Covello VT et al (1989) Effective risk communication: the role and responsibility of government and nongovernment organizations, Contemporary issues in risk analysis. Plenum Press, New York

Das M et al (2007) Asbestos Surveillance Program Aachen (ASPA): initial results from baseline screening for lung cancer in asbestos-exposed high-risk individuals using low-dose multidetector-row CT. Eur Radiol 17(5):1193–1199

De Benoist B, World Health Organization, Nutrition for Health and Development (2004) Iodine status worldwide: WHO global database on iodine deficiency. Department of Nutrition for Health and Development, World Health Organization, Geneva

Dos Santos Silva I, International Agency for Research on Cancer (1999) Cancer epidemiology: principles and methods. International Agency for Research on Cancer, Lyon

enHealth Council, Australian Department of Health and Ageing (2002) Environmental health risk assessment: guidelines for assessing human health risks from environmental hazards. Department of Health and Ageing and EnHealth, Canberra

Farmer R, Lawrenson R (2004) Lecture notes on epidemiology and public health medicine, 5th edn. Blackwell Science, Oxford

Gordis L (2009) Epidemiology, 4th edn. Saunders/Elsevier, Philadelphia

Great Britain Working Group on Inequalities in Health, Black, Douglas, Great Britain Department of Health and Social Security (1980) Inequalities in health: report. DHSS, London

Great Britain Department of Health and Social Security, Committee of Inquiry into the Future Development of the Public Health Function, Great Britain Department of Health and Social Security (1988) Public health in England: the report of the Committee of Inquiry into the Future Development of the Public Health Function. HM Stationery Office, London

Green LW, Kreuter MW (1990) Health promotion as a public health strategy for the 1990s. Annu Rev Public Health 11:319–334

Green J, Tones K (1999) Towards a secure evidence base for health promotion. J Public Health Med 21(2):133–139

Krech S, McNeill JR, Merchant C (2004) Encyclopedia of world environmental history, vol 3. Routledge, New York

Last JM, International Epidemiological Association (1995) A dictionary of epidemiology, 3rd edn. Oxford University Press, New York

Lopez AD, Disease Control Priorities Project (2006) Global burden of disease and risk factors. Oxford University Press/World Bank, New York/Washington, DC

Marmot M (2005) Social determinants of health inequalities. Lancet 365(9464):1099–1104

Marmot M, Feeney A (1997) General explanations for social inequalities in health. IARC Sci Publ 138:207–228

Mason JB, International Development Research Centre (Canada), Micronutrient Initiative (Organization) (2001) The micronutrient report: current progress and trends in the control of vitamin A, iron, and iodine deficiencies. International Development Research Centre, Ottawa

Mathers C et al (2009) Global health risks: mortality and burden of disease attributable to selected major risks. World Health Organization, Geneva

National Research Council (U.S.), Committee on the Institutional Means for Assessment of Risks to Public Health (1983) Risk assessment in the federal government: managing the process. National Academy Press, Washington, DC

Nef JU (1966) The rise of the British coal industry, vol 2, 1st edn. Cass, London

Nieuwenhuijsen MJ (2003) Exposure assessment in occupational and environmental epidemiology. Oxford medical publications/Oxford University Press, Oxford

Nriagu JO (2011) Encyclopedia of environmental health. Elsevier, Boston

Pappagianis D et al (1994) Coccidioidomycosis following the Northridge earthquake – California, 1994. JAMA 271(22):1735 (Reprinted from MMWR 43:194–195, 1994)

Porter R (1998) The greatest benefit to mankind: a medical history of humanity, 1st American edn. W. W. Norton, New York

Rose G (1985) Sick individuals and sick populations. Int J Epidemiol 14(1):32–38

Rose G (1993) The strategy of preventive medicine. Oxford University Press, Oxford/New York

Rosen G (1993) A history of public health, Expanded edn. Johns Hopkins University Press, Baltimore

Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia

Schneider E et al (1997) A coccidioidomycosis outbreak following the Northridge, Calif, earthquake. JAMA 277(11):904–908

Schulze MB, Hu FB (2005) Primary prevention of diabetes: what can be done and how much can be prevented? Annu Rev Public Health 26:445–467

Scutchfield FD, Keck CW (2009) Principles of public health practice, 3rd edn. Delmar, Cengage Learning, Clifton Park

Stern PC, Fineberg HV, National Research Council (U.S.). Committee on Risk Characterization (1996) Understanding risk: informing decisions in a democratic society. National Academy Press, Washington, DC

Szklo M, Nieto FJ (2007) Epidemiology: beyond the basics, 2nd edn. Jones and Bartlett Publishers, Sudbury

Tchounwou PB, Patlolla AK, Centeno JA (2003) Carcinogenic and systemic health effects associated with arsenic exposure – a critical review. Toxicol Pathol 31(6):575–588

Tchounwou PB, Patlolla AK, Centeno JA (2004) Health risk assessment and management of arsenic toxicity and carcinogenesis. Met Ions Biol Med 8(8):14–18

Tiitola M et al (2002) Computed tomography screening for lung cancer in asbestos-exposed workers. Lung Cancer 35(1):17–22

UN Millennium Project, Sachs J, United Nations Development Programme (2005) Investing in development: a practical plan to achieve the Millennium Development Goals: overview. Millennium Project, New York

United States Department of Health and Human Services, Committee to Coordinate Environmental Health and Related Programs, Subcommittee on Risk Management, Centers for Disease Control (U.S.) (1991) Strategic plan for the elimination of childhood lead poisoning. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Atlanta

United States Environmental Protection Agency (2012) *Asbestos*. Available from. http://www.epa.gov/asbestos/index.html. Accessed 17 Jan 2012

Wagner GR (1997) Asbestosis and silicosis. Lancet 349 (9061):1311–1315

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology: collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford/New York

Winslow CEA (1920) The untilled fields of public health. Health Service, New York County Chapter, The American Red Cross, New York

World Health Organization, (1981) Global strategy for health for all by the year 2000, Health for all series. World Health Organization, Geneva

World Health Organization (1998) The World Health Report 1998 – life in the 21st century: a vision for all. World Health Organization, Geneva

World Health Organization (2003) The World Health Report, 2003: shaping the future. World Health Organization, Geneva

World Health Organization (2005) Health and the millennium development goals. World Health Organization, Geneva

World Health Organization (2009) Global health risks: mortality and burden of disease attributable to selected major risks. World Health Organization, Geneva

World Health Organization, Division of Health Promotion (1986) Ottawa charter for health promotion. World Health Organization, Geneva (online text)

Yach D, Stuckler D, Brownell KD (2006) Epidemiologic and economic consequences of the global epidemics of obesity and diabetes. Nat Med 12(1):62–66

Part I

Environmental Biology

Ulf Lindh

Biology Education Centre, Research in Metal Biology,
Uppsala University, Uppsala, Sweden
e-mail: ulf.lindh@blixtmail.se

Introduction

Environmental biology may be characterized by interactions between geological and anthropogenic sources and life. Geological sources provide biological systems with major, minor, and trace elements. Elements present in soils are influenced by a variety of geological processes. If environmental conditions permit the elements to be available to plants, some will be taken up while others will be rejected. What is taken up becomes available to grazing animals and humans.

Anthropogenic sources provide both essential and nonessential elements. In some cases, elements do not have to be biologically available to present health problems. Some elements or compounds may impact the epithelial cells in the respiratory system merely by mechanical irritation and cause damage. Often, human activities may lead to the movement of elements from places where they reside outside of biological systems to places where their inherent chemical nature is realized.

Chapter 2 provides a comprehensive discussion of what is termed natural background. The chapter emphasizes and illustrates the importance of the biogeochemical cycle, which is intimately related to the concept of bioavailability. Numerous interpretations of bioavailability are presented.

During the past few decades, a number of environmental problems have been attributed, rightly or wrongly, to anthropogenic activities. A fundamental goal of medical geology is to provide a foundation for discussion, in which the anthropogenic sources can be distinguished from natural sources. Chapter 4 describes a variety of anthropogenic sources and reviews the known and potential hazards associated with them. In addition, current and future issues surrounding waste disposal are described, as are agricultural practices and transport of contaminants and the importance of maintaining potable water resources.

Chapter 5 reviews the chemistry of life, beginning with the unique properties of water. The chemical behavior of various elements within living cells is outlined. This chapter highlights the role of elements as chemical messengers and the requirements of multicellular organisms, while introducing two new designations, metallome and metabollome, concerning metals and nonmetals.

The biological mechanisms of element uptake into living organisms are described in Chap. 6. Following a review of some fundamental biochemical principles, this chapter highlights the uptake of iron, zinc, and copper as examples.

The essentiality of elements and particularly trace elements is often inadequately defined. Chapter 7 offers a working definition of major, minor, and trace elements. The biological functions of the major elements are reviewed, followed by an in-depth discussion of the minor elements (including calcium and magnesium) and an emphasis on the biological functions of trace elements.

The discussion in Chap. 8 takes a more physiological approach to geological sources of elements. This chapter reviews sources of essential elements and discusses their bioavailability.

Toxicity is inherent in all elements. For many essential elements, nutritional deficiency may be the common issue. Chapter 9 discusses the concepts of nutritional deficiency and toxicity, beginning with an introduction to biological responses. Various aspects of elements as toxins, and carcinogens are featured in this chapter.

Natural Distribution and Abundance of Elements 3

Robert G. Garrett

Contents

3.1 Natural Background

A widely recognized biological characteristic of a healthy and sustainable environment is diversity—as with biology, so with geology. Regions characterized by the presence of different bedrock units, and different surficial materials in areas affected by recent (geologically speaking) glaciation, develop varied landscapes that support differing ecosystems. Examples of varied landscapes range from Alpine and Cordilleran mountains, through gentler landscapes of rolling hills, to the glacial plains of Northern Europe and North America, or similarly from the high Himalayas, through verdant foothills, across fertile plains to the desert of Sind. In the parts of the world characterized by stable geological platforms, where mountain building has not taken place for many hundreds of millions of years and there have been long periods of landscape development, peneplains are the eventual outcome. Their topography is gentle without mountains. High relief areas are largely limited to inselbergs or ravines and river valleys where, due to crustal uplift, modern rivers and streams are cutting down into and eroding the old land surfaces. Examples are the Brasilian Shield, central Australia and parts of central Africa, though in the latter young volcanoes lead to local mountainous terrain. These are the physical expressions of the underlying geology, but there is another changing characteristic that cannot be seen directly—the chemistry of the underlying rocks and sediments and the soils that lie upon them.

It is the soils that either directly, or indirectly, sustain the vast majority of life on terrestrial parts of Planet Earth. The plants people eat (cereals and vegetables) or use (e.g., wood for construction, fibers for fabric and line, maize or sugar cane for ethanol production) grow in the soil. As with people, cattle, sheep and goats rely on the plants growing in the soil, and they, directly or indirectly, support human life through milk, wool, leather and flesh. Furthermore, soils interact with precipitation as it moves from surface to groundwater storage; soils are vital to sustaining life.

R.G. Garrett (✉)
Geological Survey of Canada, Natural Resources Canada, Ottawa, ON, Canada
e-mail: garrett@NRCan.gc.ca

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_3,

Soils have developed over very different time spans, from those on the peneplains of Africa, Australia, and South America that are hundreds of millions of years old, to soils developed over the last few decades on recent volcanic material and on freshly deposited silts from rivers that have overflowed their banks. Soils that have developed on glacial sediments are somewhat older. As the last ice retreated, about 8,000–12,000 years ago at the close of the Wisconsin (North America) and Weischelian or Würm (Northern Europe) Ice Ages, a bare landscape was exposed. What lay underfoot was poorly sorted glacial till, a mixture of eroded rock and sometimes previous soil, containing material from cobbles and large "rocks" down to finely ground mineral fragments. In places where glacial rivers had flowed under the ice, sinuous sand ridges called eskers were deposited. Where the rivers emerged from under the ice outwash fans were formed, and these became deltas when they flowed into glacial lakes. Sand dunes often formed near these glacial river outlets as well as back from lake shores as there was no vegetative cover to anchor the newly deposited sediments and save them from wind erosion as they dried. The soil cover had yet to form.

The soils that sustain life develop as an interaction between the solid rock or unconsolidated surface materials, the climate, and biological and other physical processes. Over time, vertical zonations called profiles develop as a function of the interaction of these processes (see also Plant et al. 2001, Figure 6). Many soil profiles are characterized by an organic, carbon-rich, black upper layer (the L, F, and A horizons); sometimes a sandy textured light-colored layer (the A_e horizon); commonly a brownish or reddish layer richer in iron and some other elements, organic matter, and minerals (the B horizon); and finally, the weathered soil parent material (the C horizon). Other characteristics develop where the soils are wet; in arid (desert) regions; or frozen in high northern and southern latitudes. In extreme northern and southern latitudes, polar deserts may form, and where there is sufficient moisture, permafrost may form. In the tropics, the upper organic-rich A horizons are often thin due to the rapid degradation of the leaf litter and other organic materials present; below these, iron-rich B horizons develop. In very old soils the B horizons may become cemented with iron oxides to form hard carapaces—variously named duricrust, ferricrete, or canga. One of the key outcomes of soil formation is that chemical elements commonly become vertically redistributed by the pedological (soil-forming) processes acting in the biogeochemical cycle. Within this major cycle many smaller cycles exist, such as that from soil to plant, back to soil, and soil to plant (see Sect. 3.5 for further discussion).

The range of element concentrations in an uncontaminated sample material is known as the background, and it is important to consider background as a range, not a single value (Reimann and Garrett 2005). Natural backgrounds characterize the chemistry of rocks and surface materials, including soils, river and lake sediments, and biological tissues. Differences in natural backgrounds arise due to landscape-forming processes, which in turn are influenced by diversity in the underlying geology. There is no one natural background level for any solid material in or on the Earth as the Earth is far too inhomogeneous (diverse). For there to be a single natural background for any substance, it would have to be homogeneously distributed throughout the planet, and this situation is only approached in the atmosphere, where the major weather systems of the Earth keep the atmosphere relatively well mixed in each hemisphere. Therefore, natural backgrounds are variable, and this chapter discusses and illustrates that reality.

Natural background concentrations of elements provide the pool of essential chemical elements required by biological processes; therefore, they are vitally important. Life on Planet Earth has developed in the presence of all the 97 naturally occurring elements of the periodic table. To varying extents biological processes employ these elements to fulfill specific biochemical tasks which ensure the continuation of life. However, in addition to essentiality there is toxicity (see also Chap. 9, this volume). A few elements, e.g., mercury, lead, and thallium, have no known essential role in sustaining life. In fact, at high levels in biota they may be toxic and cause dysfunction and eventually death. In this context, the case of mercury in fish is interesting. Although fish appear to be able to bioaccumulate mercury dominantly as highly toxic methyl-mercury species, without harm to themselves, the consumption of these fish by mammals leads to elevated mercury levels that can be cause for concern. Other elements, such as cadmium, are toxic at high levels in most animal life, but may be essential for metabolic processes that support life in some species (this is an area of current research). Still other elements appear benign, for example, bismuth and gold; the latter is even used for dental reconstruction. Finally, a great number of elements are bioessential at some level. Calcium is necessary for building bones and shells; and iron is important in blood in higher mammals and vanadium and copper for similar roles in marine biota. Other major and trace elements, e.g., sodium, potassium, magnesium, copper, nickel, cobalt, manganese, zinc, molybdenum, sulfur, selenium, iodine, chlorine, fluorine, and phosphorus are also essential for a variety of biotic processes. For most elements it is a question of balance, enough to ensure the needs of essentiality and good health, but not too much to cause toxicity. As Paracelsus stated 450 years ago and paraphrased to modern English: "The dose makes the poison." It is the imbalance between amounts available naturally and those needed to sustain a healthy existence that poses the issues of medical geology.

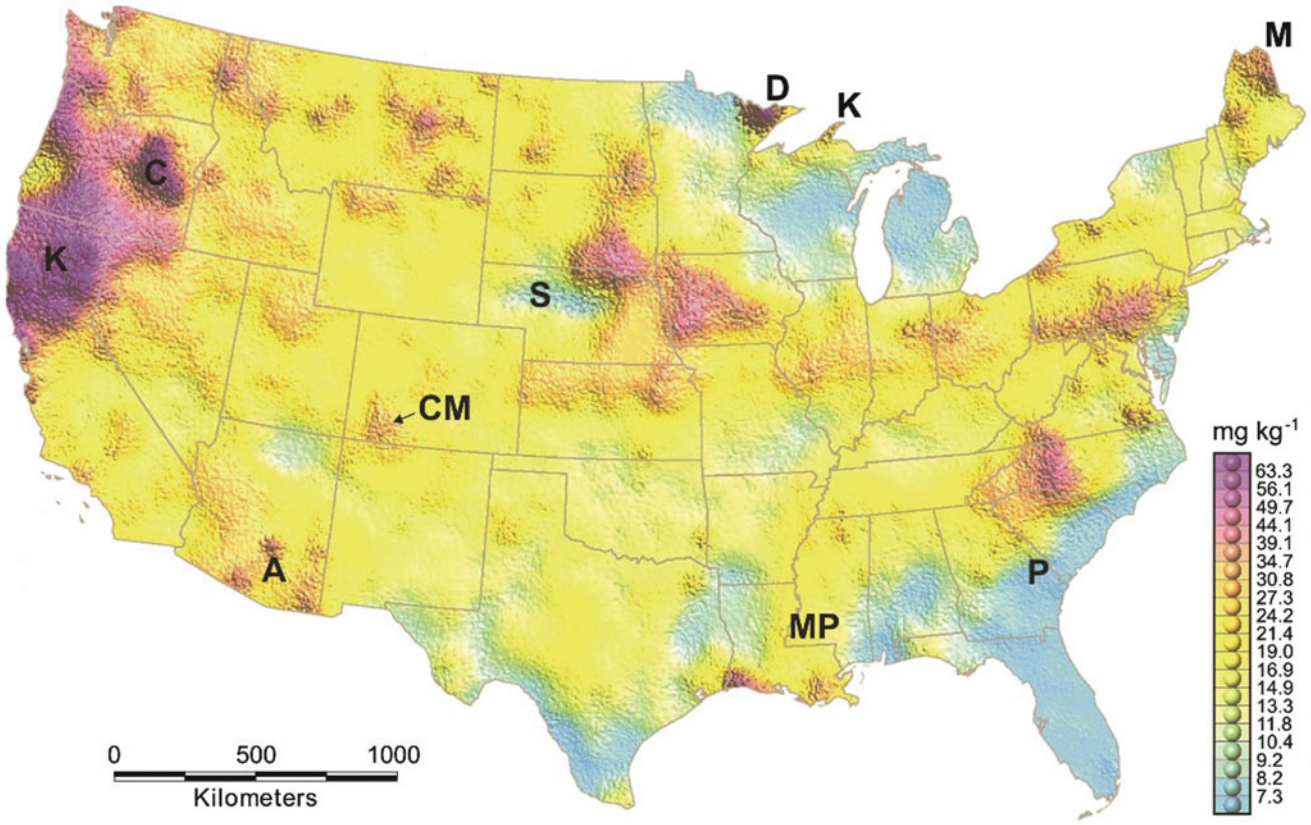


Fig. 3.1 Copper content (mg kg^{-1}) of soils in the conterminous United States (After Gustavsson et al. 2001 and Shacklette and Boerngen 1984. Reproduced with the permission of the United States Geological Survey)

3.2 A Chemically Variable Earth

An impressive way to demonstrate the chemical variability of the Earth's surface is with maps. Figure 3.1 displays the distribution of copper in the soils across the conterminous United States (7.84 million km^2), which is about 5.3% of the Earth's land surface. What is important to know when using such a map and data is how the soil samples were collected, processed, and analyzed. These are critical facts that influence the conclusions drawn from geochemical data. In this instance the soils, characterized as natural supporting native vegetation or agricultural, were collected from 20 cm below the surface at sites generally over 100 m from roads. The soils were dried, disaggregated, and the fraction that passed a 2 mm stainless steel sieve was pulverized and directly analyzed by optical emission spectroscopy. This method of analysis, which does not involve a chemical dissolution step, measures all of the copper present in the sample and is referred to as a "total" analysis. The samples used to prepare the map were collected by the U.S. Geological Survey between 1961 and 1975, and although over 25 years old, the map is still an excellent example of continental-scale variation (Shacklette and Boerngen 1984; Gustavsson et al. 2001).

One noticeable feature of these data is that they range from 7 to 63 mg kg^{-1}, which is almost an order of magnitude. However, in reality the individual 1,323 sample analyses ranged from <1 to 700 mg kg^{-1}, almost three orders of magnitude, yet they were all collected from uncontaminated natural background sites. The reduction in range in the scale from three to one order of magnitude is due to the smoothing process used to prepare the map (Gustavsson et al. 2001). Spatially, striking features are the high levels in the northwest versus the low levels in the southeast, and Fig. 3.1 is a graphic example of how natural background levels can vary spatially. The high copper background in the west is associated with the Columbia River basalts (C on Fig. 3.1) and basaltic volcanic rocks, e.g., the Klamath Mountains (K), in northern California and adjacent Oregon. The low copper background levels in the southeast in Florida and extending northward are associated with limestones in Florida and old beach sand Piedmont deposits (P) at the foot of the Appalachian mountains through Georgia and the Carolinas. Similarly low levels occur in the area of the Nebraska Sand Hills (S). Other notable features are high background levels in Minnesota associated with the Duluth gabbro (D), in Arizona (A) where major porphyry copper deposits have been mined, in southwestern Colorado associated with the Central Mineral Belt (CM), and in

northern Maine (M) adjacent to a mineral-rich region in adjoining New Brunswick, Canada. More subtle, but recognizable features are the locally elevated levels along the lower Mississippi River valley due to an abundance of overbank levee sediments deposited when the river overflowed its banks (MP). Similarly, the native (metallic) copper deposits of the Keweenaw Peninsula (K) are reflected by locally elevated values. Other areas of elevated or depressed background can be related to a variety of geological and pedological features. Although sites were avoided that may have been directly contaminated, some may have been influenced by airborne transport from local or major remote sources.

Clearly, to speak of a single range of background levels for copper in United States soils does the reality a great injustice. Instead, background levels need to be regional and reflect contiguous areas where the processes influencing them are similar. Secondly and most importantly, background levels are not single average values but are ranges reflecting the natural heterogeneity of the entity being characterized (Reimann and Garrett 2005).

However, having said this, average values are frequently published, cf. Wedepohl (1995), Reimann and de Caritat (1998), and Kabata-Pendias (2001), for different sample media such as rocks, soil, and waters. These are useful for establishing order of magnitude levels for the abundance of elements in various materials, and compilations such as Reimann and de Caritat provide a great amount of useful information. As a historical note, global averages are sometimes referred to as "Clarkes," after F.W. Clarke who was the chief chemist at the U.S. Geological Survey from 1884 to 1925. Clarke and Washington (1924) were the first persons to attempt to characterize geochemistry on a global scale with the publication of an average composition for igneous rocks based on a collection of 5,159 "superior" analyses.

Applied geochemists apply two descriptors to data in the context of background distributions: level and relief. Level is the central tendency of concentrations or measures of the amount of some property for a sampled unit. The unit could be rock, soils, waters, or vegetable matter or any discrete sample type for a specified geographic area. The central tendencies are most frequently expressed as a mean, geometric mean, or median. The median is widely used as it is unaffected by the occurrence of a high proportion of abnormally high or low values; it is robust in a statistical sense. Relief has no formal accepted numerical measure, but is an expression of the homogeneity of the data. Data with a small range, tight about the central tendency, are said to have low relief, whereas data with a large range, or exhibiting skewness, are said to have high relief. This term is also used in a spatial context, when a discrete area of a geochemical map is characterized by locally variable, "noisy" data, that area is said to have "high relief." High relief data are

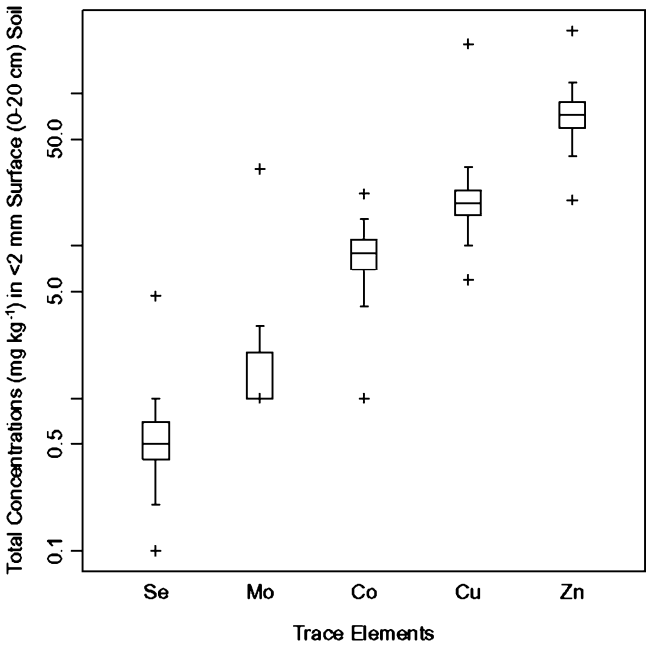


Fig. 3.2 Selenium, molybdenum, cobalt, copper, and zinc contents (mg kg^{-1}) of the <2 mm fraction of soils (N = 973) from the Canadian Prairies. These were chosen because of biological importance (Note: Crosses indicate maximum and minimum values, ends of whiskers are the 5th and 95th percentiles, the box is bounded by the 1st and 3rd quartiles, and the bar indicates the median. If a notch is present rather than a bar, the notch indicates the 95% confidence bounds around the median, which is at the narrowest point)

characteristics of areas of complex geology and/or multiple processes, and these could be associated with the formation of mineral occurrences, weathering, soil formation, etc. Statistical measures such as coefficient of variation (relative standard deviation), standard deviation, or median absolute deviation have been used to objectively quantify relief. The key issue is that background is not sufficiently characterized by a single number. Background is characterized by a range of values and some quantification is beneficial, whether it be quoting percentiles of the data distribution or the computation of some statistic. The advantage of quoting percentiles is that they involve no assumptions about the statistical distribution of the data. These data very often represent mixtures of several distributions related to different bedrock units or derived materials, together with different processes that have influence them, including the presence of mineral occurrences or weathering and pedological factors. The greater the diversity of an area described, the more likely it is that there are multiple data populations present.

It is common for trace element data from areas not characterized by the presence of anthropogenic contamination, mineral deposits, or a particularly diverse geology to span in excess of an order of magnitude. Figure 3.2 presents box-and-whisker plots for trace element concentrations in

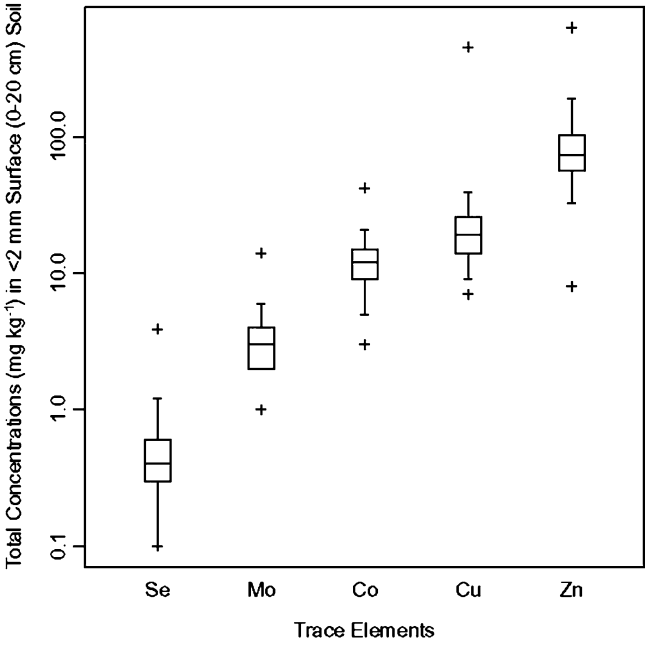


Fig. 3.3 Selenium, molybdenum, cobalt, copper, and zinc contents (mg kg^{-1}) of the <2 mm fraction of soils (N = 294) from southern Ontario

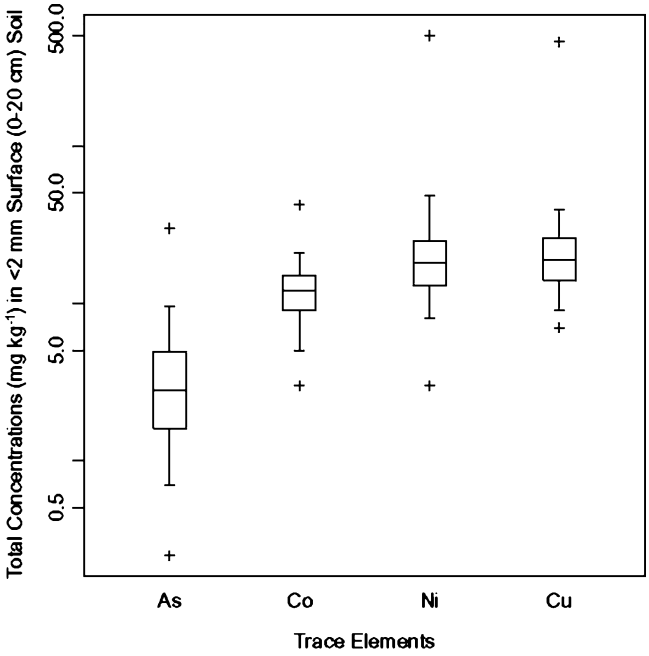


Fig. 3.4 Arsenic, cobalt, nickel, and copper contents (mg kg^{-1}) of the <2 mm fraction of soils (N = 294) from southern Ontario. These were chosen as they represent the Sudbury copper-nickel ore deposits

the <2 mm fraction of 973 surface (0–20 cm) soils collected from the Canadian Prairies in 1992. In Fig. 3.2, and those following, the box represents the middle 50% of the data and the bar the median of the data, alternately a notch indicating the 95% confidence interval on the median; the whiskers extend to the 2nd and 98th percentiles, and the minima and maxima are indicated by a '+'. These Prairie (grassland) soils developed on glacial sediments derived predominantly from the sedimentary rocks—i.e., limestones and dolomites (carbonates), shales and sandstones—of the Western Canadian Sedimentary Basin, and to a lesser extent from Canadian Shield rocks to the northeast and north, and material from the Rocky Mountains to the west. As shown in Fig. 3.2, the data span between one and one-and-a-half orders of magnitude. Most of the elements exhibit a positive skew, i.e., a greater abundance of higher values than lower values. The cobalt data are an exception and exhibit a negative skew. The element with the greatest skew is molybdenum. Cretaceous-age shales with abnormally high molybdenum levels occur in the region, and the distribution reflects this fact. The molybdenum data are characterized as having a greater relief than the other trace element data.

A soil survey using similar field sampling, sample preparation, and analytical protocols as the Prairie survey was undertaken in Ontario, Canada, in 1994. This survey included the Sudbury region which hosts some of the largest nickel deposits in the world. For comparison with the Prairie data, the same trace elements are plotted in Fig. 3.3. The differences between the mid-50% of the data are small—molybdenum and cobalt levels in surface soil are higher in Ontario, while selenium levels are lower. What are different are the ranges and skewness of the data, or their geochemical relief, which reflect the differences in geological diversity between the two survey areas. The range of the molybdenum data is greater in the Prairies, and the skew is greater reflecting the presence of the shales mentioned above. In the Ontario data, the ranges of the selenium, copper, and zinc data are larger which reflects the greater geological diversity (both older Shield rocks and younger Phanerozoic sediments) relative to the Canadian Prairies. Of particular significance is the increased positive skew or higher relief of the copper data, which reflects the presence of copper in the nickel deposits of the Sudbury basin.

The Sudbury ore deposits contain a wide range of metals present as sulfides and arsenides, and many are recovered commercially. Figure 3.4 presents the Ontario soil data for arsenic, cobalt, copper, and nickel. The deposits influence a small number of the survey sample sites, so the central parts of the distributions are not affected by their presence. Only the extremes for the major metals produced, i.e., nickel and copper are affected. Figure 3.5 shows the impact of the Sudbury basin, both as a geological and anthropogenic source (stack emissions), on what would be described as the background data distribution. The main part of the data spans one order of magnitude between 6 and 50 mg kg^{-1} nickel; two individuals (3 mg kg^{-1}) that fell below the detection limit of the analytical procedure (6 mg kg^{-1}) were arbitrarily set to half that limit; and most interestingly,

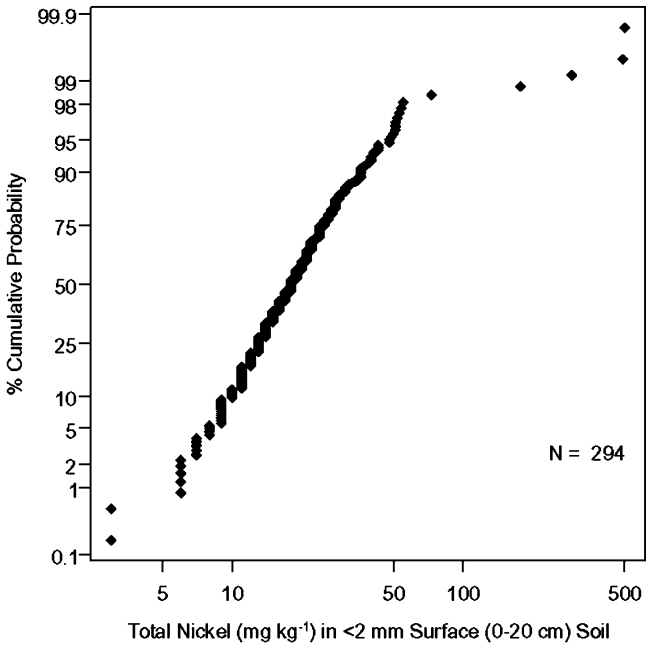


Fig. 3.5 Cumulative probability plot of nickel contents (mg kg^{-1}) of the <2 mm fraction of soils (N = 294) from southern Ontario, which demonstrate how mineral deposits are reflected by high outlying concentrations of an element in the ore

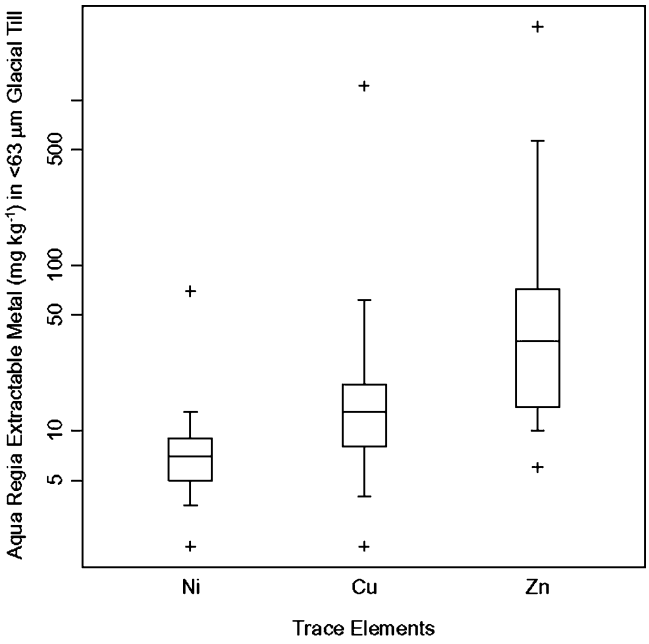


Fig. 3.6 Nickel, copper, and zinc contents (mg kg^{-1}) of the <63 µm fraction of glacial tills (N = 292) from the Nama Creek deposit, Manitouwadge, Ontario. This demonstrates the impact of a copper–zinc deposit on the data with anomalously high copper and zinc levels in till samples derived from the erosion of the ore deposit

a group of five individuals extend the range of the data a further order of magnitude to 500 mg kg^{-1} nickel. It is this latter group of samples that causes the data to have high relief and reflects the presence of the Sudbury basin, its mineral deposits, and its smelting facilities.

Background distributions may be influenced by naturally occurring high concentrations of trace elements and metals, sometimes referred to as "natural contamination", with the resultant data exhibiting a high relief. Natural processes that lead to the accumulation of elements at specific sites in the Earth's crust are what make them available to society for use. They raise concentrations to a level described as "ore", i.e., that which can be extracted from the ground at a profit (noting that there are many different economic and social models to define profit). Often in such regions there are also areas of unusually low trace element and metal concentrations called alteration zones, which reflect where natural processes have removed metals to transport and concentrate them elsewhere. When this occurs, the relief of data associated with ore elements can be very high. In general, "ore grades" exceed average crustal abundance levels by two to four orders of magnitude (McKelvey 1960), which results in natural ranges of trace element and metal concentrations of four or more orders of magnitude in areas characterized by mineral deposits. This depends on how much "ore-grade" material was incorporated in the samples collected. Two examples are provided below.

First, and to demonstrate that spatial scale has no effect on natural background levels per se, data for the nickel, copper, and zinc content of 292 glacial till samples collected from the walls of two adjacent trenches cutting across the Nama Creek copper–zinc deposit at Manitouwadge, Ontario, are presented in Fig. 3.6. These trenches, about 300 m long and up to 4 m deep, were dug and sampled prior to the development of the deposit. There are no nickel sulfides in the deposit, yet the nickel distribution spans one-and-a-half orders of magnitude, exhibiting low relief similar to the background distributions of most metals in the earlier examples. However, this is not the case for copper and zinc, which span over two-and-a-half orders of magnitude, exhibiting high relief, each due to the erosion of the mineral deposit and incorporation of ore-minerals containing these elements into the glacial till. The additional order of magnitude reflects the presence of the mineral deposit, a factor which increases the geological and geochemical diversity in the area.

The second example is a regional stream sediment survey of approximately 80,000 km^2 from Goias State in central Brazil. The area is extremely diverse geologically with a wide range of sedimentary, igneous, and metamorphic rocks present, and it is blanketed by residual soils that have developed *in situ*. Of importance to the example is that the rocks range from "metal-poor" limestones and sandstones and their metamorphic derivatives to ultramafic igneous

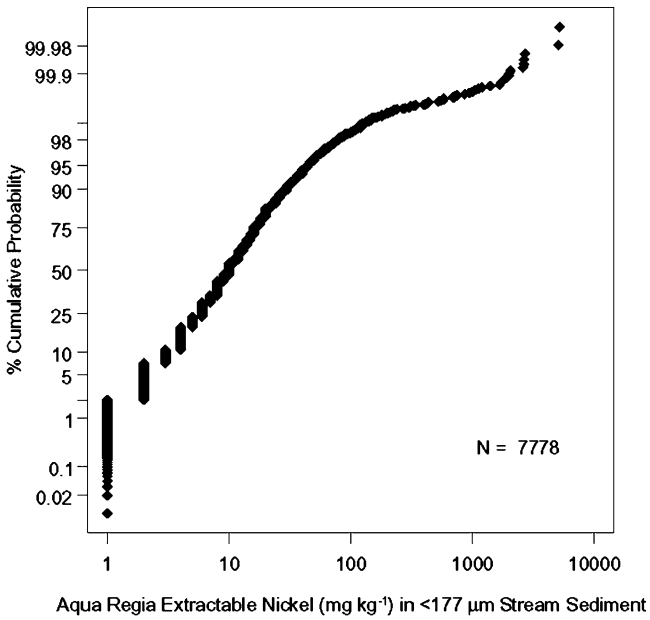


Fig. 3.7 Cumulative probability plot of nickel (mg kg^{-1}) in the <177 µm fraction of stream sediments (N = 7,778) from Goias State, Brazil. The flexure above 100 mg kg^{-1} reflects the presence of nickel-rich ultramafic rocks, and the outlying values above 1,500 mg kg^{-1} reflect the nickeliferous pyrrhotite mineral occurrences within them

intrusives containing nickel and copper mineral occurrences. The distribution of nickel in the <177 µm fraction of the stream sediments is presented in Fig. 3.7, where data span almost four orders of magnitude. These data are largely uninfluenced by anthropogenic activities, and therefore reflect natural processes. Four features are noteworthy: (1) 2% of these data were below the analytical detection limit of 2 mg kg^{-1}; (2) these and data up to 132 mg kg^{-1} (99% of all the data) reflect the variation of a wide group of different rock types and soil-forming processes active in the region; (3) the upper tail of the distribution reflects samples collected from areas underlain by nickel-rich ultramafic rocks in the Niquelândia – Barro Alto region; and (4) the highest levels, in excess of 1,000 mg kg^{-1}, reflect the presence of nickel sulfide occurrences within the ultramafic rocks.

Thus the controlling factor in determining the range of natural background is not the size of an area, but the diversity of the geology present. High diversity, due to some combination of contrasting rock types and/or the presence of mineral deposits, leads to geochemical data that are similarly diverse, i.e., they are characterized by high relief. That geochemical variability is independent of scale is widely demonstrated by the availability of national and continental scale geochemical maps; see, for example, Reimann et al. (2009, 2010). Another example of the presence of interesting patterns at widely different scales can be seen in Plate 3–1 of Darnley et al. (1995). This plate displays the copper stream sediment geochemistry of the island of St. Lucia, approximately 40 km^2, juxtaposed to the internal nickel chemistry of a grain of a platinum-bearing mineral 10 μm^2 in area as revealed by Electron Microprobe analysis. Although the two 'maps' are generally similar the difference in scale (area) is of the order of 10^{12}, yet well-designed sampling and analytical procedures at both scales reveal patterns of interest and geochemical significance.

Due to the diversity of geology and secondary environmental conditions, a vast number of regional and local background levels exist. This can be problematic when natural background distributions are used to establish national reference levels for regulatory purposes. For these to be effective they need to be very clearly defined as to what "environment" they represent, and they need to be based on adequately large sample sets. It is most important that data are not used out of the context of their collection. This issue can be exacerbated if measures of central tendency are used as national reference levels, because this immediately implies that approximately 50% of all measurements relevant to the reference level will fall above the quoted value.

To avoid this problem, reference levels associated with environmental regulations are sometimes quoted at some other level, for example, mean plus two standard deviations, which is notionally the 97.5th percentile of data for a normal Gaussian distribution derived from the estimated mean and standard deviation. In using this procedure an assumption has to be made, often implicitly with no discussion of the ramifications, as to the distribution of the data—normal, lognormal, or some other model. Various alternative procedures have been proposed, and several of these are compared in Reimann et al. (2005). Estimating background can be fraught with problems, especially as the "geographic units" from which reference levels are derived get larger and more geologically and environmentally diverse. In such cases these data are likely drawn from a number of different distributions and agglomerated into a "mixture". Often these mixtures appear to have lognormal distributions, despite the fact that many of the underlying components are more likely to have normal distributions (Vistelius 1960). An alternative is to use a percentile of the natural background distribution as a reference value. An example of such a procedure is the use of an Ontario Typical Range 98 (OTR^{98}) value by the Department of Environment of the government of the Province of Ontario, Canada, which corresponds to the 98th percentile of the background data for a specific entity, e.g., residential lands (OMOEE, 1999). Every distribution has a 98th percentile; natural processes may be the cause of higher observed levels, and anthropogenic contamination may result in levels lower than the 98th percentile. Acceptable numbers of false positives or negatives, i.e., type I and II statistical errors, are chosen during the selection of any particular percentile. However, the OTR^{98} level is used to trigger an investigation into whether the

exceedance is due to natural phenomena or is the result of anthropogenic contamination. If it is the latter, appropriate actions are taken on a site-specific basis.

3.3 Mineral Chemistry—The Key to the Distribution of Elements in Rocks

A natural question is: Why is there such diversity in the chemistry of surface materials? The answer lies in the composition of the individual minerals that compose rocks. Their properties are carried forward to other materials through erosional, weathering, and soil-forming processes, and are transferred to varying extents to waters that pass through these solid phase materials. In some respects the chemical diversity is self-fulfilling, as the main criteria that geologists use to "name" a rock, particularly in the field, are its mineralogy and texture (the shapes and interactions of and between the individual minerals). To have different names, rocks must be visibly different from each other.

The major components of the common, abundant rock types are silicates. The exceptions are rocks such as limestones and dolomites and their metamorphic derivatives (marbles), which are composed of calcium and magnesium carbonates. Other exceptions include sedimentary rocks containing phosphates and iron carbonates. Oxide, hydroxide, sulfide, and other minerals can also host the trace elements found in biological systems.

Silicates and aluminosilicates are important minerals in geochemistry, particularly ferromagnesian minerals, feldspars, and phyllosilicates (e.g., micas and clays). These minerals all contain silicon and sometimes aluminum as major components. When these components occur alone in a mineral they are present as ubiquitous quartz and the rare corundum, respectively. Corundum occurs as the gemstones ruby and sapphire where the colors are induced by trace amounts of chromium (ruby), iron, or titanium (sapphire). The ferromagnesian minerals and feldspars are important as they contain iron and magnesium, and calcium, sodium, and potassium, respectively, as major components. The presence of these elements establishes a situation where other physically similar elements may enter the lattices of the mineral crystals. It is this phenomenon that results in the wide range of trace element concentrations observed in rocks.

Minerals are rarely "pure" and are commonly contaminated with a wide range of other elements present at "trace" concentrations. Pure minerals are so rare, and often beautiful, that they are only seen in museums and mineral collections. The key to understanding which trace elements enter different mineral crystal forms, by a process known as substitution, is through knowledge of the physical properties of ionic radius (Fig. 3.8) and electronegativity (Fig. 3.9).

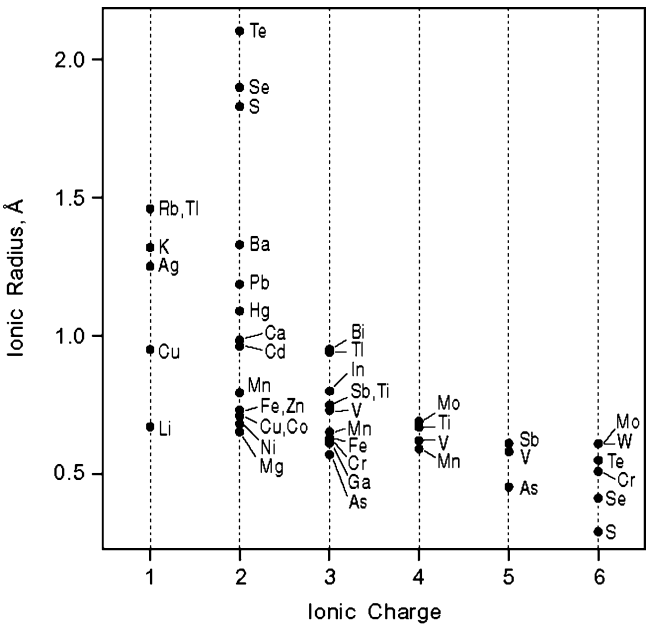


Fig. 3.8 The relationship of ionic radius to ionic charge (valence) for major and trace elements of mineralogical and geochemical interest

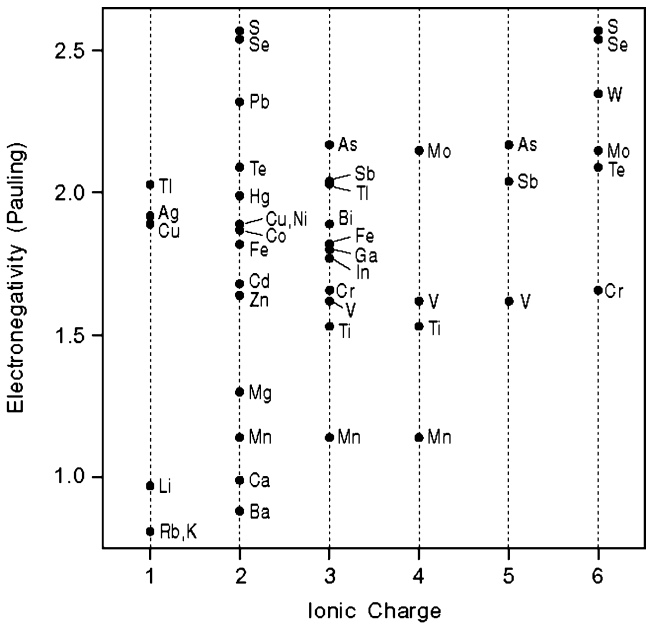


Fig. 3.9 The relationship of electronegativity to ionic charge (valence) for major and trace elements of mineralogical and geochemical interest

One example is the ferromagnesian mineral olivine, which is a major component of many ultramafic and mafic rocks and forms the essentially monomineralic rock dunite. Its composition is $(Mg,Fe)_2SiO_4$, which like many minerals is an intermediate form on a continuous solid solution of two other minerals—pure Mg_2SiO_4, forsterite, and pure Fe_2SiO_4, fayalite—which themselves are rare. The ionic radii of iron

Table 3.1 Classification of elements as lithophile, chalcophile, siderophile, or atmophile

Lithophile	Chalcophile	Siderophile	Atmophile
C, O, P, H, F, Cl, Br, I, Si, Al, Fe, Mg, Ca, Na, K, Ti, Sc, Cr, V, Mn, Th, U, Nb, Ta, Sn, W, Be, Li, Rb, Cs, Ba, Sr, B, Y, Zr, Hf, rare earths (REEs), Ga, (Cd), (Zn), (Pb), (Cu), (Ni), (Co), (Mo), (Tl)	S, Se, Te, As, Sb, Bi, Ag, In, Ge, Tl, Hg, Cd, Zn, Pb, Cu, Ni, Co, Mo, Re, (Fe), (Sn), (Au)	Pt, Ir, Os, Ru, Rh, Pd, Au, (Fe)	N, O, C (as CO_2), H, He, Rn, and other noble gases, (S as oxides), (Hg)

and magnesium in their two-valent states are 0.74 and 0.66 Å, respectively. Thus in the solid solution form, olivine, there are crystal lattice sites that can just as easily be occupied by other two-valent ions, such as nickel (0.69 Å), copper (0.72 Å), cobalt (0.72 Å), zinc (0.74 Å), and with a bit more crystal stretch, manganese (0.8 Å). The inclusion of manganese is facilitated by the fact that there is a further solid solution between olivine and monticellite, (Ca, $Mg)_2SiO_4$, and the ionic radius of two-valent manganese lies between that for calcium (0.99 Å) and magnesium (0.66 Å).

When iron occurs in the three-valent form it can be incorporated into garnet group minerals, $Fe_3Al_2(SiO_4)_3$, which are common accessory minerals in many metamorphic rocks, and to a lesser extent, igneous rocks. Garnet chemistry can also be expressed as $3R^2O.R^3{}_2O_3.3SiO_2$, where the superscript number indicates the valence and the letter R indicates a metal. In this latter form it can be seen that three-valent aluminum (0.51 Å), iron (0.64 Å), chromium (0.63 Å), and titanium (0.76 Å) can enter the garnet crystal lattice. Garnets are truly remarkable in this fashion, and they can host a wide range of divalent and trivalent metal ions. It is known that they occur widely at the base of the continental crust, because they come to the surface in rocks named eclogites entrained in certain volcanic extrusives, and they are believed to be the host and "reservoir" for many of the trace elements stored deep within the crust.

Similar examples can be provided for the other important ferromagnesian minerals, and their capacity to host trace elements by substitution is most easily understood when their formulae are expressed in the same way as for garnets. Thus, pyroxenes expressed as $R^2O.R^3{}_2O_3.\ SiO_2$ and amphiboles expressed as $R^2O.\ SiO_2$ can also contain aluminum and ferric iron and have very complex chemistries. They may also include hydroxyl groups and fluorine, epidotes expressed as $2R^2O.R^3OH.R^3{}_2O_3.3SiO_2$, and micas, e.g., biotite, expressed as $K_2O.3(Mg,Fe)O.3(Al,Si)O_2.(OH)_2$. These examples again can have very complex chemistries. Because of the abundant sites for divalent and trivalent metal ions in sixfold coordination, and the ability for small cations such as aluminum to replace silicon in fourfold coordination, these ferromagnesian minerals are hosts for a wide range of trace elements.

The aluminosilicate feldspars also play an important role as hosts for larger ionic radius metal ions. Feldspar chemistry lies between three end members: anorthite ($CaO.Al_2O_3.2SiO_2$), albite ($Na_2O.Al_2O_3.6SiO_2$), and orthoclase ($K_2O.Al_2O_3.6SiO_2$). The physical structure of feldspars consists of SiO_4 and AlO_4 tetrahedra (silicon and aluminum in fourfold coordination) in a three-dimensional network. This network is elastic and accommodates within its interstices not only the large positively charged cations, calcium (0.99 Å), sodium (0.97 Å), and potassium (1.33 Å), but also strontium (1.12 Å), lead (1.2 Å), barium (1.34 Å), rubidium (1.47 Å), and thallium (1.47 Å).

Phyllosilicates are an important group of minerals in both rocks and their weathering products. They include silicate (e.g., talc) and aluminosilicate minerals commonly known as micas (e.g., phlogopite) and clays (e.g., montmorillonite), and may contain sodium, potassium, calcium, iron, and magnesium. These minerals have a sheeted mineral structure with sheets of SiO_4 tetrahedra held together tightly by cations; and pairs of these sheets held together loosely by other cations. In the case of the mica biotite, iron and magnesium provide the tight bonds and potassium the looser bonds. Muscovite mica is similar, but the tight bonds are provided by aluminum. A wide variety of cations may replace the iron, magnesium, and aluminum in the tight binding sites, and other elements may substitute for the potassium that loosely binds the sheets together. The breakdown of silicates and aluminosilicates due to alteration or weathering leads to the formation of a range of clay minerals, which host a wide range of cations in addition to silicon and/or aluminum. The mica and clay minerals with the greatest ability to support substitution with metal cations are generally those that employ iron and magnesium at interlayer sites, e.g., montmorillonite. In contrast, kaolinite or gibbsite contains only aluminum and silicon or aluminum, respectively.

The elements discussed above are commonly called lithophile (rock loving) and are distinct from other elements referred to as chalcophile, siderophile, or atmophile (Table 3.1). The lithophile elements may occur as silicates, aluminosilicates, oxides, carbonates, sulfates, halides, phosphates, and vanadates, among other mineral forms in the natural environment.

An important second group are the chalcophile elements (Table 3.1), which are characterized by forming sulfides, arsenides, antimonides, selenides, and tellurides. It is these compounds that form the ore minerals that are the source of the nonferrous metals used by society. Some of these minerals, particularly iron sulfides such as pyrite and marcasite (FeS_2), pyrrhotite ($Fe_{(1-x)}S$), and the sulfarsenide

arsenopyrite (FeAsS), and to a lesser extent copper, zinc, lead, and molybdenum sulfides like chalcopyrite ($CuFeS_2$), sphalerite ($ZnFeS_2$), galena (PbS), and molybdenite (MoS_2), occur in many igneous and metamorphic and some sedimentary rocks. Due to the large amounts of these trace elements that can be held in sulfide and related minerals, it is not necessary to have abundant sulfides, etc., present in order to raise the levels of the chalcophile trace elements in rocks to quite high levels. As with silicate and aluminosilicate minerals, the chalcophile trace elements are often present as substitutions in commonly found minerals rather than in their own unique minerals. Again this is the result of fundamental physical properties, in this case electronegativity (Fig. 3.9). As examples, silver and mercury replace copper in many copper minerals, cadmium and indium replace zinc in sphalerite, selenium and tellurium replace sulfur, and arsenic and antimony occur interchangeably in others and with sulfur. In iron sulfides copper, cobalt, and nickel commonly substitute for iron. In igneous and metamorphic rocks sulfides exist as blebs and crystals along the boundaries between the majority silicate and aluminosilicate minerals. In addition to this, they may occur within the rock-forming minerals along fracture planes. This is the result of a process known as exsolution, which occurs as igneous and metamorphic rocks cool down and the individual rock-forming crystals are less able to accommodate incompatible components. The offending substances are then rejected to form discrete minerals along internal lines of crystal weakness or completely to discrete mineral grain boundaries.

Siderophile and atmophile elements are less important in the following discussion. The siderophile elements form alloys with iron and these are important sources of platinum group metals, together with gold, to society. The atmophile elements are ubiquitous in relatively hemispherically homogeneous atmospheres, with less exchange between northern and southern hemispheres that there is mixing within them. Mercury is the only metal that occurs as a gas at "normal temperatures and pressures," and this permits its transport over long distances independent of fluvial systems.

Although not strictly a crystalline mineral phenomenon, many elements are associated with organic matter in sedimentary rocks and metamorphic rocks derived from them. This is due to two general processes: (1) the ability of organic compounds to sequester and adsorb trace elements, e.g., copper and mercury; and (2) the actual formation of metallo-organic compounds to fulfill particular biochemical functions, such as copper and vanadium in the heme of marine invertebrates. In the geological context, organic matter is only preserved in rocks under anoxic conditions, which due to the prevalent redox conditions are also sulfur-reducing environments that lead to the presence of sulfides. This is particularly important in the formation of rocks described as "black shales" that can become enriched in many trace elements.

Table 3.2 Common geochemical associations

Group	Associations
Generally associated elements	K-Rb Ca-Sr Al-Ga Si-Ge Zr-Hf Nb-Ta Rare earths (REEs), La, Y Pt-Ru-Rh-Pd-Os-Ir Au-Ag
Plutonic rocks	
Generally associated elements	Si-Al-Fe-Mg-Ca-Na-K-Ti-Mn-Cr-V Zr-Hf-REEs-Th-U-Sr-Ba-P B-Be-Li-Sn-Ga-Nb-Ta-W-Halides
Specific associations	
Felsic igneous rocks	Si-K-Na
Alkaline igneous rocks	Al-Na-Zr-Ti-Nb-Ta-F-P-Ba-Sr-REEs
Mafic igneous rocks	Fe-Mg-Ti-V
Ultramafic igneous rocks	Mg-Fe-Cr-Ni-Co
Some pegmatites	Li-Be-B-Rb-Cs-REEs-Nb-Ta-U-Th
Some contact metasomatic deposits	Mo-W-Sn
Potassium feldspars	K-Rb-Ba-Pb
Many other potassium-rich minerals	K-Na-Rb-Cs-Tl
Ferromagnesian minerals	Fe-Mg-Mn-Ni-Co-Cu-Zn
Sedimentary rocks	
Fe-oxide rich	Fe-As-Co-Ni-Se
Mn-oxide rich	Mn-As-Ba-Co-Mo-Ni-V-Zn
Phosphatic limestones	P-F-U-Cd-Ag-Pb-Mo
Black shales	Al-As-Sb-Se-Mo-Zn-Cd-Ag-U-Au-Ni-V

After Rose et al. (1979)

As a result of these relationships, geochemists have observed consistent patterns in the distribution of many elements. Some of the more interesting of these elements are presented in Table 3.2. Several of the associations are related to mineral deposits, which are major natural sources of elements to the Earth's surface environment, and the processing of the ores can be major anthropogenic sources of contamination if appropriate emissions controls are not installed at processing plants and smelters.

In the secondary, surficial weathering, environment most of the ferromagnesian and aluminosilicate minerals are unstable and break down to more hydrated forms, e.g., phyllosilicates, oxides, and hydroxides and residual silica (quartz). During this process the trace elements held in the rocks are liberated: some are removed in solution as surface runoff or enter groundwater, and others are incorporated into new minerals or sequestered by organic matter. Two key mineral forms capable of retaining trace elements are the phyllosilicates, minerals such as smectites and chlorites, and the oxides and hydroxides of iron and manganese. The phyllosilicates sequester trace elements by two processes: by cation exchange to constant electrical-charge sites on the tabular surfaces of the minerals, and by adhering to the

broken edges of the clay particles where variable charge sites occur. Smectites are particularly effective in this role as they have large cation exchange capacities. The oxyhydroxides of iron and manganese formed during weathering are also effective in sequestering cations. This ability is enhanced when the oxyhydroxides are linked to humic or fulvic acids, which raises the charge on the oxyhydroxide surfaces. This effect is even more pronounced with the formation of humic colloids.

In contrast to the trace elements that are associated with minerals that break down in the weathering zone, those associated with resistate minerals that do not weather to any significant extent are retained in that mineral form in the soil and subsequent erosion products. Examples of resistate minerals are chromium in the mineral chromite; tin in cassiterite; niobium and tantalum in columbite-tantalite; zirconium and hafnium in zircons; and cerium, lanthanum, yttrium, and thorium in monazite.

3.4 Diversity in the Chemistry of Rocks

The combination of the chemistry of minerals and their abundances in different rock types, which are defined upon a mineralogical and textural basis, leads to a varied rock geochemistry. Table 3.3 provides examples of estimated average values for the trace elements in the Earth's crust and different rock types. These estimated averages give no indication of the actual range observed in these individual rock types; such an estimate has never been made on a global basis, but it is likely at least one or two orders of magnitude. The variability behind the Continental Crust estimates can be implied from the variability for the estimates of individual rock types. These range from less than an order of magnitude for mercury to almost three orders of magnitude for nickel.

To illustrate the variability associated with a single rock type, the example of black shales from Manitoba, Canada, is presented. These data come from 54 surface outcrop sites and drill holes in an area approximately 300 km wide (ENE–WSW) and 500 km long (NNW–SSE) along the eastern margin of the Western Sedimentary Basin. This area includes shales varying in age from Ordovician to Tertiary, spanning some 360 million years (Ma). Figure 3.10 displays the molybdenum data for the 476 samples subdivided by geologic period, oldest to the right and youngest to the left. The oldest (Ordovician, O) and youngest (Tertiary, T) black shales have lower molybdenum contents than the generally similar median-valued Devonian (D) to Cretaceous (K) shales. However, what is notable is the variability of the Cretaceous (K) black shales, which extend over three orders of magnitude.

The Cretaceous shales span approximately 85 Ma of deposition in a sea that went through various transgressive (deepening) and regressive (shallowing) stages. These 333 black shales have been subdivided by stratigraphic

Table 3.3 Compilation of average geochemical background data for the earth's crust and selected rock types

	Hg ($\mu g\ kg^{-1}$)	Pb ($mg\ kg^{-1}$)	Cd ($mg\ kg^{-1}$)	Cr ($mg\ kg^{-1}$)	Ni ($mg\ kg^{-1}$)	As ($mg\ kg^{-1}$)	Cu ($mg\ kg^{-1}$)	Zn ($mg\ kg^{-1}$)	Ref.
Earth's crust									
	80	13	0.2	100	75	2	55	70	Taylor (1964)
	90	12	0.2	110	89	2	63	94	Lee and Yao (1970)
Upper continental crust									
		20	0.1	35	20	1.5	25	71	McLennan (1992)
	80	13	0.2	77	61	1.7	50	81	Lee and Yao (1970)
Igneous rocks									
Ultramafic	4	1	0.1	1,600	2,000	1	10	50	Turekian and Wedepohl (1961)
Mafic	13	6	0.2	170	130	2	87	105	Turekian and Wedepohl (1961)
Intermediate	21	15	0.1	22	15	2	30	60	Turekian and Wedepohl (1961)
		10		55	30		60		McLennan (1992)
Felsic (4)	39	19	0.1	4	5	1	10	39	Turekian and Wedepohl (1961)
Sedimentary rocks									
Sandstone	57	14	0.02	120	3	1	15	16	Faust and Aly (1981)
Limestone	46	16	0.05	7	13	2	4	16	Faust and Aly (1981)
Shale	270	80	0.2	423	29	9	45	130	Faust and Aly (1981)
Black shale		15	4	18	68	22	50	189	Dunn (1990)
		100		700	300		200	1,500	Vine and Tourtelot (1970)

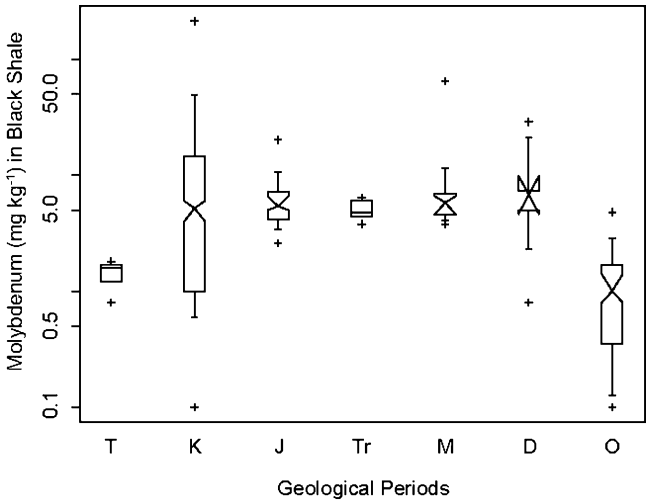


Fig. 3.10 Distribution of molybdenum (mg kg^{-1}) by age in black shales, Manitoba, Canada. From youngest to oldest: *T* Tertiary, *K* Cretaceous, *J* Jurassic, *Tr* Triassic, *M* Mississippian, *D* Devonian, *O* Ordovician

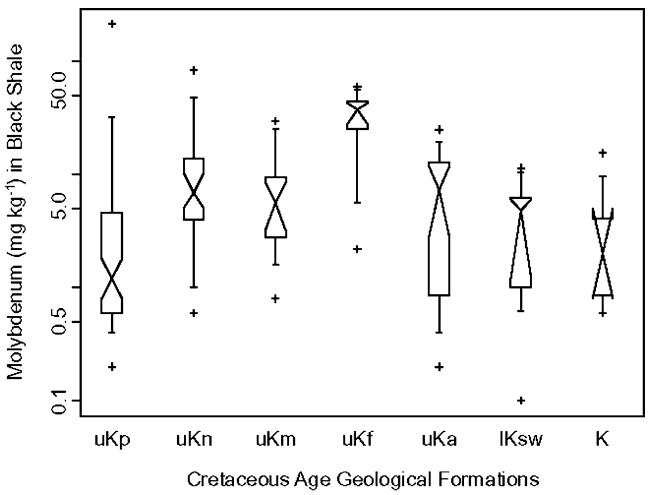


Fig. 3.12 Distribution of molybdenum (mg kg^{-1}) in Cretaceous age black shales, Manitoba, Canada, subdivided by formation. See Fig. 3.11 for legend

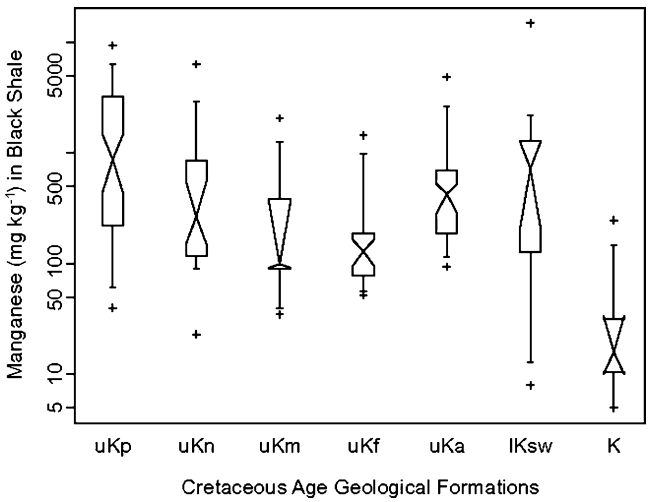


Fig. 3.11 Distribution of manganese (mg kg^{-1}) in Cretaceous age black shales, Manitoba, Canada, subdivided by formation. Formations from youngest to oldest, where u as a prefix indicates upper Cretaceous and l indicates lower Cretaceous: uKp Pierre, uKn Niobrara, uKm Morden, uKf Favel, uKa Ashville, lKsw Swan River, K undifferentiated, mostly lowermost Cretaceous, black shale

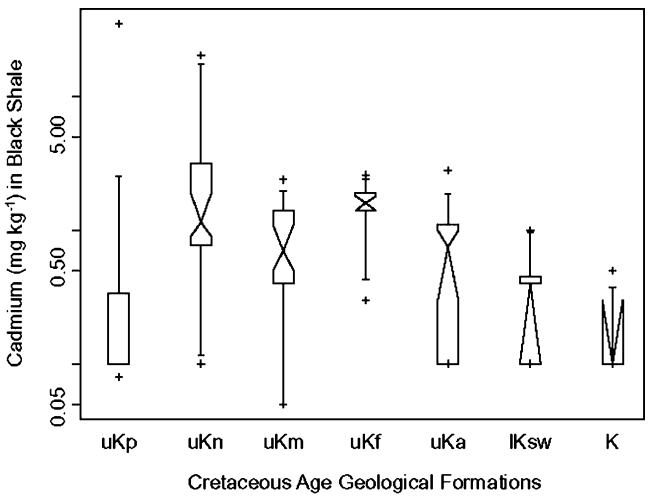


Fig. 3.13 Distribution of cadmium (mg kg^{-1}) in Cretaceous age black shales, Manitoba, Canada, subdivided by formation. See Fig. 3.11 for legend

formation (except for undivided Cretaceous rocks, K) in Figs. 3.11, 3.12, and 3.13. Figure 3.11 displays the by-age distribution of manganese (mg kg^{-1}), which forms a bowl shape with lowest manganese levels in the Favel (uKf) and Morden (uKm) Formations when the Western Interior Seaway was at it deepest. Manganese levels increase in older Ashville (uKa) and Swan River (lKsw) and younger Niobrara (uKn) and Pierre (uKp) shales, which were deposited in shallower water. These variations reflect the fact that in the deep anoxic carbonate-poor waters of the Western Interior Seaway manganese was preferentially retained in the seawater. Figure 3.12 illustrates the distribution of molybdenum (mg kg^{-1}) in shales. Levels are highest in the deep water shales of the Favel Formation (uKf) where anoxic sulfide- and organic-rich sediments favored the sequestration of molybdenum, and the median molybdenum level is almost two orders of magnitude higher than for the youngest Pierre shales (uKp). Some of the Cretaceous shales, particularly the Favel (uKf), Morden (uKm), and Niobrara (uKn), are enriched in cadmium (Fig. 3.13). At the surface, these shales were eroded from the Manitoba Escarpment and transported westward across the Canadian

Prairies about 400 km during the Wisconsin glaciation at the close of the last Ice Age, and they have contributed to higher levels of cadmium in the agricultural soils of the region.

In some instances, large segments of the Earth's crust may exhibit consistent patterns of trace element enrichment. These features are known as geochemical provinces. Rose et al. (1979) provide the definition of these geochemical provinces as "a relatively large segment of the Earth's crust in which the chemical composition is significantly different from the average." They go on to state: "One of the criteria of a bona fide geochemical province is that the characteristic chemical peculiarities should be recognizable in rocks covering a substantial period of time." Examples of geochemical provinces are the Bear uranium province in the northwestern part of the Canadian Shield; the central European uranium province that includes parts of Germany, Poland, and the Czech Republic; the great tin province that spans from eastern Australia, through Indonesia, western Malaysia, Thailand, and into China where rocks of many ages are enriched in tin; and the lesser known manganese province in northeastern North America including Maine and New Brunswick.

A second type of province, metallogenic, is of particular economic importance, and may have environmental and health consequences. In the context of risk assessment, the concept of metalloregions embodies both geochemical and metallogenic provinces (Fairbrother and McLaughlin 2002). Metallogenic provinces are regions of the Earth's crust that are characterized by the presence of mineral deposits and occurrences for particular metals of a particular mineral deposit type. In any mineral district, there are many more small mineral occurrences than economically exploitable mineral deposits. A distinction is made here between mineral occurrences and mineral deposits. Mineral deposits contain ore, i.e., that which can be extracted from the ground at a profit, using a sufficiently broad definition of profit to meet various societal needs; whereas mineral occurrences are similar geologically, but have no mining value. Metallogenic provinces may also be considered "mineral" provinces, as it is the presence of the metals at high concentrations in specific mineral forms that makes deposits exploitable. Examples are the copper deposits that occur in the western Americas extending from the southwestern United States through Central America to Peru, Bolivia, Chile, and Argentina; the prolific tin deposits coincident with the tin geochemical province of southeast Asia in Indonesia, western Malaysia, Thailand, and China; the Copper Belt deposits of Zambia and Zaire in central Africa; and the gold and base-metal deposits of the Abitibi Greenstone Belt in Quebec and Ontario, Canada.

The mineral forms for base- and precious-metals, apart from native (metallic) gold and platinum alloys, are sulfides, arsenides, antimonides, tellurides, and selenides that are far less stable in the surface weathering environment than silicates, aluminosilicates, oxides, etc., which host the metals in common rocks. As a result, metals in mineral occurrences and deposits weather more easily and are transferred to soils, sediments, and waters. From a mineral exploration viewpoint, their dispersion away from point sources facilitates exploration geochemical surveys. However, from a biological viewpoint this may be a good or bad outcome. If the metals are essential for life this may be beneficial; high metal levels may permit species that require such high levels, or are resistant to them, to flourish. If the metal species present is toxic, this will inhibit some life forms which may lead to an absence of biota or permit only certain hardy species to exist. It was well known that in the central African Copper Belt the high copper in the stream waters derived from the weathering of copper deposits was sufficient to be toxic to the snails that were critical in the *Bilharzia* cycle (Ali 2011). As a result, schistosomiasis was largely absent in the region. For those interested in the interaction of geology, geochemistry, and botany, readers are referred to Brooks (1972, 1998) and Dunn (2007).

The relationship between geochemical and metallogenic provinces is of some interest. By their nature geochemical provinces are low-entropy phenomena. In contrast, metallogenic provinces are high-entropy phenomena with extreme segregation of metals into spatially small discrete high concentration zones. Despite these contrasting characteristics, many metallogenic provinces lie within geochemical provinces, and it is assumed that the increases in regional background levels of the metals have provided part of the pool of metals that have been concentrated into the mineral deposits and occurrences. Reimann and Melezhik (2001) provided a discussion of the relationships between metallogenic and geochemical provinces in the context of a large (188,000 km^2), low sampling density (1 site per 300 km^2) regional geochemical study of surficial materials in Arctic Europe called the Kola Ecogeochemistry Project (Reimann et al. 1998). The authors noted that some significant metallogenic provinces were not recognized in the project they undertook, and also cast doubt on the value of the term "geochemical province." However, considering the very different nature of the two province types in terms of entropy, and the interaction of sampling density through search theory to the probability of recognition of the mineral occurrences of a metallogenic province with low-density field sampling programs (see Sect. 3.5 and Garrett 1983), the death of these provinces as useful concepts is premature.

It is this underlying variability in rock chemistry, that is in turn due to the mineralogical and compositional variability of the rocks composing the Earth's crust, as well as the mineral deposits and occurrences that occur in them, that causes the geochemical variability in natural background levels in the surficial materials and weathering products discussed in the first section of this chapter.

Although tables of averages (Table 3.3) are useful as general indicators of the element content of rocks, the variability behind them must never be forgotten. As noted and demonstrated above, that variability is considerable and can easily span up to three orders of magnitude for trace elements. Readers requiring average estimates for composition of both the Earth's crust, individual rock types, and other materials are referred to Wedepohl (1995), Reimann and de Caritat (1998), Kabata-Pendias (2001), and Rudnick and Gao (2003).

3.5 The Biogeochemical Cycle

The discussion in this chapter thus far has concerned the mineral kingdom. However, mineral-related processes only form one part of what is known as the biogeochemical cycle—the sum of the biotic and abiotic processes that move elements from rocks, to soils, sediments, and waters where they are incorporated into plants and animals and become parts of food chains. As these processes proceed, elements are returned to soils, sediments, and waters and, given sufficient geological time, are incorporated into various marine and terrestrial deposits that will be transformed to newly formed rocks. However, of more immediate interest is the small scale cycling that occurs at local levels.

Plants play a key role in the biogeochemical cycle; they are critical to soil formation in all but desert regions. The acids produced by their roots to liberate nutrients from the minerals in the soil contribute to the breakdown of those minerals. Another source of organic acids that contributes to mineral decomposition is the decay of litter-fall as plant material decomposes in the surface layers of soil. As these acids percolate downward they solubilize and carry many elements with them to lower levels in the soil. This process is vividly demonstrated in the podzolic soils (Spodosols) of humid temperate zones. Organic matter accumulates in the topmost layers of the soils (the L, F, H, and A_h horizons), giving them their characteristic dark brown-black color. Immediately below this there is a "sandy," colorless, leached (eluviated, A_e horizon) zone composed of mineral grains that have resisted corrosion and decomposition by downward percolating organic acids. Below the eluviated zone the soils are enriched in iron-oxyhydroxides, clay-sized materials, and to a lesser extent organic substances. Known as the B horizon, it is well developed in podzols and exhibits a rich red-brown coloration. Together, the humic-rich (A_h) top layer of the soil and the B horizon are sites of trace element, especially metal, accumulation due to the abundance of organic matter, smectites, clay minerals, and iron- and manganese-oxyhydroxides. The retention of trace elements in these horizons introduces a barrier in the biogeochemical cycle that halts circulation through the cycle for varying periods of time. Goldschmidt (1937) introduced the term "geochemical barrier" specifically for the retention of trace elements in humic-rich surface soil as it was a barrier to "flow" in the biogeochemical cycle. Once elements, metals, are sequestered and stabilized in organic compounds that are resistant to degradation, levels naturally increase. The rates of elemental flow around the biogeochemical cycle are varied, and when accumulation exceeds degradation compositional levels increase.

As a result of the biogeochemical cycle, trace elements are dispersed into different materials. Three examples are provided below, i.e., one for stream waters (Wales), one for organic stream plant material (Sweden), and one for terrestrial plant material and a foodstuff, hard red spring wheat (Canada). A fourth example of mineral rich stream sediment is provided in Sect. 3.5.

In recent years, the British Geological Survey has applied new ultrasensitive water analysis procedures in the preparation of hydrogeochemical maps. Figure 3.14, drawn from Simpson et al. (1996), is an example of the distribution of fluoride in stream waters from Wales and adjacent parts of England as revealed by a suite of 17,416 analyses. There is a clear boundary between high F^-waters draining dominantly Permo-Triassic rocks in the east and low F^- waters draining older Paleozoic rocks to the west. These two rock types were laid down in very different environments: the older rocks in a variety of marine environments, and the Permo-Triassic rocks in a terrestrial environment that was desertic to the east along the western margins of the Zechstein Sea and later along the northwestern margin of the Tethyan Ocean. What is important here is that the different environments that either favored the retention of F^-in terrestrial environments or its retention in seawater have survived in excess of 200 Ma to influence ground and surface water chemistry today and have epidemiological consequences (Simpson et al. 1996).

The Swedish Geological Survey is unique among geological surveys in employing organic stream material for its national−scale geochemical survey program (Fredén 1994). The reason for this choice of sample material, as distinct from the mineral-rich stream sediments commonly collected in other countries (see for example Fig. 3.17), is that it provides data that better estimate the bioavailable amounts of trace elements present in the environment. Aquatic plants such as mosses and sedge (*Carex* L.) roots are in equilibrium with the stream sediments and waters in which they grow, and their composition reflects the available amount of trace elements. Although stream water compositions may vary seasonally, the composition of the mosses and sedge roots varies more slowly, smoothing out temporal variations in the water chemistry. This makes the survey data particularly effective for monitoring anthropogenic impacts that result in the dispersion of trace elements into the surficial environment, and for detecting natural

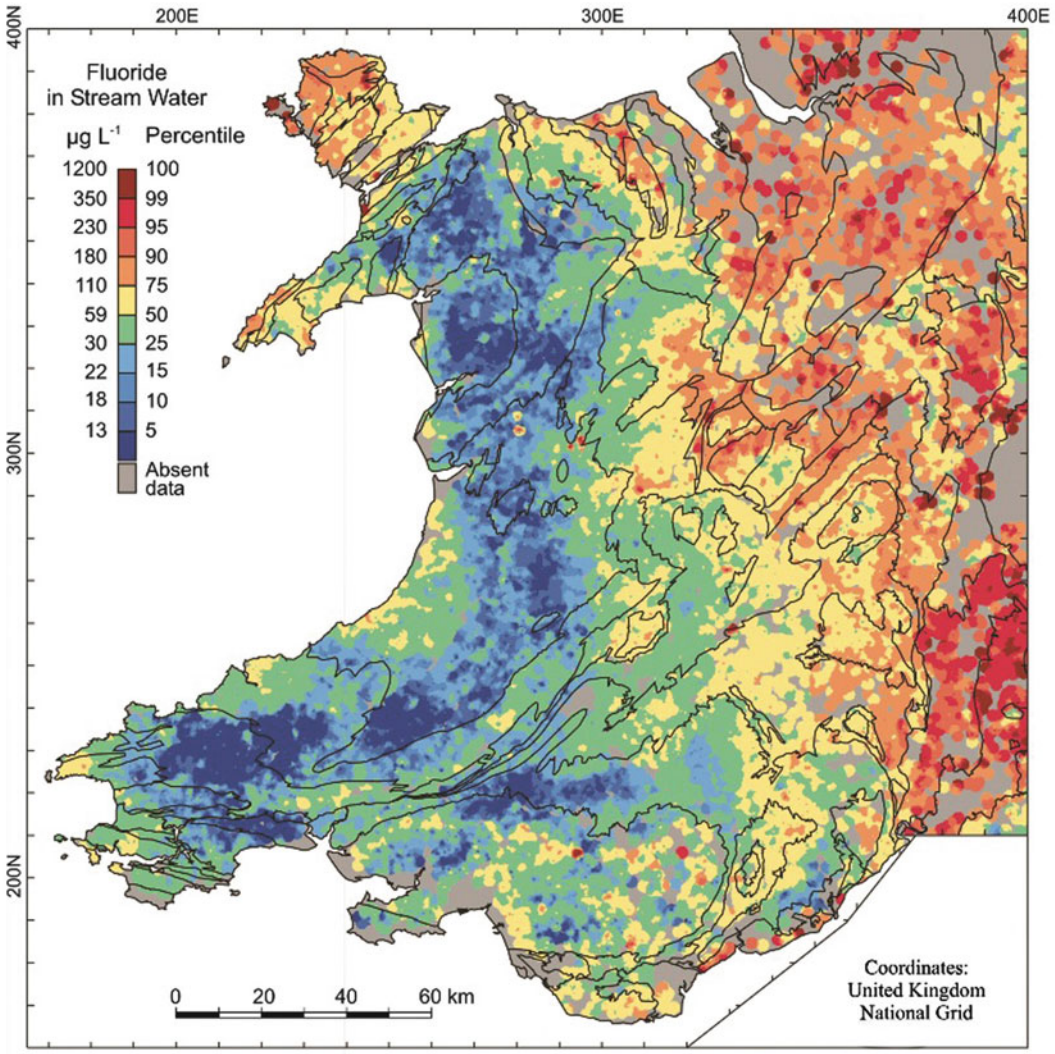


Fig. 3.14 Fluoride (µg/L^{-1}) in stream waters (N = 17,416) from Wales and adjacent parts of England (Reproduced with the permission of the British Geological Survey and Pergamon Press)

geological sources that may have an impact on the ecosystem. Figure 3.15 presents the stream plant chemistry for chromium in part of southern Sweden. The elevated levels in the western part of the map around Vänersborg and Trollhättan are associated with contamination from a smelter. Yet in the central part of the area near Nässjö equally high levels are observed which relate to the presence of dolerites, a mafic rock that is naturally enriched in chromium relative to most other rocks (see Table 3.3). To the east-southeast, elevated chromium levels occur due to the presence of small gabbro and ultramafic rock bodies. Another notable pattern occurs in a triangular area east of a line joining Norrköping, Linköping, and Västervik where elevated chromium levels characterize an area underlain by old basement granitoid, likely reflecting even older rocks that were incorporated into the granitoids by the processes of granitization. These chromium-elevated areas lie within an extensive area of low chromium rocks dominantly composed of orthogneisses and granites. Thus it can be seen that the organic stream sediment geochemistry reflects both recent anthropogenic processes and a variety of regional geological features. In this context, the high chromium patterns east of Linköping and Norrköping are of particular interest. These rocks are granitic, like so much of the area characterized by low chromium levels; however, a relict geochemistry reflecting earlier rocks is retained and indicates that these rocks are "different." This demonstrates one of the strengths of regional geochemistry. Although rocks may look the same, i.e., granitic, they may be geochemically different in their trace element composition, which demonstrates their different geological histories.

As a last example, the regional distribution of selenium in Canadian Prairie hard red spring wheat (*Triticum*

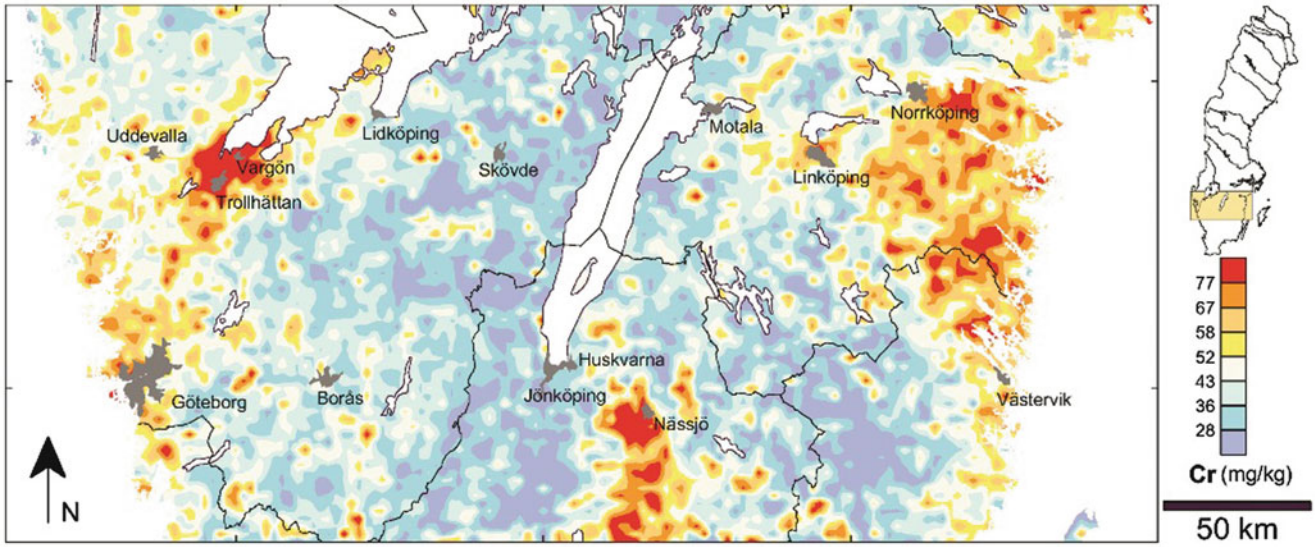


Fig. 3.15 Distribution of chromium (mg kg^{-1} in ash material) in organic stream sediments in southern Sweden (Reproduced with the permission of the Geological Survey of Sweden)

aestivem L.) is displayed in Fig. 3.16. These data plotted are averages of the selenium content of grain harvested from 1996 to 1998 (Gawalko et al. 2001). This is an unusual geochemical map, but it represents an important end pause, rather than an end point, in the biogeochemical cycle. In terms of the full biogeochemical cycle involving human populations, the next step is milling and incorporation into foodstuffs. Unlike the data for copper, zinc, iron, and manganese, the data for selenium and cadmium show much greater variability (Gawalko et al. 2001). The reason for this variability is that the wheat plants interact with their soil environment to ensure uptake of the essential micronutrients copper, zinc, iron, and manganese. They do not appear to regulate cadmium and selenium which are taken up with the regulated micronutrients to be sequestered in various plant tissues. The spatial distribution of these data is strongly influenced by soil properties, and to a lesser extent the varying selenium content of the soil parent material. The area of highest regional selenium levels is in southwestern Saskatchewan and adjoining Alberta where the soils are dominantly Brown Chernozems (Mollisols, Aridic Borolls), with some Solonetzic (Natric) soils. These are relatively organic carbon poor in comparison with the Dark Brown, Black, and Dark Gray Chernozems (Typic, Udic, and Boralfic Borolls) and Gray Luvisols (Boralfs) that underlie most of the remainder of the Canadian Prairie agricultural region. Organic compounds in soils form seleno-complexes with labile selenium, thus restricting its availability to the plants. The likely cause of the spatial pattern of high selenium is low organic carbon content. The absence of seleno-complexes favors selenium partitioning into soil pore waters where it is available to the plants. A similar situation has been reported from China (Fordyce et al. 2000; Johnson et al. 2000; Wang and Gao 2001), where higher soil organic-carbon contents are associated with lower selenium levels in rice and the increased incidence of Keshan disease. Interestingly, a link was made by MacPherson et al. (1997) between the selenium content of Canadian wheat and the selenium status of the Scottish population. The authors related the lowering of selenium status over time to the shift to importing lower selenium European wheat in preference to higher selenium Canadian bread wheat.

The examples above represent the many regional geochemical and biogeochemical studies that have been undertaken in the twentieth century. In some cases these data are presented spatially, and in others temporally. They clearly demonstrate the spatial and temporal chemical variability of the natural environment. Life is supported on the Earth's chemically inhomogeneous surface, and there are ecosystem consequences to that reality. To fully understand these consequences and manage any associated risks, a sound knowledge of baseline geochemistry is required.

3.6 Establishing Geochemical Baselines

Applied geochemical surveys are undertaken for one of two basic reasons: (1) to detect geochemical anomalies both natural, e.g., related to mineral occurrences, and anthropogenic, e.g., related to industrial releases; and (2) to map and establish natural background levels or baselines. In the first instance, surveys are designed as search exercises so that a feature of known size and geochemical contrast from the

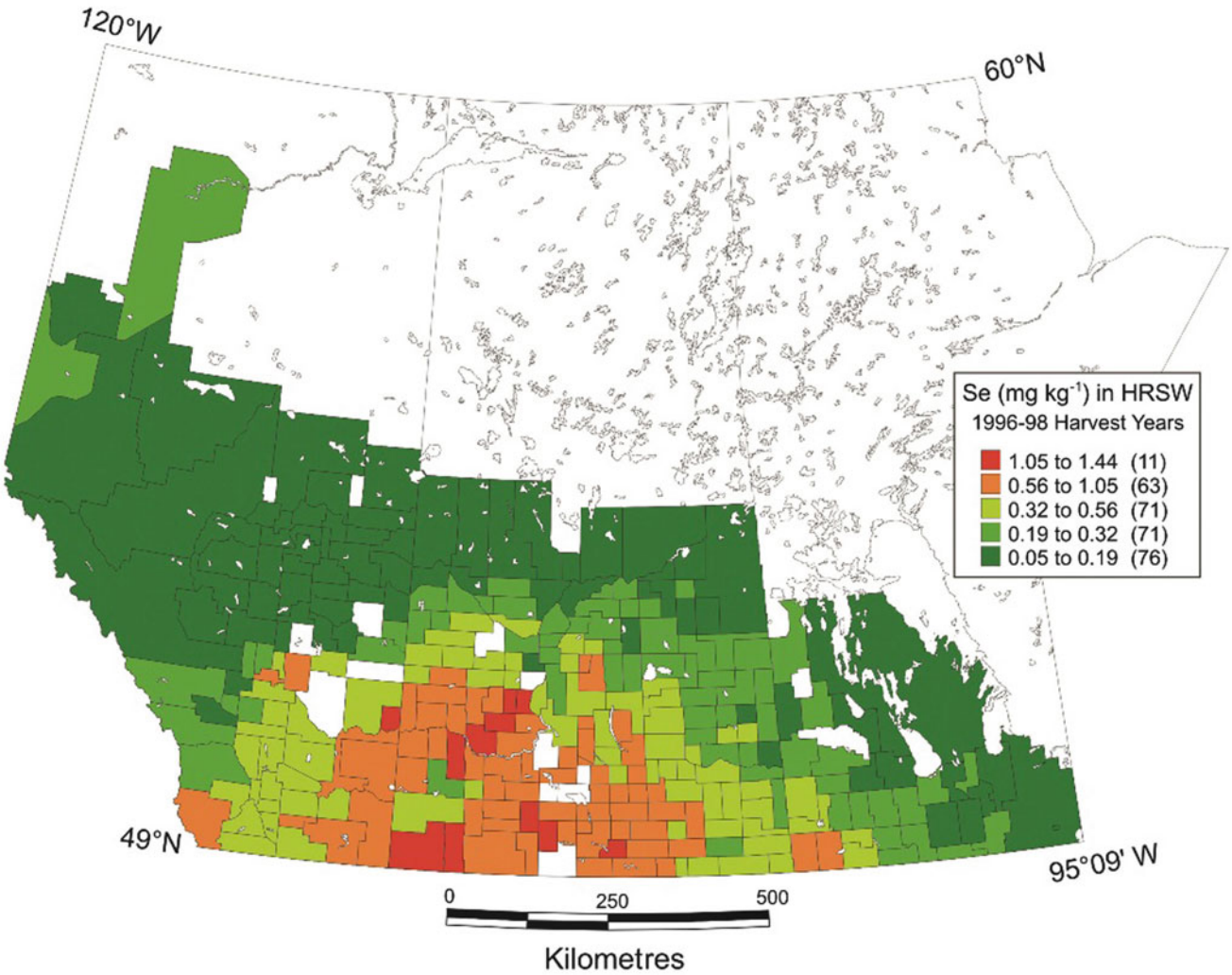


Fig. 3.16 Distribution of selenium (mg kg^{-1}) in hard red spring wheat for the 1996–1998 harvests in the Canadian Prairies

local natural background concentrations can be detected. The sampling is undertaken over an area of fixed extent, perhaps a map sheet, a particular geological terrain, or a particular jurisdiction or economic zone. Surveys undertaken to establish natural background and baselines are designed to sample the area of interest, such as a particular jurisdiction or terrain (e.g., an eco-district), in an unbiased manner so that an average may be estimated together with a measure of the data variability. Depending on the size of the area, determining if there is systematic spatial variability across the survey area may be required. The sampling considerations for these two types of survey have been discussed by Garrett (1983).

To be able to reliably estimate the geochemical level and relief (e.g., mean and variance) for a study area, a sufficient number of samples must be collected and analyzed. If the study area can be treated as a homogeneous entity, a useful rule of thumb based on the formula for the standard error of the mean ($SE = s.n^{-0.5}$) is that a minimum of 30 sites should be sampled, 60 would be better, and it is probably a waste of resources to sample more than 120. In order to obtain unbiased estimates the samples sites should be distributed randomly across the study area. A common strategy used by geoscientists is to use a square grid with a cell size such that the required number of cells, a minimum of 30 for example, are present in the study area, and then to use a random number table or generator to locate an x–y position (site) in each cell to be sampled. An alternate strategy is to use prior knowledge of rock or soil type distribution and collect material from sufficient sites from these postulated homogeneous units in an unbiased manner to reliably estimate their geochemical level and relief (e.g., mean and standard deviation). For national-scale surveys many more sites are sampled as the goal is to reveal broad-scale regional geochemical variability.

Clearly, factors such as resource availability influence decisions concerning sample design. Although the ability of widely spaced sampling to yield reliable estimates is sometimes questioned, examples from many parts of the

world have demonstrated that low-density sampling can be successfully used to map natural backgrounds on a regional scale, e.g., Darnley et al. (1995), Xie and Cheng (2001), Gustavsson et al. (2001), Reimann et al. (1998, 2003), Salminen et al. (2004, 2005), de Caritat and Cooper (2011), and Reimann et al. (2012a, b). It is important to note that if the sampling is unbiased (random), all parts of the study area have an equal opportunity to be sampled, and there is a finite probability that small features that do not form a "significant" proportion of the study area will be missed. If finding such small features is important, resources are required to increase the sample density. The cruel reality for surveys may be depicted as a triangle. The apices are survey area, detail of information generated, and resources—time, staff, and funds. One can fix any two of these apices, but one cannot fix all three. To minimize survey costs and maximize the return in information, Xie and Cheng (2001) recommended the widely spaced sampling of major river floodplain sediments. They demonstrated that sampling densities between 1 site per 1,000–6,000 km^2, i.e., about 520 samples representing all of mainland China, can provide reliable estimates of regional background levels. The field work for this survey was completed in 1 year (1992–1993) by a sampling team of three people. On this basis, they recommend that surveys of this type be completed before more detailed, time-consuming, and expensive surveys are undertaken. The National Geochemical Survey of Australia also employed similar river sediment sampling to map the continent (de Caritat and Cooper 2011).

In any baseline survey consistent protocols are essential for both field and laboratory work. The "target population" needs to be specifically defined, e.g., river/stream water collected midstream 10 cm below surface or surface soils collected from 0 to 25 cm (0–10 inches) at sites at least 100 m from a road that are visibly uncontaminated. Adequate field notes, photographs and accurate location (the latter two are facilitated by digital cameras and global positioning systems, GPS) information need to be recorded. The procedures for on-site treatment such as filtering and acidification of waters, and storage, shipping, and preparation of samples like drying and screening of soil (2 mm or in the range 0.1–0.18 mm), or drainage sediment (in the range 0.1–0.18 mm), and the retention of the fine fraction for analysis, need to be clearly laid out. The difference in retained fractions for soils is historical, and care needs to be taken in selecting an appropriate size fraction. Traditionally soil (as prescribed by the United Nations Food and Agriculture Organization, UN-FAO) and environmental scientists have employed the <2 mm fraction for analysis, while applied geochemists have tended to employ a finer fraction of <0.150 or <0.177 mm, as used for drainage sediment surveys in mineral exploration, or <0.063 mm (<63 μm) as used for glacial sediments and the soils developed on them. Compatibility with prior data sets is a major consideration, as are considerations of plant–soil relationships. Appropriate analytical protocols and QA/QC procedures need to be in place. Procedures and considerations for baseline surveys are discussed in Darnley et al. (1995), Salminen and Gregorauskiene (2000), and by the Forum of European Geological Survey (FOREGS) in Salminen et al. (1998). Procedures for geochemically mapping agricultural and grazing lands in Europe are discussed by the EuroGeoSurveys Geochemistry Working Group (EGA 2008), and those for North American soils by the Geological Survey of Canada (Friske et al. 2010; Rencz et al. 2011).

An important issue to consider in planning baseline surveys is temporal variability. For soils this is not a major consideration; however, for surface and groundwater changes in flow rates resulting from climatic variations such as seasonal rains (e.g., monsoons) or snow-melt events affect the elemental levels observed in the waters. Therefore, it is necessary to sample a region under similar conditions, and to subsequently only use the data for comparison with data sets collected under similar conditions. In the case of stream and river sediments, strong seasonal differences in water flow can modify the bed load composition, and therefore its chemistry (Chork 1977; Rose et al. 1979; Steenfelt and Kunzendorf 1979). In general, if an effect is present, seasonal rains or spring freshets mass-waste bank material into streams and rivers, and fine sediments are then winnowed from the streambed in higher energy (faster flow) environments and deposited in lower energy (slower flow) environments over the period of the subsequent dry season or lower flow period. The result is that levels of trace elements that occur in the finer fraction decrease in higher energy stream environments and increase in lower energy environments. The converse is true for trace elements that occur in coarser or heavier fractions of the sediment. Thus, if seasonal variation is expected, sampling programs should be restricted to longer periods of steady stream flow. If severe weather events occur, e.g., cyclones or hurricanes, the complete bed load may be changed which results in major changes in sediment geochemistry (Ridgway and Dunkley 1988; Ridgway and Midobatu 1991; Garrett and Amor 1994). Thus, if a specific long-term study is undertaken where knowledge of baseline levels is important, catastrophic weather events will likely require a post-event survey to determine if baseline levels have changed significantly.

Once the analytical data are in hand, the accuracy and precision of the data need to be estimated consistent with international standards (Darnley et al. 1995), to determine if they are adequate. If the field sampling has been structured so that analysis of variance (Garrett 1983) or geostatistical procedures (e.g., Issaks and Srivastava 1989) can be applied, the presence of significant spatial trends across the study area can be investigated.

As discussed earlier, geochemical data are often drawn from multiple populations. Therefore it is prudent to summarize the data as percentiles (e.g., minimum, 2nd, 5th, 10th, 25th (first quartile), 50th (median), 75th (3rd quartile), 90th, 95th, 98th percentiles, and maximum), as well as arithmetic means and standard deviations, and robust estimates such as the median and mean absolute deviation. A complication arises because geochemical data are compositional, i.e. they have a constant sum to 100% or 1 million mg kg^{-1}. For the computation of univariate statistical summaries, a logistic transform of the data is required. However, for trace elements and concentrations up to 10%, a logarithmic transform will suffice (Filzmoser et al. 2009). Thus, the geometric mean and the standard deviation in logarithmic units may be presented. Other properties of the data set to be reported are the number of samples analyzed, the lower quantification limit, and how many samples are below the limit. The reporting of the data as percentiles is a non-parametric procedure that avoids any assumption concerning the distribution of the data, and their inspection quickly reveals whether the distribution is skewed. In addition, percentiles can assist in setting realistic, in the sense of the natural distribution of an element, regulatory levels as described in Sect. 3.2.

Often for jurisdictional or regulatory reasons, the geographical entities over which baseline surveys are undertaken have no direct, or only an indirect, relation to geology and pedology. The Commission for Environmental Cooperation of the North American Free Trade Agreement (NAFTA−CEC) has prepared a geographic eco-classification for North America (see http://www.cec.org/pubs_info_resources/publications/enviro_conserv/ecomap.cfm and Marshall et al. 1996), which is likely to see increasing use as a way to sub-divide natural background data into entities of ecological and environmental relevance. The eco-district boundaries are strongly influenced by soil (soil series) properties that reflect the underlying geological and biological processes, which in turn reflect climate. An example of presentation of natural background data using this framework is presented in Fig. 3.17 for reverse aqua-regia soluble zinc determined in the <0.177 mm fraction of stream sediments from Yukon Territory, Canada. The eco-district medians vary by a factor of 6, again demonstrating the spatial variability in natural background levels. There are sound geological reasons for the spatial patterns: the highest levels relate to zinc-rich black shales in the Selwyn basin in the northeast; low levels relate to the Yukon crystalline terrain in Central Yukon with generally higher levels to the southwest and northeast; and numerous smaller details may be explained by the presence of particular rock units.

3.7 Total Analyses and Bioavailability

Geoscientists traditionally determine the total or near-total amounts of elements in the samples they collect, except for specific geochemical exploration procedures where a wide range of analytical protocols are employed. Amongst the near-total procedures, methods based on aqua regia digestions are widely used, for example, the US-EPA 3050B aqua regia variant (US-EPA, 1996). The total, or near-total, amount of metal in a rock, soil, or sediment is a poor estimator of what may become phyto- or bioavailable and be able to cross cellular barriers (for example, Allen 2002a).

A wide variety of protocols have been developed by agronomists and ecotoxicologists to better estimate the phyto- and bioavailable amounts of an element in soils, waters, plants, etc. An additional issue is: Bioavailable to what? (Allen 2002a, b). This has led to a great variety of protocols, some of which are locally adequate for specific studies (see, for example, Sauvé 2002); however, some extractants, e.g., 0.01 M $CaCl_2$ and 0.1 M $NaNaO_3$, have been proposed for more general application (Menzies et al. 2007). No extraction has universal application, though the estimation of free-ion concentrations in free or pore waters (the Free Ion Activity Model, FIAM) approaches that requirement (see Campbell 1995). Another important factor influencing the availability of metals in solutions to biota is the presence of dissolved organic carbon compounds, many of which are capable of sequestering trace elements so that they remain unavailable. This situation has been addressed by Tipping (1994) in the WHAM model which includes recognition of humic acid complexes. Finally, it must be remembered that the interaction between the biotic and abiotic realms is not passive. Biota are capable of mediating the passage of ions across cellular barriers. In the case of plants, they are capable of acidifying or releasing low molecular weight organic acids, e.g., phytosiderophores (deoxymugineic acid), to the immediate root environment (rhizosphere) to liberate or chelate micronutrients and other trace elements into the proximal pore waters (moisture) so they are available to the plant.

However, much geoscience-driven baseline work is still undertaken using total or near-total aqua-regia or hot nitric acid digestions, and for many sample materials these procedures overestimate the amount of metal that may be available to biota. Notable exceptions are the 1:2,500,000 scale soil maps found in the geochemical atlas of Poland (Lis and Pasieczna 1995), which employed a 1:4 HCl mixture, and the Baltic Soil Survey (Reimann et al. 2003) that generated data using an ammonium acetate ($CH_3COO.NH_4$) extraction. A simple water extraction at natural soil

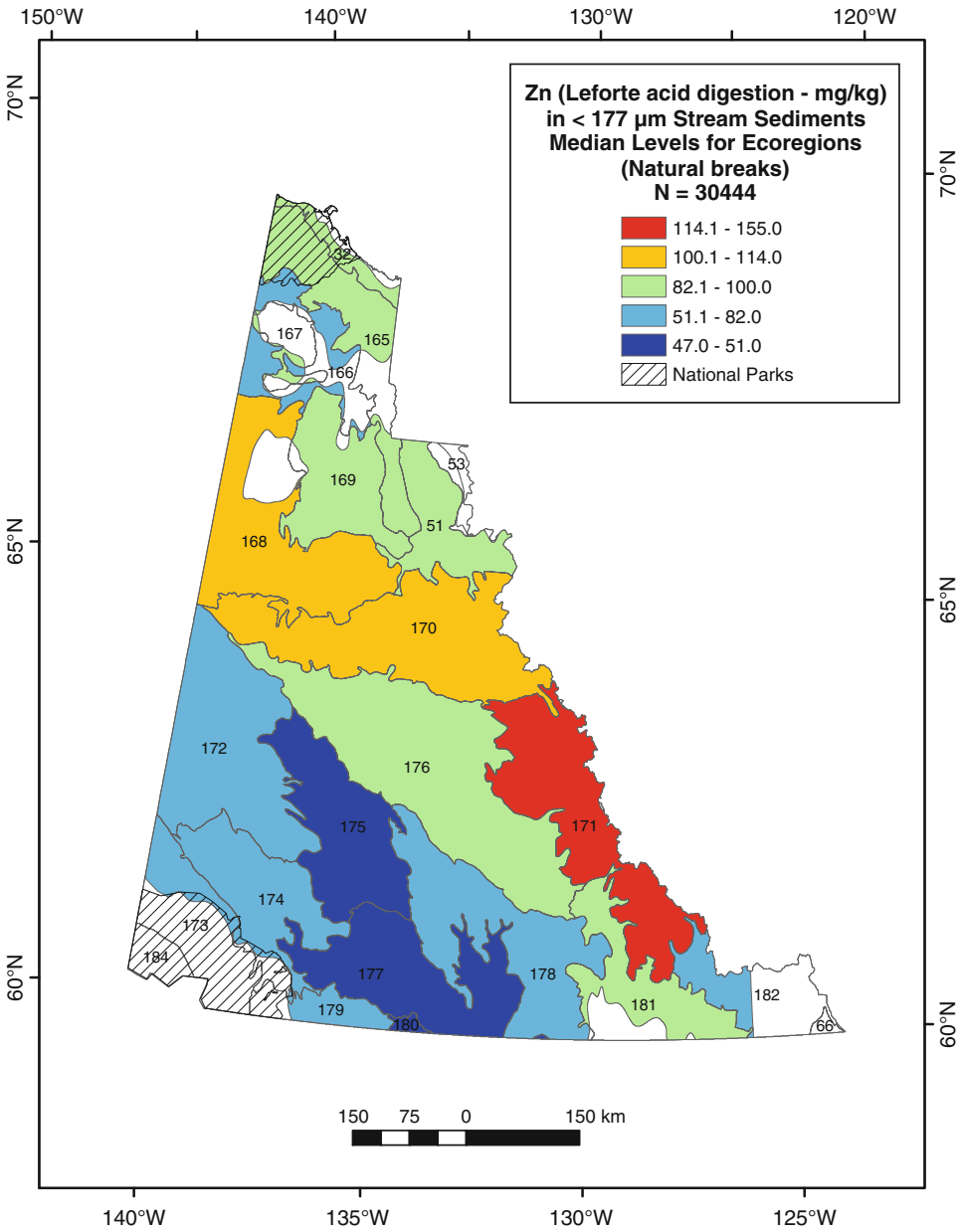


Fig. 3.17 Distribution of zinc (mg kg^{-1}) in the <0.177 µm fraction of 26,862 stream sediments displayed as eco-district median values, Yukon Territory, Canada

pHs has been used in North American studies (Garrett 2009; Friske et al. 2012) Although there are sound reasons for these choices, the data are no longer comparable with other international data sets. In an ideal world the best procedure would be to have a total/near-total determination and an appropriate weaker extraction that would better estimate the phyto- or bioavailable amount in the sample material. A further important point is that all biota are not created equal, and an extraction that may be good for agricultural crops might be quite inappropriate for soil invertebrates.

The prime controls on phyto- and bioavailability are the mineralogical or material form of an element in question and its mobility in the aqueous environment, e.g., soil or sediment pore waters, fresh stream or lake water, or marine sediments and water. Table 3.4 provides an indication of the mobility of many trace elements of interest as a function of pH and redox conditions (see also Plant et al. 2001, Figure 6).

With reference to solid phases, elements tend to be bioavailable when they are loosely held on mineral surfaces or present as metallo-organic complexes. Thus, weak extractants: acetic acid; sodium or ammonium acetate; calcium chloride; potassium, sodium or ammonium nitrate, and chelating agents such as EDTA, DTPA, and sodium pyrophosphate are commonly used in analytical protocols to estimate phytoavailable amounts of elements in soils. Sauvé (2002) and Menzies et al. (2007) have provided reviews of these methods in the context of determining metal speciation (chemical bonding) in soils. In contrast to the soil protocols,

Table 3.4 Mobility of elements in the surface environment

	Oxidizing (pH 5–8)	Oxidizing (pH < 4)	Reducing
Relative mobility			
Highly mobile	Cl, Br, I, S, Rn, He, C, N, Mo, B, Se, Te	Cl, Br, I, S, Rn, He, C, N, B	Cl, Br, I, Rn, He
Moderately mobile	Ca, Na, Mg, Li, F, Zn, Ag, U, V, As, Sb, Sr, Hg	Ca, Na, Mg, Sr, Li, F, Zn, Cd, Hg, Cu, Ag, Co, Ni, U, V, As, Mn, P	Ca, Na, Mg, Li, Sr, Ba, Ra, F, Mn
Slightly mobile	K, Rb, Ba, Mn, Si, Ge, P, Pb, Cu, Ni, Co, Cd, In, Ra, Be, W	K, Rb, Ba, Si, Ge, Ra	K, Rb, Si, P, Fe
Immobile	Fe, Al, Ga, Sc, Ti, Zr, Hf, Th, Sn, REEs, Pt metals, Au, Cr, Nb, Ta, Bi, Cs	Fe, Al, Ga, Sc, Ti, Zr, Hf, Th, Sn, REEs, Pt metals, Au, As, Mo, Se	Fe, Al, Ga, Sc, Ti, Zr, Hf, Th, Sn, Fe, REEs, Au, Cu, Pt metals, Ag, Pb, Zn, Cd, Hg, Ni, Co, As, Sb, Bi, U, V, Se, Te, Mo, In, Cr, Nb, Ta

After Rose et al. (1979)

dilute hydrochloric acid with various additions has been used to simulate gastric fluids in estimating the amount of trace elements that could be solubilized from soils in the digestive tract of an animal.

3.8 Future Challenges

The great challenge of the future is to provide consistent, relevant, high-quality geochemical data to support epidemiological research, environmental regulation, and other studies such as agricultural and forestry resource management that are concerned with the transfer of trace elements through potable water and the food chain into human and animal populations.

Attention has been drawn to this by publication of the report on International Geological Correlation Projects 259 and 360 (Darnley et al. 1995) and by Plant et al. (2001). Funding of such regional- and continental-scale mapping activities poses a major challenge. To date significant progress has been made in China, Europe, Australia and parts of North America where there is a national commitment to monitor the surficial environment in sufficient detail to yield useful maps (Xie and Cheng 2001). Progress has also been made in Europe through the collaborative efforts of nations working through the Forum of European Geological Surveys. The challenge for the future is to create the interdisciplinary teams that can generate the critical mass to organize and execute systematic baseline surveys at continental scales with the support of agencies with the resources and vision to appreciate the value of a global geochemical background database.

The advent of rapid global change will stress, some would argue is stressing, the world's resource base, and makes sustainable development an even more important issue than it is now. The introduction of the new European Chemicals Regulation, REACH (Registration, Evaluation and Authorization of Chemicals), and the implementation of the EU Soil Protection Directive provided the impetus and funding for the GEMAS project (Reimann et al. 2012a, b) to augment knowledge concerning soil quality. The study of Pb-isotopes on a continental scale (Reimann et al. 2012b) provides context for the use of Pb-isotopes to infer natural or anthropogenic sources for lead in the environment and is a valuable new addition to geochemical mapping. The role of human activity as a causative factor may be argued by some, but it remains that global economic development has radically increased the rate of change in the environment (Fyfe 1998). The changes to the global environment wrought by humans has led to the recognition of the current geological age as the Anthropocene (Revkin 1992; Zalasiewicz et al. 2010). As Plant et al. (2001) noted: "The problem is most acute in tropical, equatorial, and desert regions where the surface environment is particularly fragile because of its long history of intense chemical weathering over geological time scales." Change needs to be monitored, but how can it be monitored if baselines are not known? Concerted international action is required to acquire the data essential for managing the risks that the natural environment poses to the world's population (Zoback 2001).

See Also the Following Chapter. Chapter 4 (*Anthropogenic Sources*)

Further Reading

Ali SA (2011) Natural products as therapeutic agents for schistosomiasis. Res J Med Plant 5:1–20

Allen HE (ed) (2002a) Bioavailability of metals in terrestrial ecosystems: importance of partitioning for bioavailability to invertebrates, microbes and plants. Society of Environmental Toxicological and Chemistry (SETAC), Pensacola

Allen HE (2002b) Chapter 1: Speciation of metals in soil. In: Allen HE (ed) Bioavailability of metals in terrestrial ecosystems: importance of partitioning for bioavailability to invertebrates, microbes and plants. Society for Environmental Toxicology and Chemistry (SETAC), Pensacola, pp 1–5

Brooks RR (1972) Geobotany and biogeochemistry in mineral exploration. Harper & Row, New York
Brooks RR (ed) (1998) Plants that hyperaccumulate heavy metals: their role in phytoremediation, microbiology, archaeology, mineral exploration, and phytomining. CAB International, Oxford
Campbell PGC (1995) Interactions between trace metals and aquatic organisms: a critique of the free-ion activity model. In: Tessier A, Turner DR (eds) Metal speciation and bioavailability in aquatic systems. Wiley, New York, pp 45–102
Chork CY (1977) Seasonal sampling and analytical variations in stream sediment surveys. J Geochem Explor 7:31–47
Clarke FW, Washington HS (1924) The data of geochemistry. U S Geol Surv Bull 770:841
Darnley AG, Björklund A, Bølviken B, Gustavsson N, Koval PV, Plant JA, Steenfelt A, Tauchid M, Xie X, with contributions by Garrett RG, Hall GEM (1995) A global geochemical database for environmental and resource management. Earth Science Series no. 19. UNESCO, Paris
de Caritat P, Cooper M (2011) National geochemical survey of Australia: the geochemical atlas of Australia. Geoscience Australia, Canberra, p 557, Record 2011/20 (2 vols)
Dunn CE (1990) Lithogeochemical study of the Cretaceous in central Saskatchewan—perliminary report. Sask Geol Surv Misc Rep 90–4:193–197
Dunn CE (2007) Biogeochemistry in mineral exploration. In: Hale M (Series ed) Handbook of exploration and environmental geochemistry, vol 9. Elsevier, Amsterdam
EGA (2008) EuroGeoSurveys Geochemistry Working Group. EuroGeoSurveys geochemical mapping of agricultural and grazing land in Europe (GEMAS)—Field manual. Report, 2008.038, Norges Geologiske Undersøkelse, Trondheim, p 46
Fairbrother A, McLaughlin MJ (2002) Metalloregions. Fact sheet on environmental risk assessment 12. International Council on Mining and Metals (ICMM), London
Faust SD, Aly OM (1981) Chemistry of natural waters. Ann Arbor Science, Ann Arbor
Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Tot Environ 407:6100–6108
Fordyce FM, Zhang G, Green K, Liu X (2000) Soil, grain and water chemistry in relation to selenium-responsive diseases in Enshi district, China. Appl Geochem 15:117–132
Fredén C (ed) (1994) Geology. In: Wastenson L (ed) National atlas of Sweden Series. SNA Publishing, Stockholm
Friske PWB, Ford KL, Kettles IM, McCurdy MW, McNeil RJ, Harvey BA (2010) North American soil geochemical landscapes project: Canadian field protocols for collecting mineral soils and measuring soil gas radon and natural radioactivity. Open File 6282, Geological Survey of Canada, Ottawa
Friske PWB, Ford KL, McNeil RJ (2012) Soil geochemical, mineralogical, radon and gamma ray spectrometric data from the 2007 North American soil geochemical landscapes project in New Brunswick, Nova Scotia and Prince Edward Island. Open File 6433, Geological Survey of Canada, Ottawa
Fyfe WS (1998) Towards 2050; the past is not the key to the future; challenges for environmental geology. Environ Geol 33:92–95
Garrett RG (1983) Sampling methodology. In: Howarth RJ (ed) Handbook of exploration geochemistry, vol 2, statistics and data analysis in geochemical prospecting. Elsevier, Amsterdam, pp 83–110
Garrett RG (2009) A water-leach procedure for estimating bioaccessibility of elements in soils from transects across the United States and Canada. Appl Geochem 24:1438–1453
Garrett RG, Amor SD (1994) Temporal variations in stream sediment data from Jamaica. In: Bloom L (ed) Prospecting in tropical and arid terrains. Prospectors and Developers Association of Canada, Toronto, pp 117–137 (Section 2)
Gawalko EJ, Garrett RG, Nowicki TW (2001) Trace elements in western Canadian hard red spring wheat (*Triticum aestivem* L.): levels and quality assurance. J Assoc Anal Intl 84:1953–1963
Goldschmidt VM (1937) The principles of distribution of chemical elements in minerals and rocks. J Chem Soc (Lond) 1937 (Pt. 1):655–673
Gustavsson N, Bolviken B, Smith DB, Severson RC (2001) Geochemical landscapes of the conterminous United States—New map presentations for 22 elements. U S Geol Surv Bull 1645:38
Issaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York
Johnson CC, Ge X, Green KA, Liu X (2000) Selenium distribution in the local environment of selected villages of the Keshan disease belt, Zhangjiakou district, Hebei province, People's Republic of China. Appl Geochem 15:385–401
Kabata-Pendias A (2001) Trace elements in soils and plants, 3rd edn. CRC Press, Boca Raton
Lee T, Yao C-L (1970) Abundance of chemical elements in the earth's crust and its major tectonic units. Intl Geol Rev 12(7): 778–786
Lis J, Pasieczna A (1995) Geochemical atlas of Poland. Polish Geological Institute, Warsaw
MacPherson A, Barclay MNI, Scott R, Yates RWS (1997) Loss of Canadian wheat imports lowers selenium intake and status of the Scottish population. In: Fischer PWF, Abbé MRL, Cockell KA, Gibson RS (eds) Trace elements in man and animals-9. National Research Council Press, Ottawa, pp 203–205
Marshall IB, Smith CAS, Selby CJ (1996) A national framework for monitoring and reporting environmental sustainability in Canada. Environ Monit Assess 39:25–38
McKelvey VE (1960) Relation of reserve of the metals to their crustal abundance. Am J Sci 258–A:234–241
McLennan SM (1992) Continental crust. In: Nierenberg WA (ed) Encyclopedia of earth sciences, vol 1. Kluwer, Dortrecht, pp 581–592
Menzies NW, Donn MJ, Kopittke PM (2007) Evaluation of extractants for estimation of phytoavailable trace elements in soils. Env Poll 145:121–130
Ontario Ministry of Environment and Energy (1999) Ontario typical range of chemical parameters in soil, vegetation, moss bags and snow, version 1.0a. Queen's Printer for Ontario, Toronto
Plant J, Smith D, Smith B, Williams L (2001) Environmental geochemistry on a global scale. Appl Geochem 16:1291–1308
Reimann C, de Caritat P (1998) Chemical elements in the environment: fact sheets for the geochemist and environmental scientist. Springer, Berlin
Reimann C, Garrett RG (2005) Geochemical background—concept and reality. Sci Tot Env 350:12–27
Reimann C, Melezhik V (2001) Metallogenic provinces, geochemical provinces and regional geology—what causes large-scale patterns in low density geochemical maps of the C-horizon of podzols in arctic Europe. Appl Geochem 16:963–983
Reimann C, Äyräs M, Chekushin VA, Bogatyrev IV, Rognvald B, de Caritat P, Dutter R, Finne TE, Halleraker JH, Jæger Ø, Kashulina G, Niskavaara H, Lehto O, Pavlov VA, Räisänen ML, Strand T, Volden T (1998) Environmental geochemical atlas of the central Barents region. Geological Survey of Norway, Trondheim
Reimann C, Siewers U, Tarvainen T, Bityukova L, Erikson J, Gilucis A, Gregorauskiene V, Lukashev VK, Matinian NN, Pasieczna A (2003) Agricultural soils in northern Europe: a geochemical atlas. Bundesanstalt für Geowissenschaften und Rohstoffe: Sonderhefte Reihe D, Hefte SD5, Hannover
Reimann C, Filzmoser P, Garrett RG (2005) Background and threshold: critical comparison of methods of determination. Sci Tot Env 346:1–16

Reimann C, Matschullat J, Burke M, Salminen R (2009) Arsenic distribution in the environment: the effects of scale. Appl Geochem 24:1147–1167

Reimann C, Matschullat J, Burke M, Salminen R (2010) Antimony in the environment: lessons from geochemical mapping. Appl Geochem 25:175–198

Reimann C, de Caritat P, GEMAS Project Team, NGSA Project Team (2012a) New soil composition data for Europe and Australia: demonstrating comparability, identifying continental-scale processes and learning lessons for global geochemical mapping. Sci Tot Env 416:239–252

Reimann C, Flem B, Fabian K, Borke M, Ladenberger A, Négrel P, Demetriades A, Hoogewerff J, The GEMAS Project Team (2012b). Lead and lead isotopes in agricultural soils of Europe—the continental perspective. Appl Geochem 27:532–542

Rencz AN, Garrett RG, Kettles IM, Grunsky EC, McNeil RJ (2011) Using soil geochemical data to estimate the range of background element concentrations for ecological and human health risk assessments. Current Research 2011–9, Geological Survey of Canada: Ottawa

Revkin A (1992) Global warming: understanding the forecast. Abbeville, New York

Ridgway J, Dunkley PN (1988) Temporal variations in the trace element content of stream sediments: examples from Zimbabwe. Appl Geochem 3:609–621

Ridgway J, Midobatu C (1991) Temporal variations in the trace element content of stream sediments: an example from a tropical rain forest regime, Solomon islands. Appl Geochem 6:185–193

Rose AW, Hawkes HE, Webb JS (1979) Geochemistry in mineral exploration, 2nd edn. Academic, London

Rudnick RL, Gao S (2003). Composition of the continental crust. In: Rudnick RL (ed) The crust. Treatise on geochemistry, vol 3 (Holland HD, Turekian KK (Series eds)). Elsevier-Pergamon, Oxford, pp 1–64

Salminen R, Gregorauskiene V (2000) Considerations regarding the definition of a geochemical baseline of elements in the surficial materials in areas differing in basic geology. Appl Geochem 15:647–653

Salminen R, Tarvainen T, Demetriades A, Duris M, Fordyce FM, Gregorauskiene V, Kahelin H, Kivisilla J, Klaver G, Klein H, Larson JO, Lis J, Locutura J, Marsina K, Mjartanova H, Mouvet C, O'Connor P, Odor L, Ottonello G, Paukola T, Plant JA, Reimann C, Schermann O, Siewers U, Steenfelt A, Van der Sluys J, Williams L (1998) FOREGS geochemical mapping: field manual. Geol Surv Finland Guide 47:42

Salminen R, Chekuskin V, Tenhola M, Bogatyrev I, Glavatskikh SP, Fedotova E, Gregorauskiene V, Kashulina G, Niskavaara H, Polischuok KA, Rissanen L, Selenok L, Tomilina O, Zhdanova L (2004) Geochemical atlas of the eastern Barents region. Elsevier, Amsterdam (reprinted from Jour Geochem Explor vol 83)

Salminen T (Chief ed), Batista MJ, Bidovec M, Demetriades A, De Vivo B, De Vos W, Duris M, Gilusis A, Gregorauskiene V, Halamic J, Heitzmann P, Lima A, Jordan G, Klaver G, Klein P, Lis J, Locatura J, Marsina K, Mazreku A, O'Connor PJ, Olsson SA, Otteen R-T, Petersell V, Plant JA, Reeder S, Salpeteur I, Sandstrom H, Siwers U, Steenfelt A, Tarvainen T (2005) Geochemical atlas of Europe, part 1—Background information, methodology, and maps. Geological Survey of Finland, Espoo

Sauvé S (2002) Chapter 2: Speciation of metals in soil. In: Allen HE (ed) Bioavailability of metals in terrestrial ecosystems: importance of partitioning for bioavailability to invertebrates, microbes and plants. Society for Environmental Toxicology and Chemistry (SETAC), Pensacola, pp 7–58

Shacklette HT, Boerngen JG (1984) Element concentrations in soils and other surficial materials of the conterminous United States. U S Geol Surv Bull 1270:105

Simpson PR, Breward N, Flight DMA, Lister TR, Cook JM, Smith B, Hall GEM (1996) High resolution regional hydrogeochemical baseline mapping of stream waters in Wales, the Welsh borders and West Midlands region. Appl Geochem 11:621–632

Steenfelt A, Kunzendorf H (1979) Geochemical methods in uranium exploration in northern east Greenland. In: Watterson JR, Theobald PK (eds) Geochemical exploration 1978, Spec. vol 7. Association of Exploration Geochemists, Toronto, pp 429–442

Taylor SR (1964) Abundance of chemical elements in the continental crust, a new table. Geochim Cosmochim Acta 28(8):1273–1285

Tipping E (1994) WHAM—a chemical equilibrium model and computer code for water, sediments, and soils incorporating a discrete site/electrostatic model of ion-binding by humic substances. Comput Geosci 21:973–1023

Turekian KK, Wedepohl KH (1961) Distribution of the elements in some major units of the Earth's crust. Bull Geol Soc Am 72 (1):175–192

U.S. Environmental Protection Agency (1996). Method 3050b—Acid digestion of sediments, sludges, and soils. http://www.epa.gov/wastes/hazard/testmethods/sw846/pdfs/3050b.pdf. Accessed 10 Apr 2012

Vine JD, Tourtelot EB (1970) Geochemistry of black shale deposits; a summary report. Econ Geol 65(3):253–272

Vistelius AB (1960) The skew frequency distributions and the fundamental law of the geochemical processes. J Geol 68:1–22

Wang Z, Gao Y (2001) Biogeochemical cycling of selenium in Chinese environments. Appl Geochem 16:1345–1351

Wedepohl KH (1995) The composition of the continental crust. Geochim Cosmochim Acta 59:1217–1232

Xie X, Cheng H (2001) Global geochemical mapping and its implication in the Asia-Pacific region. Appl Geochem 16:1309–1321

Zalasiewicz I, Williams M, Steffan W, Crutzen P (2010) The New world of the anthropocene. Environ Sci Technol 44:2228–2231

Zoback ML (2001) Grand challenges in earth and environmental sciences—science, stewardship, and service in the 21st century. Geol Soc Am Today 11:41–47

Anthropogenic Sources

4

Ron Fuge

Contents

4.1 Introduction

As outlined in Chap. 3, the geochemistry of environmental media is largely dependent on the chemistry of the natural sources from which they have been derived, or with which they have interacted. Thus soil and surficial sediment chemistry are strongly influenced by the composition of their parent materials. Similarly stream and river waters, derived initially from precipitation, depend on the rocks, sediments, and soils from which they come into contact and interact with for their chemical composition. However, with the evolution of humans in the relatively recent geological past there have been anthropogenic impacts on the environment, which have increased dramatically with increasing population, urbanization, and industrialization (Fyfe 1998). Thus humans have contaminated or polluted the once pristine environment, and this impact is manifested in the chemistry of environmental materials that reflect anthropogenic signals superimposed on the natural composition.

Many human activities have resulted in environmental contamination and these include:

1. Mineral extraction and processing
2. Smelting and refining of mineral ores and concentrates
3. Power generation—fossil fuel, nuclear, geothermal, and hydroelectric
4. Other industrial and manufacturing activities—metallurgical and chemical industries, brick and pipe manufacture, cement manufacture, the ceramics and glass industry, plastics and paint manufacture, and fertilizer manufacture
5. Waste disposal—household refuse, fly ash, sewage, nuclear, and the open burning of refuse
6. Agricultural practices—application of mineral-based fertilizers and manure together with sewage sludge, application of pesticides and herbicides, farmyard runoff including such materials as sheep dip, etc., and deforestation, which has contributed to problems of mercury contamination of Brazilian rivers

R. Fuge (✉)
Institute of Geography and Earth Sciences, Aberystwyth University, Aberystwyth, UK
e-mail: rrf@aber.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_4,

7. Transportation,—motor vehicle derived contamination; this is particularly important in urban environments
8. Treatment and transport (through metal pipes and fittings) of potable water

Nowhere is the impact of environmental pollution more apparent than in the urban environment. Road dust and soils in the urban environment can be heavily contaminated, particularly by metals. Road dust and fine soil particles are the major sources of house dust and as such represent a potential pathway into the human body from inhalation and inadvertent ingestion from hands, etc., particularly for children.

There are many examples of anthropogenically derived substances having marked effects on human health; for example, in 1956 the mercury poisoning experienced in the Minimata region of Japan, known as "Minimata disease." This resulted from a factory situated on the coast releasing mercury, which was used as a catalyst in plastics manufacture, into Minimata Bay. From there it was passed through the marine food chain in a methylated form [CH_3Hg^+ or $(CH_3)_2Hg$], and into humans via consumption of fish. Mercury attacks the central nervous system and causes irreversible brain damage. The methyl form of mercury represents a particularly serious toxic threat, as it is able to cross membranes and accumulate in the central nervous system. It has been estimated that over 20,000 people were affected.

Although the main thrust of this book concerns the impact of the natural environment on human health, this chapter deals briefly with anthropogenic impacts on the composition of environmental media together with some brief considerations of health impacts. This is included for the sake of completeness and as such it is not a detailed account of the topic; indeed, such a detailed account would require a second book.

4.2 Impact of Mineral Extraction and Refining

4.2.1 Mining and Mineral Extraction

Mineral deposits represent concentrations of an element or elements to a level at which they can be profitably extractable. As such these deposits represent concentrations of several elements well above crustal abundance. Such naturally occurring high concentrations are reflected in the chemistry of the soils, waters, sediments, plants, etc., in the immediate vicinity of the deposit. Indeed, measurement of the concentrations of various metals and non-metals in media such as soils, sediments, and waters have been used to locate mineral deposits. This practice is described as geochemical exploration.

Although there are likely to be significant natural enrichments of several elements in the vicinity of mineral deposits, mining and extraction of the deposit will add greatly to these enrichments. The mining and subsequent beneficiation of minerals and the separation and refining of their various components is one of the most serious sources of contamination of soils, waters, and the biosphere.

Humans have extracted minerals, particularly the metalliferous ores, since ancient times, and the extraction and refining of metals have played a major role in human development. The mining and processing of minerals have increased through time, due to population growth and the greater utilization of raw materials for manufacture. Many areas of past mining activity, in both the Old and New Worlds, bear witness to these extractions in the form of abandoned workings and extensive waste tips. Modern mineral extraction technology is generally far more efficient than past practices, and in many countries such processes are heavily regulated to limit the degree of contamination from extractive industries. However, historical mineral extraction involved less efficient technologies, and in those times virtually no environmental regulations were in place. Long-abandoned mineral workings are currently the cause of serious environmental pollution in many countries.

A large number of different materials are extracted from the Earth ranging from fuels such as oil and coal, industrial minerals such as clays and silica, aggregates for building and roadstone, and minerals for fertilizers as well as sources of non-metals. However, the major cause of concern are the metalliferous ores that are used as sources of metals and metalloids. Some of the more important metalliferous ore minerals are listed in Table 4.1.

The extraction and subsequent processing of ores can be summarized as follows:

Mining	→	crushing/grinding
	→	concentration of ore mineral
	→	smelting/refining

Ores are extracted from the Earth by either subsurface mining, open pit surface techniques, or in a few cases by solution mining, which carries with it risks of groundwater pollution. Both subsurface and surface extraction result in waste material, which is generally piled on the surface in the vicinity of the mine. However, while such waste piles, which frequently contain ore minerals, are sources of environmental contamination, it is the subsequent processing of the ores that results in the greatest environmental problems. The crushing and grinding (comminution) of mineral processing has the objective of separating the ore minerals

Table 4.1 Some of the important, mostly metalliferous minerals

Mineral	Composition	Comments
Arsenopyrite	FeAsS	Frequently occurs as a gangue mineral
Barite	$BaSO_4$	Major use in drilling muds
Bauxite	Mainly $Al(OH)_3$	Only ore of aluminum
Bornite	Cu_5FeS_4	Important ore of copper
Carrollite	$Cu(Co,Ni)_2S_4$	Important ore of cobalt
Cassiterite	SnO_2	Main ore of tin
Chalcocite	Cu_2S	Important ore of copper
Chalcopyrite	$CuFeS_2$	Major ore of copper
Chromite	$FeCr_2O_4$	Main ore of chromium
Cinnabar	HgS	Main ore of mercury
Galena	PbS	Main ore of lead
Gold (native)	Au	Main source of metallic gold
Haematite	Fe_2O_3	Major ore of iron
Ilmenite	$FeTiO_3$	Important ore of titanium
Magnetite	Fe_3O_4	Important ore of iron
Molybdenite	MoS_2	Main ore of molybdenum
Pentlandite	$(Fe,Ni)_9S_8$	Major ore of nickel
Platinum	Pt (with other metals)	Main source of platinum
Pyrite	FeS_2	Common gangue mineral
Pyrrhotite	$Fe_{(1-x)}S_2$	Common gangue mineral
Rutile	TiO_2	Important ore of titanium
Scheelite	$CaWO_4$	Important ore of tungsten
Sphalerite	ZnS	Major ore of zinc
Stibnite	Sb_2S	Main ore of antimony
Tetrahedrite	$(Cu,Fe)_{12}Sb_4S_{13}$	Copper ore—often contains silver
Uraninite (Pitchblende)	UO_2	Main ore of uranium
Wolframite	$(Fe,Mn)WO_4$	Important ore of tungsten
Zircon	$ZrSiO_4$	Main ore of zirconium

from the waste, generally referred to as gangue. To effect separation, the mined ore is finely crushed to liberate individual ore mineral grains to enable concentration of the sought after ore minerals. The very fine waste material left after this concentration process is referred to as tailings, and this material can contain, along with the gangue minerals, residual amounts of the ore minerals and can be a serious source of pollution. The tailings are very fine and are subject to wind ablation and can easily be transported by surface runoff. At many long-abandoned mine sites tailings have been left open to the environment resulting in serious contamination of surrounding soils and waters. In more recent mining operations, tailings are stored wet in tailings ponds often behind artificial dams. However, leakage of metal ions into both surface and subsurface waters has, in some cases, resulted in serious contamination of these waters, some of which have been sources of potable water. In addition, after mine closure tailings ponds will dry out unless arrangements are made to keep them permanently wet, which renders the fine material susceptible to spreading across the neighboring area. In modern mining operations, upon closure, remedial action such as isolation of the tailings material by covering serves to limit subsequent environmental pollution.

A further problem of tailings dams is the possibility of dam failure releasing large quantities of highly contaminated sediments and waters into the local environment. There have been several such dam failures in the last decade such as at the Omai gold mine in Guyana in 1995 and at the Mar copper mine on Marinduque Island, Philippines. A major failure of the Los Frailes tailings dam at Aznalcóllar, Spain, in 1998, spilled 6.8 million m^3 of water and pyrite-rich tailings that covered approximately 2,000–3,600 ha of agricultural land. In 2000 a gold mine tailings dam at Borsa in northwestern Romania released large quantities of cyanide and metals into local rivers, the Vaser and Tizla, which ultimately drain into the Danube River. This caused the death of many fish and birds that ate the fish.

In addition to the major metal components of the various ore minerals, many trace constituents are included and these frequently cause as much environmental concern as the major elements in the ores (see Table 4.2). Of the trace elements perhaps the most notorious is cadmium. It is ubiquitous in zinc ores with concentrations of up to 4.4% having been recorded in some ores. Weathering of the ore minerals within the abandoned mines and in waste and tailings tips results in the release of the trace constituents along with the major components of the ore. Soils and waters in the vicinity

Table 4.2 Some trace constituents of selected sulfide minerals (values in mg kg^{-1})

Element	Normal range	Maximum found
Galena (PbS)		
Ag	500–5,000	30,000
As	200–5,000	10,000
Bi	200–5,000	50,000
Cu	10–200	3,000
Sb	200–5,000	30,000
Tl	<10–50	1,000
Sphalerite (ZnS)		
As	200–500	10,000
Cd	1,000–5,000	44,000
Cu	1,000–5,000	50,000
Hg	10–50	10,000
Sn	100–200	10,000
Tl	10–50	5,000
Chalcopyrite ($CuFeS_2$)		
Ag	10–1,000	2,300
Co	10–50	2,000
Ni	10–50	2,000
Sn	10–200	770
Pyrite (FeS_2)		
As	500–1,000	50,000
Co	200–5,000	>25,000
Cu	10–10,000	60,000
Ni	10–500	25,000
Pb	200–500	5,000
Sb	100–200	700
Tl	50–100	100
Zn	1,000–5,000	45,000

From Levinson (1980)

of disused metal mines are frequently heavily contaminated. Figure 4.1 is a geochemical map for cadmium in stream plants in Sweden. The high cadmium concentrations in the central area of the country reflect metal mining in the area.

The most important ores of several base metals such as lead, zinc, and copper are sulfide minerals. The sulfide ore minerals represent the most serious threat of environmental contamination, because they are fairly easily oxidized in the presence of air to the considerably more soluble sulfates such as:-

$$ZnS + 2O_2 \rightarrow ZnSO_4$$

As a result surface and groundwaters in the vicinity of the weathering sulfide minerals can be seriously impacted.

A particular problem concerning weathering of sulfide minerals is that of pyrite and marcasite (both FeS_2). These iron sulfides oxidize to give various iron oxides and hydroxides together with some sulfates; these oxidation products are collectively called ocher. In addition to ocher,

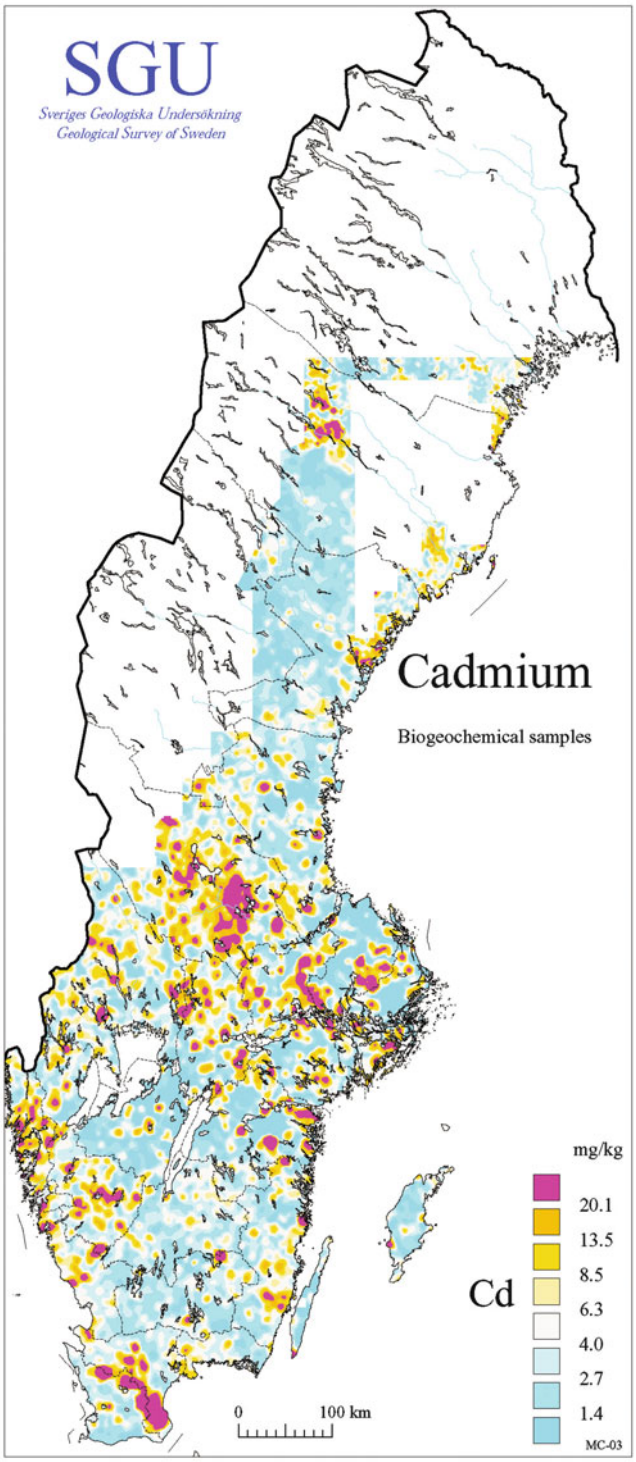


Fig. 4.1 Geochemical map for cadmium in stream plants in part of Sweden. (Reproduced with permission of the Swedish Geological Survey.) The high cadmium concentrations in south central Sweden are due to anthropogenic sources, those on the east coast are due to long range aerial deposition, those in the central area are derived from metal mining, and those in the central eastern area are derived from fertilizer. The high cadmium concentrations in the extreme south of the country are natural and are derived from cadmium-rich sediments

a by-product of the oxidation process is sulfuric acid; the resultant runoff from mines and waste piles is called acid mine drainage or acid rock drainage. Such acidic solutions can chemically attack other ore minerals and rocks to produce a cocktail of elements that can have a serious environmental impact on receiving rivers and streams. The reactions resulting in acid drainage can be summarized in the following equations:

$$FeS_2 + \frac{7}{2}O_2 + H_2O \rightarrow FeSO_4 + H_2SO_4$$

$$2FeSO_4 + H_2SO_4 + \frac{1}{2}O_2 \rightarrow Fe_2(SO_4)_3 + H_2O$$

$$FeS_2 + Fe_2(SO_4)_3 + 2H_2O + 3O_2 \rightarrow 3FeSO_4 + H_2SO_4$$

The ferrous iron produced is oxidized to ferric iron and hydrolysis of this results in ocher precipitation:

$$4Fe^{2+} + O_2 + 4H^+ \rightarrow 4Fe^{3+} + 2H_2O \quad \text{followed by}$$
$$Fe^{3+} + 3H_2O \rightarrow Fe(OH)_3 + 3H^+$$

Recently, Kimball et al. (2010) demonstrated that chalcopyrite, the major ore mineral of copper, can produce very significant acidity when oxidized by ferric iron:

$$CuFeS_2 + 16Fe^{3+} + 8H_2O \rightarrow Cu^{2+} + 17Fe^{2+} + 2SO_4{}^{2-} + 16H^+$$

Most of the world's economic source of copper is that occurring in Porphyry copper deposits, where chalcopyrite is finely disseminated throughout a large volume of rock. Such deposits are mainly extracted using open pit mining operations, with huge volumes of rock being extracted, and result in very large waste piles. As a result of the acidity generated by chalcopyrite and associated pyrite, waste tips are serious potential sources of acid mine drainage and disused open pits frequently fill up with highly acidic waters. Rice and Herman (2012) have estimated that globally a minimum of 4×10^6 tonnes of H^+ are produced annually from the extraction of copper.

Acid mine drainage from metalliferous mines is generally strongly enriched in potentially harmful elements. In acid mine drainage impacted rivers, lakes, and estuaries the contaminants can cause extreme damage to the biosphere. This is exacerbated by the precipitation of ocher, which is enriched in metals such as cadmium, copper, zinc, and aluminum, and metalloids such as arsenic. The ocher is frequently fine enough to be ingested by fish. With decreasing acidity which results from dilution and neutralization of the acid by interaction with rocks, the concentration of contaminant metals in the ocher increases. This results in sediments in the impacted streams becoming heavily enriched in potentially harmful elements.

In addition to its occurrence in metalliferous ore deposits, pyrite is frequently associated with coal, with drainage from coal mines and associated waste piles being heavily ocherous, but generally not as strongly enriched in environmentally harmful elements.

As stated above, historical mineral extraction processes were much less efficient than current methods and were not subject to strict environmental regulation. In many orefields where mining occurred in river valleys, the plentiful supply of water was used to separate ore minerals by gravity. As a result of this inefficient process, considerable amounts of highly contaminated sediment and fine ore grains were released into rivers. Due to subsequent flooding, they were also released onto the floodplains of the river valleys concerned. Thus many floodplain soils in old mining areas are highly enriched with contaminant metals.

One problem concerning gold mining, which deserves a special mention, is that of associated mercury pollution. Mercury has long been used to extract gold from ores and gold-rich sediments due to its ability to amalgamate with gold. As a result of this process many areas where gold has been extracted, particularly from river sediments, have suffered serious mercury contamination.

4.2.2 Smelting and Refining

Following the mining and processing of the ores, the resultant concentrate is transported to a smelter. Pyrometallurgical smelting involves roasting of the ore concentrate at high temperatures with the consequent emission of large quantities of potentially harmful elements. The smelter emissions can be in the form of gases, such as sulfur dioxide, aerosols, and larger particulates. Modern smelter stacks are fitted with electrostatic precipitators and other dust recovery mechanisms, which results in the retention of most of the particulates, but some gaseous and aerosolic emissions are still released into the atmosphere. Although any larger particulates released are likely to be deposited close to the source, aerosols and gases can be transported long distances and as a result the smelting of ores has far wider aerial impact than the mining and processing of these ores.

As many of the ores are sulfides and the smelting process is designed to be oxidizing, the gases sulfur dioxide (SO_2) and sulfur trioxide (SO_3) are produced. These gases react with water vapor in the atmosphere to produce sulfuric acid (H_2SO_4), which results in acid rain with potentially serious consequences for the environment.

During the smelting of metalliferous ores many metals and metalloids are released to the environment. The elements released depend on the ore being smelted and

include antimony, arsenic, bismuth, cadmium, chromium, cobalt, copper, lead, mercury, nickel, thallium, selenium, and zinc. Extremely high concentrations of some of these elements have been recorded in close proximity to the smelters. Rieuwerts and Farago (1996) showed that soils within 500 m of a lead smelter in the Czech Republic contained up to 3.73% lead and 2.76% zinc. It has also been shown that high concentrations of elements can be transported considerable distances from the smelter. Reimann et al. (1998) demonstrated that nickel from smelters in the Kola Peninsula of Russia was transported over 100 km in the direction of the prevailing wind. These same authors state that one of the smelters in the Kola Peninsula (Monchegorsk) emitted 1,619 tonnes of nickel during 1994.

While present day smelters emit potentially harmful elements, areas of past smelting operations also continue to be sites of environmental contamination. In the area of the lower Swansea Valley, South Wales, UK, a major smelting center for copper from 1717 until 1925, soils contain up to 200 mg kg^{-1} copper with the affected area extending 25 km north and 20 km east, the prevailing wind direction being from the southwest. The contamination is thought to have resulted from smelter fumes and wind ablation of smelter waste piles. More recently, in the same area, extreme selenium and tellurium contamination has been identified around a site of a nickel refinery with concentrations of selenium in soils ranging up to 200 mg kg^{-1} and tellurium up to 11 mg kg^{-1} (Perkins 2011).

Many specialist refining facilities have also been shown to cause environmental contamination. Perhaps the best documented is the release of fluorine during aluminum production. Aluminum is derived from alumina (aluminum oxide), which is recovered from bauxite ore by high temperature electrolysis with the alumina in a fluorine-rich electrolyte. The fluorine released during the process has been found to cause toxicity in plants in the vicinity of the refineries, and high fluorine occurs in the skeletons of animals in the same areas.

Smelter contaminants have been shown to have a detrimental effect on human and animal health. Copper toxicity was found to occur in cattle in the vicinity of a copper smelter in Mpumalanga (eastern Transvaal), South Africa (Gummow et al. 1991).

4.3 Power Generation

4.3.1 Fossil Fuel

Globally, fossil fuel (coal, oil, peat) combustion provides most of the power generated for industrial and domestic use. Burning of these fuels has achieved notoriety in recent years due to the large volumes of carbon dioxide (CO_2) produced, the consequent buildup of this gas in the atmosphere, and its possible contribution to the greenhouse effect on the Earth. In addition, combustion of high sulfur-containing fuels in some areas has resulted in production of sulfur dioxide and sulfur trioxide that, as in the case of smelter emissions (see Sect. 4.2.2 above), results in acid rain.

An additional environmental consequence of the use of fossil fuels for power generation derives from the many trace elements contained in the fuels. Although fossil fuels are predominantly made up of organic matter—the combustion of the carbon in these fuels being the source of energy—they also contain variable amounts of inorganic constituents retained in the ash left after the combustion process with some emitted in fine combustion products into the atmosphere. The ash residue, called fly ash, can contain many potentially harmful elements and therefore needs to be carefully disposed of (see Sect. 4.5). However, some of the inorganic components are emitted during combustion and can impact the soils, water, and biosphere in the vicinity of the power plant.

Many trace elements have been detected in fossil fuels. The actual concentrations of individual elements are variable and depend on the source of the fuel. However, it has been suggested that in Europe oil and coal combustion contribute significantly to atmospheric deposition of arsenic, cadmium, chromium, copper, nickel, and vanadium (Rühling 1994). Coal combustion is thought to have made a significant contribution to atmospheric lead deposition in the UK, with the coals containing up to 137 mg kg^{-1} (Farmer et al. 1999). It has also been suggested that coal combustion is the major anthropogenic source of selenium in the environment, and United States coals contain up to 75 mg kg^{-1} (Coleman et al. 1993). Enhanced concentrations of uranium in many coals have resulted in enrichments of this element around coal-fired power stations. Elevated concentrations of mercury occur in some oils with values up to 72 mg kg^{-1} recorded (Al-Atia 1972). Elements enriched in oils such as vanadium have also been found to be elevated in the environment in the vicinity of oil refineries (Rühling 1994).

The occurrence of fluorine in coal has been shown to impact the biosphere. Fluorine is highly phytotoxic and combustion of fluorine-rich coals has caused toxicity in vines downwind of a power station in New South Wales, Australia (Leece et al. 1986). Dental fluorosis occurring in cattle in the vicinity of large power stations in Yorkshire, England, was suggested by Burns and Allcroft (1964) to be due to the use of high-fluorine coals. Human fluorosis from burning fluorine-rich coals as a domestic fuel occurs in Guizhou Province, China (Zheng and Hong 1988).

Arsenic is enriched in coals from various parts of the world; however, in Guizhou Province, China, extreme

enrichments have been found with up to 3.5% arsenic found in some samples (Ding et al. 2001). More than 3,000 cases of arsenic poisoning have occurred from the combustion of these coals for domestic heating (Ding et al. 2001). Abandoned coal gasification sites are another source of contamination related to fossil fuels. Such sites have residual waste piles that are sources of elements such as arsenic, cadmium and copper, and cyanide and organic compounds such as phenols and tars which may have leaked into the subsoil.

4.3.2 Nuclear

Nuclear power generation has been utilized since the mid-1950s and accidental leakages and permitted effluent releases have impacted the environment. The nuclear industry is now strictly regulated, but in the early years this was not so and authorized discharges of radioactivity were considerably larger. Radionuclides released during these early years still pose a problem. For example, radioactive elements such as americium and plutonium released from Sellafield nuclear power station in Cumbria, northwestern England, are still retained in nearby marshy areas.

Although much of the contamination released from nuclear power plants affects only the immediate environment of the nuclear installation in question, the catastrophic explosion at Chernobyl in April 1986 caused widespread contamination, which seriously affected the Ukraine and Belarus with radioactivity spreading over much of Europe and many other parts of the world. Some of the more important radioactive species released in the explosion are listed in Table 4.3. Radioactive cesium from the Chernobyl accident rained out over upland areas of the UK and high concentrations were found in sheep in the area. The accident at the Fukushima nuclear plant in Japan resulting from an earthquake and subsequent tsunami, in March 2011, resulted in major release of several radioactive species. Estimates suggest that the amount of ^{137}Cs released was about 42% of that released from Chernobyl, while the release of ^{133}Xe was the largest non-bomb release of radioactive noble gas in history (Stohl et al. 2011). About 4/5 of the radioactive cesium was deposited in the sea while about a fifth was deposited on Japan. However, radioactive iodine released from Fukushima Dai-Ichi was detected in the western USA within 4 weeks of the disaster.

Anthropogenic radioactivity in the environment poses a serious threat to human health. Of particular concern is radioactive iodine, mainly ^{131}I, which has been found to move through the food chain rapidly. As a result of exposure to radioactive iodine, humans are prone to increased incidences of thyroid cancer, as evidenced in the aftermath of Chernobyl (see Soils and Iodine Deficiency, Chap. 17, this volume).

Table 4.3 Some of the more important radioactive isotopes released by the chernobyl accident

Isotope	$T_{1/2}$
Relatively short half-life isotopes	
^{132}Te	78 h
^{133}Xe	5.27 days
^{131}I	8.07 days
^{95}Nb	35.2 days
^{89}Sr	52 days
Isotopes with longer half-lives	
^{134}Cs	2.05 years
^{85}Kr	10.76 years
^{90}Sr	28.1 years
^{137}Cs	30.23 years
Actinides (some of the more problematic)	
^{241}Pu	13.2 years
^{244}Pu	8×10^7 years
^{241}Am	458 years

4.3.3 Geothermal

Geothermal energy has often been assumed to be a "clean" form of power generation. However, many geothermal areas are associated with volcanic activity and many of the hot springs actively precipitate arsenic, antimony, mercury, and thallium, whereas some geothermal waters contain very high concentrations of boron. Thus spent waters from geothermal production in areas such as New Zealand contain very high concentrations of arsenic (see Arsenic in Groundwater and the Environment, Chap. 12 this volume), which pose serious environmental problems for the receiving rivers.

4.3.4 Hydroelectric

Hydroelectric power generation has led to problems resulting from flooding of areas for dams. In flooded areas where soils have been inundated it has been found that fish frequently contain elevated concentrations of mercury. The source of the mercury has been found to be the waterlogged soil where this element becomes converted to a methylated form which is bioavailable (see also Sect. 4.1 of this chapter).

4.4 Other Industrial Activities

There are a large number of other industrial activities that have the potential to cause environmental contamination. In this section only a small selection typifying such processes is discussed.

4.4.1 Metallurgical

Many metals and metalloids are used in the manufacture of alloys, with steel being the major product. The manufacture and recycling of steel results in the release of many elements into the atmosphere with subsequent deposition and contamination of the local environment. Figures obtained for atmospheric pollutants in the north of the former Soviet Union show that 79% of the chromium, 76% of the manganese, and 75% of the zinc together with significant quantities of antimony, arsenic, cadmium, lead, nickel, and vanadium were derived from the iron and steel industry (Pacyna 1995). More modern electric arc furnaces processing steel scrap give rise to a large amount of metal-rich flue dusts.

In the proximity of a steel plant in Mpumalanga (eastern Transvaal), South Africa, chronic vanadium toxicity occurs in cattle as a result of air pollution from the steelworks (Gummow et al. 1994). Another health problem related to steel production could possibly stem from the use of the mineral fluorite (calcium fluoride) as a flux in the process. An occurrence of fluorosis in cattle in northern England was suggested by Burns and Allcroft (1964) to be due to fluorine releases from local steel plants. In addition the waste, or slag, from steel plants can contain many metals that can have a detrimental effect on the environment, and this needs to be disposed of appropriately. Some slags have been used in fertilizers and Frank et al. (1996) reported vanadium poisoning in cattle after grazing on pasture land that had been amended with basic slag.

4.4.2 Brick and Pipe Manufacture

Bricks and pipes are very important in the construction industry. Both are produced from naturally occurring clays and clay-containing rocks. Manufacture of these materials involves shaping the clays and then roasting them at temperatures up to 1,200°C. As a result of this roasting, elements, which are contained in the clay minerals such as copper, lead, and zinc, are released to the environment (Fuge and Hennah 1989). However, a more significant health problem is the release of fluorine, which is frequently enriched in the minerals of the clays and rocks used during the manufacture. Very serious fluorosis in farm animals occurring in the East Anglia region of the UK in the 1950s was due to the extensive brick manufacture in this area (Burns and Allcroft 1964).

4.4.3 Cement Manufacture

Cement is also manufactured by a process involving the roasting of clay-rich rocks. A mixture of limestone and shale rocks is heated in a furnace at temperatures around 1,450°C. Shale is a clay-rich rock and can contain elevated concentrations of potentially harmful elements, and as in brick and pipe manufacture, these can be released at high temperatures.

4.4.4 Others

Many other industrial activities have been shown to result in environmental contamination. The production of enamel, ceramics, and glass releases appreciable concentrations of fluorine into the environment (Koritnig 1972). Another industry shown to release large quantities of fluorine into the environment is the manufacture of phosphate fertilizer from the fluorine-containing mineral apatite [$Ca(PO_4)_3(OH,F,Cl)$] (Dabkowska et al. 1995).

Of the many other industrial processes that impact the environment, it is perhaps relevant to mention the serious impact of chromium in the environment from tanning plants (Davis et al. 1994). In addition, the chlor-alkali industry, which manufactures sodium hydroxide and chlorine from seawater by electrolysis with mercury electrodes has resulted in serious environmental contamination of the atmosphere and waterways with mercury. Spillages and leakage from the electrodes is also an environmental problem. Although the chlor-alkali process is being phased out, disused chlor-alkali works are a continuing source of contamination.

4.5 Waste Disposal

Large quantities of waste materials are produced on a daily basis from urban and industrial sources. Although there is a move to recycle and reuse waste products, the vast majority of waste has to be treated and subsequently disposed of. Many of the waste products created in modern society have a detrimental effect on the environment.

4.5.1 Refuse

Most household refuse together with waste materials from industrial sources is consigned to landfill sites. Such landfill sites produce leachate that contaminates soils, surface waters, and groundwaters in the vicinity. Landfill leachate chemistry is variable and depends on the components of the landfill plus varying conditions within the buried material, such as the differing redox conditions of various parts of the landfill and the percolation and flow of water into and through the site. Modern landfill sites are lined with low permeability materials such as high-density plastics to eliminate leachate leakage. However, old landfill sites made use

of suitable depressions, such as disused quarries, and were not lined. These sites have become serious sources of contamination. Because landfills can contain a wide range of waste materials, many elements and organic compounds can be released, which include toxic metals such as cadmium and mercury and substances such as ammonia, nitrate, and sulfate. Due to the generally anaerobic nature of landfills, methane gas is generated and in modern facilities this is collected and used as fuel.

Incineration is the other commonly used method of refuse disposal. This process, which is utilized to provide a source of energy in some instances, results in the production of ash that contains many elements that are potentially harmful to the biosphere so necessitating disposal in a landfill. Additionally, incinerators can vent various elements and organic compounds such as dioxins into the atmosphere which impact local soils and waters. In this context it is perhaps relevant to note that uncontrolled burning of refuse, particularly in the developing world, and the consequent emissions into the atmosphere are of particular concern because they are likely to increase in the future.

4.5.2 Fly Ash

The residue from the combustion of coal is called fly ash and large quantities are created which result in disposal problems. This ash is composed of finely powdered glass-like particles and is highly reactive. It contains high concentrations of many potentially harmful elements, some of which could be readily mobilized in the environment. The composition of fly ash depends on the composition of the original coal, and it can contain relatively high concentrations of arsenic, barium, boron, cadmium, chromium, copper, fluorine, germanium, lead, molybdenum, nickel, selenium, and zinc together with high quantities of sulfur in the form of the sulfate anion.

Generally, fly ash is disposed of as a water-based slurry into lagoons or as a dry powder in landfills. In the case of slurry disposal, water from the lagoon is drained to local rivers after treatment and dilution. The soluble components of the ash are then transferred to local drainage systems.

A modern approach to fly ash disposal is to reuse it economically. Thus fly ash has been used in the manufacture of blocks for use in building; it has also been added to agricultural soils to neutralize acidity and to add essential elements to the soils.

4.5.3 Sewage

Sewage effluent is transported to treatment plants where the solid material, sewage sludge, is separated and the remaining liquid portion is discharged into rivers and seas. The liquid effluent can greatly impact the aqueous environment. The effluent can be enriched in nitrate, which results in algal blooms and eutrophication in coastal waters (see Sect. 4.6 for discussion of problems of eutrophication). This problem has been exacerbated due to effluent enriched in phosphate due to the addition of detergents. These additions have since been limited or banned in many countries.

Although liquid effluent can cause serious environmental problems, it is the solid waste or sludge that is the larger problem. Sewage sludge is produced in very large quantities and its disposal can have serious environmental implications. The methods used for its disposal include application to agricultural land (see Sect. 4.6), burial in landfills, and incineration. Disposal at sea, which was once a common practice, has been banned in most countries since the late 1990s. While several methods of sludge disposal are used, it is its use on agricultural land that causes the greatest concern.

The composition of sewage sludge is variable because it is derived from many sources such as domestic effluent, surface runoff that includes soil and street dust, and from different industrial effluents. Thus sludge derived from a predominantly urban area will be very different compositionally from that derived from highly industrialized areas. The composition of sewage sludge is variable and it has been found to contain metals such as cadmium, chromium, copper, lead, mercury, molybdenum, nickel, and zinc along with elements such as arsenic and fluorine and several organic contaminants such as polyaromatic hydrocarbons (PAHs) and chlorophenols. Nicholson et al. (2003) noted that in the UK controls on discharges from industrial sources mean that domestic sources of metals have become relatively more important. Thus copper and zinc from water pipes (see Sect. 4.10) and elements such as zinc and selenium, which are contained in many shampoos, are mainly derived from urban sources.

The impact of sewage sludge application to agricultural land is discussed in Sect. 4.6.

4.5.4 Nuclear

The disposal of nuclear waste is a relatively modern issue that is and will continue to be a serious problem for mankind. This waste is classified as high-, medium-, and low-level waste depending on its radioactivity. The favored method for disposal of high- and medium-level waste is long-term storage in underground vaults, and the potential for leakage and transport from these is currently the subject of much research.

Medium- and high-level waste needs to be isolated in vaults, but low-level waste is frequently disposed of in

surface locations. It is of note that drainage from one such site in northwestern England, which is the low-level waste disposal site for the nuclear installation at Sellafield, contains 61 μg L^{-1} uranium with associated stream sediments containing 48 mg kg^{-1} uranium, far higher concentrations than are found in surrounding streams (British Geological Survey 1992).

4.6 Agricultural Practices

The well-documented environmental contaminants from agricultural practices are the nitrate and phosphate derived from fertilizer and farmyard manure that are applied to soils. These soluble ions are washed into surface and groundwaters. Nitrate and phosphate are essential nutrients and in lakes and coastal marine waters they can cause algal blooms and ultimately eutrophication when massive algal growths cause all of the oxygen in the water to be depleted. This results in the death of the algae and other biota, which leads to masses of rotting organic matter. In addition, the leaching of nitrates into groundwater in some agricultural areas has caused problems when the groundwaters are sources of potable waters. Such nitrate-rich waters have been shown to have caused methemoglobinemia in young babies due to nitrate interfering with the iron in blood which results in defective transport of oxygen. This is the cause of so-called "blue blood," and methemoglobinemia is sometimes referred to as "blue baby syndrome." It has also been suggested that ingestion of nitrates can contribute to gastric cancer when some of the nitrates are converted to nitrosamines, which are thought to be carcinogens (Mirvish 1991).

In addition to the problems of nitrate and phosphate derived from fertilizers, many other contaminants are added to soil, surface, and groundwaters due to agricultural activities. Thus many substances applied to soil as fertilizers contain trace elements known to be detrimental to the biosphere and potentially toxic to humans. These substances can accumulate in soils and be leached into local waters (Table 4.4). Of the inorganic fertilizers used, phosphate fertilizers manufactured from apatite-rich rocks are of particular interest because they can contain elevated concentrations of several potentially harmful elements. High cadmium in some soils and surface waters in agricultural areas has been linked to the application of phosphate fertilizer, as shown in Fig. 4.1 by stream plant chemistry for an area of Sweden, where the high concentrations in the east central part of the country reflect fertilizer sources. Phosphate fertilizers have also been linked to high uranium concentrations in waters draining agricultural areas and have been found to be the source of elevated concentrations of the element in surface peats and waters in the Florida Everglades (Zielinski et al. 2000). It should be mentioned that modern phosphate fertilizers contain less contaminant metals than the earlier varieties such as superphosphate.

Table 4.4 Concentrations of some trace elements in fertilizers added to agricultural land (mg kg^{-1})

	Phosphate fertilizer	Nitrate fertilizer	Manure	Sewage sludge
As	1–1,200	2–120	3–25	2–30
Cd	0.1–190	0.05–8.5	0.1–0.8	<1–3,400
Cr	66–245	3.2–19	1.1–55	8–41,000
Cu	1–300	–	2–172	50–8,000
Hg	0.01–2	0.3–3	0.01–0.4	<1–55
Mo	0.1–60	1–7	0.05–3	1–40
Ni	7–38	7–34	2–30	6–5,300
Pb	4–1,000	2–27	11–27	30–3,600
Se	0.5–25	–	0.2–2.4	1–10
U	20–300	–	–	<2–5
V	2–1,600	–	–	–
Zn	50–1,450	1–42	15–570	90–50,000

Data from several sources but predominantly from Alloway (1995)

The application of sewage sludge has been shown to be a source of elemental and organic contamination of agricultural soils. As mentioned in Sect. 4.5.4, one of the methods of disposal of sewage sludge is its application to agricultural land as a fertilizer. The composition of sewage sludges is extremely variable depending on their geographical source, and their application to soils is generally controlled with limits that are a function of their elemental compositions. However, some sludge applied to agricultural land has, in the past, resulted in health effects in livestock; for example, Davis (1980) found that an outbreak of fluorosis in cattle on a farm in the UK was due to the application of sewage sludge containing 33,000 mg kg^{-1} of fluorine.

The composition of livestock manure is also variable with its elemental composition essentially derived from their food and water intake. However, in some cases feeds are supplemented with metals, as in pig and poultry diets where copper and zinc are added; thus, pig and poultry manure are particularly rich in these two metals. In the past arsenic was added to pig diets, but this has recently ceased. It is worth noting also that both sewage sludge and animal manure can add pathogens to soil and subsequently to surface and groundwaters.

From a study of metal inputs to agricultural soils in the UK, Nicholson et al. (2003) suggested that of the fertilizer sources, livestock manure represents the greatest source of zinc and copper. Sewage sludge was the most important source for lead and mercury, and inorganic fertilizers (mainly phosphate fertilizer) were the major source of cadmium. However, Nicholson et al. (2003) concluded that the major input of metals to agricultural soils in the UK was atmospheric deposition (see Sect. 4.8).

The application of pesticides—insecticides, herbicides, and fungicides—contributes to contamination of soils and surface and groundwaters. Many organic and inorganic compounds are utilized in pesticides. Some of the insecticides are chlorinated hydrocarbons (organochlorines) that include DDT and lindane, both of which were found to be toxic to humans. The use of these insecticides has been banned for some years, but they still persist in the environment as they decay slowly. Other insecticides such as organophosphates have replaced the organochlorines and, together with other organic compounds, are extensively used. There are a large number of organic herbicides and fungicides that show varying degrees of persistence in the environment ranging from a few weeks to many years. Many of these compounds have been found in groundwater, some of which is used as a source of drinking water. Several pesticides are based on inorganic compounds containing elements such as arsenic, lead, manganese, and zinc. Copper salts are used extensively as pesticides, for example, the fungicide Bordeaux mixture.

Another agricultural practice that can impact soils is the addition of lime, which is used to reduce acidity. Lime derived from limestone can contain many trace elements that can influence the composition of the soils. However, the major influence of lime is in increasing soil pH, which in turn can reduce the uptake of some elements by plants. Lime is sometimes added to soils to limit uptake of potentially harmful elements; thus, in sewage sludge amended soils, where high concentrations of some metals occur in the sludge, liming of soils is used to limit the uptake of metals. In some cases liming of soils has a detrimental effect because it limits plant uptake of elements that are essential to grazing animals. Some animals have suffered copper deficiency when grazing on recently limed pasture. In addition, copper deficiency is exacerbated if there are high amounts of molybdenum in the diet, which can arise from liming soils. This increases alkalinity (pH) and enhances molybdenum uptake by plants. Molybdenum-induced copper deficiency (molybdenosis) has been suggested to be the cause of unexplained moose deaths in Sweden, which have occurred since the mid-1980s in an area that had previously been heavily limed (Frank 1998) (see also Chap. 21, this volume).

Many other agricultural processes impact soils locally, for example, the application of steelworks slag (see Sect. 4.4.1) and fly ash (see Sect. 4.5.2). The use of wood preservative, both in agricultural and urban areas, can add elements such as copper, chromium, and arsenic to soils and runoff. In addition, metal fences and gates are frequently coated with zinc (galvanized) and can act as a source of this metal in the environment. In arid areas the practice of irrigation can sometimes lead to problems. High selenium in waters in the San Joaquin Valley, California, which has lead to toxicity in birds in the area, was found to be due to leaching of natural selenium from soils by irrigation waters (Kharaka et al. 1996). Kharaka et al. (1996) also suggested that several other western states in the United States have similar leaching problems with selenium-rich irrigation waters.

A particularly interesting problem resulting from an agricultural practice has been manifested in Brazil. Elevated mercury concentrations occurring in tributaries of the Amazon were originally ascribed to artisanal gold mining in river sediments, which result from the use of mercury to amalgamate gold (see Sect. 4.2.1). However, it has subsequently been shown that the mercury in the rivers is derived from soil as a result of deforestation.

4.7 Transportation-Derived Contamination

It has been suggested that motor vehicles represent the greatest single source of atmospheric contamination. Motor vehicles are powered by gasoline or diesel, and their combustion in vehicle engines results in the production of exhaust gases in which there are large quantities of carbon dioxide. However, due to incomplete combustion, carbon monoxide, hydrocarbons, and nitrogen oxides are also evolved. These gases pollute the urban atmosphere and are thought to contribute, along with emitted particulates, to respiratory diseases. In addition the nitrogen oxide gases destroy ozone in the stratosphere, which aids in thinning of the ozone layer.

Vehicle exhaust gases can also contain metallic elements such as lead, manganese, nickel, and vanadium. Nickel and vanadium are derived from diesel fuel and these metals can be very rich in diesel exhaust fumes, which represent a major component of atmospheric nickel and vanadium loads. Roadside soils have been found to be markedly enriched in nickel, and diesel locomotives can emit significant quantities of nickel and vanadium.

Lead has been added to gasoline since the early 1920s as tetraethyl and tetramethyl lead to make its combustion more efficient. However, research showed that the lead emitted through the exhaust system had a detrimental effect on the environment with estimates that over 75% of environmental lead was derived from this source. As a result the use of lead-containing fuel has been phased out in many countries. This is not true of all countries and gasoline with lead additives is still used in many developing countries. Because of the phasing out and subsequent ban on the use of leaded gasoline in the UK, lead concentrations in the atmosphere

Table 4.5 Vehicular sources of metals

Metal	Source
Lead	Gasoline
Manganese	Gasoline
Nickel	Diesel + alloys
Vanadium	Diesel + alloys
Zinc	Tires + galvanized items
Cadmium	Tires + lubricating oils (minor amounts)
Chromium	Chrome plating, brake linings, etc.
Copper	Electrical wiring, thrust bearings, etc.
Platinum groupmetals	Catalytic converters

have decreased markedly. Although lead concentrations in roadside soils have also decreased, these soils together with road dusts still contain high concentrations that continue to be pathways into the biosphere. In some countries where lead additives have been banned methylcyclopentadienyl manganese tricarbonyl (MMT) is used as an alternative which causes emissions of manganese, a suspected neurotoxin, in exhaust emissions.

Many other pollutant metals in roadside soils are derived from motor vehicles (see Table 4.5). Zinc is added to the rubber used for tires and some steel components of cars are coated with zinc (galvanized) to minimize rusting. Cadmium, a component of zinc ores (see Sect. 4.2.1), finds its way into tires and is also used as an antioxidant in lubricating oils. Chromium is used in steel alloys and in chrome plating, and nickel is also used in some steel alloys, which provides a second potential source of nickel. Motor vehicles are also thought to be a source of copper pollution, it being released from copper wiring, thrust bearings, and brakes. Similarly, metals such as tungsten used in high-temperature alloys in turbojet components and airplanes may be found in the vicinity of airfields.

A fairly modern group of pollutant metals derived from motor vehicles are the platinum group elements mainly platinum, palladium, and rhodium, which have been used in catalytic converters. These converters were first introduced in the 1970s in Japan and the United States and subsequently in other countries to clean up motor vehicle exhaust. The catalytic converter essentially converts carbon monoxide to carbon dioxide, nitrogen oxides to nitrogen, and hydrocarbons to carbon dioxide and water vapor. These converters were originally assumed to be fairly indestructible, but a study by Hodge and Stallard (1986) showed that roadside dust in San Diego, California, was enriched in platinum and palladium. Subsequently there have been strong enrichments of the platinum group elements in roadside dust, urban river sediments, roadside drain sediments, sewage sludges, etc. The concentration of these elements in some of these sediments is similar to that found in platinum group metal ores.

4.8 Atmospheric Deposition of Contaminants

Many contaminants are released into the atmosphere and carried as gases, aerosols, and particles, and they are subsequently deposited on the surface where they are incorporated into soils and waters, absorbed by plants, or inhaled by animals and humans. Contaminant sources augment the many natural sources of elements and compounds contributing to the atmospheric load, such as volcanic and hot spring activity, forest and grassland fires, and wind-ablated soil particles. The sources of the atmospheric pollutants are listed in the proceeding sections of this chapter (Sects. 4.2, 4.3, 4.4, 4.5 and 4.7) but include mining, smelting and refining, power generation, various industrial processes, waste incineration, and transportation related activities. In addition, fine soil particles carried into the atmosphere can have contaminants adsorbed on their surfaces.

The degree of atmospheric contamination is geographically variable and depends markedly on meteorological conditions and potential sources. The degree of transport of contaminants varies depending on the form in which they are held. Many contaminants from vehicles are carried as fairly large particles and are not carried very far from their source, which accounts for much of the contamination of road dusts (see Sects. 4.7 and 4.9). However, gaseous and aerosol transport can carry contaminants very great distances, and such forms of atmospheric transport are the major causes of contamination of the Arctic regions where metals such as lead have been found to be enriched in ice cores.

Deposition of atmospheric metals and arsenic in Europe has been estimated using moss analysis (Rühling 1994). This has demonstrated the importance of various industries and power generation on the distribution of the elements, with coal and oil combustion and mining and smelting accounting for the majority of the deposition of arsenic, cadmium, and copper and much of the nickel and zinc. Almost all of the vanadium is derived from coal and oil combustion and oil refining, whereas the steel industry accounted for most of the chromium and iron and some of the nickel. Lead was also derived from industrial and power generation sources, but vehicular contamination was found to be a major source. Pacyna (1995) estimated that in the heavily industrialized northern part of the former Soviet Union, fossil fuel combustion accounted for 75% of the lead in the atmosphere, 87% of the vanadium, and 34% of the selenium with 79% of the chromium and 75% of the zinc derived from the steel industry. Eighty-five percent of the arsenic, 53% of the cadmium, and 51% of the antimony came from the mining and smelting industries.

Deposition from the atmosphere is a major source of elements for the surface environments. Nicholson et al.

(2003) showed that for UK agricultural soils atmospheric deposition was the major source of several metals and accounted for 85% of the total mercury input, 53% of the cadmium, 55–77% of the arsenic, nickel, and lead, and 38–48% of the zinc.

In Fig. 4.1 the high cadmium on the west coast of Sweden reflects long-range atmospheric transport.

4.9 Contamination of Urban Environments

Contamination of the urban environment is of major importance because over half of the world's population live in this environment. Urban contamination is derived from many sources and motor vehicles are a major present-day source (see Sect. 4.7); however, many urban areas also house industrial activities with consequent contamination. Even in modernized cities in many countries, even though industries have moved out of the urban area, historical industrial activities have left large areas of contaminated land. Additionally, in many urban areas fossil fuels are still burned as sources of household heating. Contamination can also arise from fertilizers applied in parks and gardens, pesticide use, and garden fires and bonfires adding to atmospheric and soil pollution. A further source of urban contamination is the disposal of waste and refuse in gardens and open areas, and uncontrolled burning in the latter.

Many contaminants have derived from urban sources with many elements noted as enriched in urban soils relative to their rural counterparts (Table 4.6). Thus arsenic, antimony, boron, cadmium, chromium, copper, lead, mercury, molybdenum, nickel, tin, vanadium, and zinc, among others, have been listed as serious contaminants of the urban environment. The soils and street dusts of the urban environment are enriched in many potentially harmful metals (Table 4.6) and are the major sources of household dust. They represent a major pathway into the human body via inhalation and inadvertent ingestion from hands, which is a particular problem for young children. Although many of these elements are potentially harmful to humans, it is lead that has attracted the most attention because it has been shown to cause serious health problems in children from urban environments.

Lead in inner city areas is derived from a variety of sources. The use of leaded fuel in vehicles (see Sect. 4.2) either has been or is being phased out in most developed countries; however, it is still used in many less developed countries. Even in countries where leaded fuel is banned, much of the urban dust and soil lead is a legacy of its former usage (Table 4.6). Even though the major source of urban lead is likely to be the past or present use of leaded fuel, Mielke (1994) in a study of New Orleans, Louisiana, found that a major source of lead in soils of inner city areas was lead-containing paint. Although such paints are no longer used, the exteriors of older houses have layers of lead-based paint which have been painted over. Weathering of the paint or renovation of the houses where the old paintwork is scraped off results in the addition of lead to the soils. Old lead paintwork has also been shown to be a source of lead in children who ingest paint flakes that have been either picked off the paintwork or scraped off during renovation.

Table 4.6 Metals in the urban environment—Birmingham, England

Urban area—street dust	Range (mg kg^{-1})
Cadmium	0.4–25
Chromium	9–228
Copper	36–3,160
Lead	32–4,820
Nickel	11–683
Tin	3–332
Zinc	79–5,210
Urban area—soil	
Cadmium	0.7–1.6
Copper	38–715
Lead	75–350
Zinc	53–450
Rural area—soil	
Cadmium	0.2–0.5
Copper	5–57
Lead	14–74
Zinc	10–180

Data for street dust are from Brothwood (2001) and data for soils are from Davies and Houghton (1984)

Note: Data for dust was obtained with a strong acid attack and represents a virtually total extraction of metals. Soil data is for a weak acid attack and represents metals which are easily extractable

Lead is an extremely toxic element and attacks the central nervous system. Young children are particularly vulnerable. In extreme cases of lead toxicity, although now very rare, the brain swells (encephalopathy) and death can result. The obvious symptoms of lead toxicity are rare, but the subclinical effects are a cause for concern as they have been shown to impair learning and cognitive response. Surveys of children residing in inner cities have shown that relatively low intelligence quotients (IQ) correlate with high body burdens of lead. These levels are sometimes estimated from the lead content of the milk teeth of young children and monitored by blood lead levels. Studies have shown that exposure to Pb at an early age has a detrimental effect on educational achievement. It is also suggested that elevated lead in children's blood can increase the incidence of attention deficit hyperactivity disorder, while high blood lead in children/teenagers has been linked with antisocial and violent behaviour. Finally research has indicted that exposure to lead at an early age speeds the ageing process.

In addition to the elemental contaminants mentioned above, it is worth pointing out that many organic contaminants, such as dioxins and PAHs, are added to the urban environment from industry, fossil fuel combustion, and refuse burning.

4.10 Treatment and Transport of Potable Waters

Drinking water for humans comes from surface and groundwater sources and, hence, its initial chemistry is governed mainly by the chemistry of the surface sediments and rocks with which it interacts. The initial water is generally subjected to various treatments to ensure that it is fit for human consumption and this process modifies its composition. Subsequently, the water is distributed through a network of pipes that are frequently metallic; this transport can also modify the water composition.

Surface water frequently carries solid matter that needs to be removed. Although coarse material can be removed easily by filtering, finer material poses more of a problem and a coagulant is added to cause the flocculation and precipitation of this material. As a result of the coagulation process traces of the coagulant remain in the water. Several chemicals are used as coagulants and one of the most important is aluminum salts. However, concern has been expressed concerning elevated aluminum in drinking water, because some studies suggest that it is a contributory factor in Alzheimer's disease (premature senility). Iron compounds have also been used as coagulants, but their use can lead to subsequent strong leaching of metals from pipe surfaces.

Following removal of solid material the water is disinfected to remove any potentially harmful organisms. The most common disinfectant is chlorine. However, it has been shown that chlorine can react with organic matter in the water to produce chlorine-containing organic compounds (chlorophenols) that have been shown to be carcinogenic.

Following treatment the water is distributed through pipes, which are frequently metallic, and as a result traces of the pipe metal are transferred to the water. For example, where copper pipes are used copper and zinc are found in the water. The concentrations of these metals are generally low and Fuge and Perkins (1991) recorded <65 μg L^{-1} copper and <40 μg L^{-1} zinc for most samples analyzed in the main drinking water in north Ceredigion, Wales. However, in low-calcium containing (soft) waters from springs and wells where the drinking waters were untreated, much higher copper and zinc levels were present, with up to 500 μg L^{-1} of copper and 125 μg L^{-1} of zinc recorded.

Lead pipes have been extensively used in the past for transporting water, but in view of the known serious health effects of lead (see Sect. 4.9), this metal is no longer used and in many cases older lead piping has been replaced. However, due to its historical use some lead pipes are still in place and Fuge and Perkins (1991) record 29 μg L^{-1} lead in a supply with lead pipes. Lead in drinking water is highly bioavailable, and it has been suggested that this is a major pathway into the human body in some areas where lead piping is still in use. To this end the World Health Organization (WHO) suggests that drinking water contain no more than 10 μg L^{-1} of lead.

4.11 Summary

The geochemistry of environmental media is strongly dependent on the chemistry of the natural sources from which they have been derived or with which they have interacted. However, the chemistry of environmental materials demonstrates a superimposed anthropogenic signal. It has been suggested that no environment on Earth is free of contamination and even the polar ice caps have elevated lead concentrations. Anthropogenic impacts on the environment result from many human activities that range from the extraction of raw materials from the Earth and manufacturing the many products needed by society that provide the food and generate the power for the world's population and subsequent disposing of the resultant waste materials. With much of the world's population living in major conurbations, nowhere is the impact of anthropogenic activities more evident and much of this is derived from the use of motor vehicles.

The anthropogenic influence on the environment has resulted in many elements and inorganic and organic compounds impacting soils, waters, and the atmosphere. As humans and animals are dependent on these media for food, drinking water, and air, any contaminants can enter the biosphere and can potentially have a serious impact with resultant health problems. A summary of the major sources of elements that can have serious effects on the biosphere is given in Table 4.7.

See Also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements) • Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 21 (Animals and Medical Geology)

Table 4.7 Important anthropogenic sources of some elements known to have detrimental effects on the biosphere

Element	Sources
Antimony	Mining, smelting, fossil fuel combustion
Arsenic	Mining, smelting, steel making, fossil fuel combustion, geothermal energy production, phosphate fertilizer, pesticides
Cadmium	Mining, smelting, fossil fuel combustion, incineration, phosphate fertilizer, sewage sludge, motor vehicles
Chromium	Smelting, steel making, fossil fuel combustion, phosphate fertilizer, sewage sludge
Cobalt	Mining, smelting, fossil fuel combustion
Copper	Mining, smelting, fossil fuel combustion, manure, sewage sludge, pesticides
Fluorine	Mining, aluminum refining, steel making, fossil fuel combustion, brick making, glass and ceramic manufacture, phosphate fertilizer
Lead	Mining, smelting, fossil fuel combustion, sewage sludge, pesticides, motor vehicles
Mercury	Smelting, fossil fuel combustion, incineration, sewage sludge
Nickel	Mining, smelting, steel making, fossil fuel combustion, oil refining, sewage sludge, motor vehicles
Selenium	Smelting, fossil fuel combustion
Thallium	Smelting, fossil fuel combustion
Uranium	Fossil fuel combustion, phosphate fertilizer
Vanadium	Steel making, fossil fuel combustion, oil refining
Zinc	Mining, smelting, steel making, fossil fuel combustion, phosphate fertilizer, manure, sewage sludge, pesticides, motor vehicles, galvanized metal

References

Al-Atia MJ (1972) Trace elements in Iraqi oils and their geological significance. In: Mineral exploitation and economic geology. University of Wales, Aberystwyth

Alloway BJ (ed) (1995) Heavy metals in soils, 2nd edn. Blackie Academic & Professional, London

British Geological Survey (1992) Regional geochemistry of the Lake District and adjacent areas. British Geological Survey, Keyworth

Brothwood S (2001) Vehicle related emissions of platinum group elements and other heavy metals in the urban environment, PhD thesis, University of Wales, Aberystwyth

Burns KN, Allcroft R (1964) Fluorosis in cattle: occurrence and effects in industrial areas of England and Wales 1954–57, industrial disease surveys, reports 2, part 1. Ministry of Agriculture Fisheries and Food, London

Coleman L, Bragg LJ, Finkelman RB (1993) Distribution and mode of occurrence of selenium in US coals. Environ Geochem Health 15:215–227

Dabkowska E, Machoy-Mokrzynska A, Straszko J, Machoy Z, Samujlo D (1995) Temporal changes in the fluoride levels of jaws of European deer in industrial regions of Western Pomerania, Poland. Environ Geochem Health 17:155–158

Davies BE, Houghton NJ (1984) Distance-decline patterns in heavy metal contamination of soils and plants in Birmingham, England. Urban Ecol 8:285–294

Davis RD (1980) The uptake of fluoride by ryegrass grown on soil treated with sewage sludge. Environ Pollut B 1:277–284

Davis A, Kempton JH, Nicholson A, Yare B (1994) Groundwater transport of arsenic and chromium at a historical tannery, Woburn, Massachusetts, USA. Appl Geochem 9:569–582

Ding Z, Zheng B, Long J, Belkin HE, Finkelman RB, Chen C, Zhou D, Zhou Y (2001) Geological and geochemical characteristics of high arsenic coals from endemic arsenosis areas in Southwestern Guizhou province, China. Appl Geochem 16:1353–1360

Farmer JG, Eades LJ, Graham MC (1999) The lead content and isotopic-composition of British coals and their implications for past and present releases of lead to the UK environment. Environ Geochem Health 21:257–272

Frank A (1998) "Mysterious" moose disease in Sweden. Similarities to copper deficiency and/or molybdenosis in cattle and sheep. Biochemical background of clinical signs and organ lesions. Sci Total Environ 209:17–26

Frank A, Madej A, Galgan V, Petersson LR (1996) Vanadium poisoning of cattle with basic slag. Concentrations in tissues from poisoned animals and from a reference, slaughter house material. Sci Total Environ 181:73–92

Fuge R, Hennah TJ (1989) Fluorine and heavy metals in the vicinity of brickworks. Trace Subst Environ Health 23:183–197

Fuge R, Perkins WT (1991) Aluminum and heavy metals in potable waters of the North Ceredigion area, Mid-Wales. Environ Geochem Health 13:56–65

Fyfe WS (1998) Towards 2050; the past is not the Key to the future; challenges for science and chemistry. Environ Geol Health 33:92–95

Gummow B, Botha CJ, Basson AT, Bastianello SS (1991) Copper toxicity in ruminants: air pollution as a possible cause. Onderstepoort J Vet Res 58:33–39

Gummow B, Bastianello SS, Botha CJ, Smith HJC, Basson AT, Wells B (1994) Vanadium air pollution: a cause of malabsorption and immunosuppression in cattle. Onderstepoort J Vet Res 61:303–316

Hodge VF, Stallard MO (1986) Platinum and palladium in roadside dust. Environ Sci Technol 20:1058–1060

Kharaka YK, Ambats G, Presser TS, Davis RA (1996) Removal of selenium from contaminated agricultural drainage water by nanofiltration membranes. Appl Geochem 11:797–802

Kimball BE, Rimstidt JD, Brantley SL (2010) Chalcopyrite dissolution rate laws. Appl Geochem 25:972–983

Koritnig S (1972) Fluorine. In: Wedepohl KH (ed) Handbook of geochemistry. Springer, Berlin, chap. 9

Leece DR, Scheltema JH, Anttonen T, Weir RG (1986) Fluoride accumulation and toxicity in grapevines *Vitus vinifera* L. in New South Wales. Environ Pollut A 40:145–172

Levinson AA (1980) Introduction to exploration geochemistry, 2nd edn. Applied Publishing, Wilmette

Mielke HW (1994) Lead in New Orleans soils: New images of an urban environment. Environ Geochem Health 16:123–128

Mirvish SS (1991) The significance for human health of nitrate, nitrite and N-nitroso compounds. In: Bogárdi I, Kuzelka RD, Ennenga

WG (eds) Nitrate contamination exposure, consequences and control. Springer, New York, pp 253–266

Nicholson FA, Smith SR, Alloway BJ, Carlton-Smith C, Chambers B (2003) An inventory of heavy metals inputs to agricultural soils in England and Wales. Sci Total Environ 311:205–219

Pacyna JM (1995) The origin of Arctic air pollutants: lessons learned and future research. Sci Total Environ 160/161:39–53

Perkins WT (2011) Extreme selenium and tellurium contamination of soils – an eighty year-old industrial legacy surrounding a Nickel refinery in the Swansea Valley. Sci Total Environ 412/413:162–169

Reimann C, Äyräs M, Chekushin VA, Bogatyrev IV, Boyd R, de Caritat P, Dutter R, Finne TE, Halleraker JH, Jæger Ø, Kashulina G, Lehto O, Niskavaara H, Pavlov VA, Räisänen ML, Strand T, Volden T (1998) Environmental geochemical atlas of the central Barents region. Geological Survey of Norway, Trondheim

Rice KC, Herman JS (2012) Acidification of earth: an assessment across mechanisms and scales. Appl Geochem 27:1–14

Rieuwerts JS, Farago ME (1996) Heavy metal pollution in the vicinity of a secondary lead smelter in the Czech republic. Appl Geochem 11:17–23

Rühling Å (ed) (1994) Atmospheric heavy metal deposition in Europe—estimations based on moss analysis. Nordic Council of Ministers, Copenhagen

Stohl A, Seibert P, Wotawa G, Arnold D, Burkhart JF, Eckhardt S, Tapia C, Vargas A, Yasunari TJ (2011) Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-Ichi nuclear power plant: determination of the source term, atmospheric dispersion, and deposition. Atmos Chem Phys Discuss 11:28319–28394

Zheng B, Hong Y (1988) Geochemical environment related to human endemic fluorosis in China. In: Thornton I (ed) Geochemistry and health. Science Reviews Limited, Northwood, pp 93–96

Zielinski RA, Simmons KR, Orem WH (2000) Use of ^{234}U and ^{238}U isotopes to identify fertilizer-derived uranium in the Florida everglades. Appl Geochem 15:369–383

Uptake of Elements from a Chemical Point of View 5

Robert J.P. Williams

Contents

5.1 Introduction: The Essential Chemistry of all Living Cells

In order to appreciate the initial link between the geosphere and the biosphere about 4×10^9 years ago, certain characteristic, essential features of the *organic chemistry* of all known living organisms need to be noted and put into the context of the nature of primitive Earth. The chemicals concerned are in the *cytoplasm*, which is the central compartment of all cells (Fig. 5.1). The common features of cells of this cytoplasmic chemistry are the syntheses of polymers such as the production of lipids (fats), polysaccharides (sugars), proteins, and nucleotides (DNA and RNA). They are all products of a few non-metal elements extracted from the environment; namely carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorus. Now the very nature of the polymers and their small molecule building blocks is that all of these six elements except phosphorus have to be in reduced states relative to $CO(CO_2)$, H_2O, N_2, (O_2), and elemental sulfur. This means that they are bound not just to one another but are bound, except for phosphorus, mainly by hydrogen. There is then a requirement in all cells for the cytoplasm to have a reductive synthesis capacity (below -0.5 V). Because the required reductions cannot be brought about by organic chemicals, a transition metal ion redox catalyst, such as iron and copper, is also essential. This cytoplasmic chemistry is demanding in other ways because the initial small building blocks, such as simple aliphatic acids, sugars, amino acids, and bases, need to be polymerized. There is then a second requirement for a suitable Lewis acid/base catalyst and again metal ions, especially divalent ions magnesium and zinc, are extremely valuable in reactions such as condensation polymerization.

$$XOH + YH = XY + H_2O$$

A pH close to neutral is also necessary to avoid hydrolysis and the solution must therefore be buffered by small molecules. Of course all the activities of a cell require

R.J.P. Williams (✉)
Oxford University, Oxford, UK
e-mail: jellywilliams@btinternet.com

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_5,

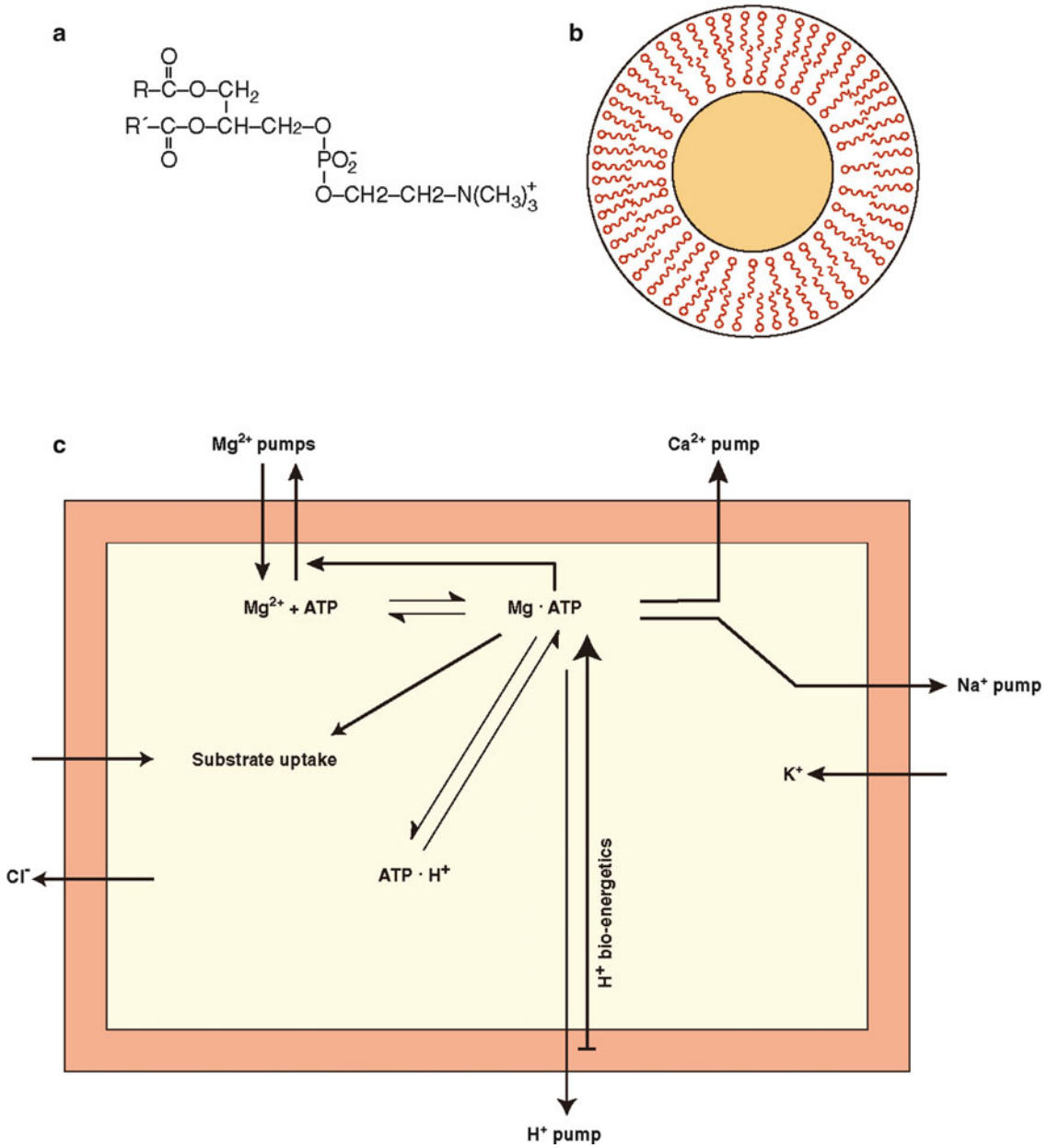


Fig. 5.1 The simplest possible cell had a single membrane that enclosed an aqueous solution called the cytoplasm. The membrane was made from lipids; see (**a**) and (**b**). The cytoplasm had to have a supply of energy shown as adenosine triphosphate (ATP) in (**c**) which is bound pyrophosphate (see text). As far as simple ions were concerned, the cell in the sea had to maintain relatively low Na^+, Ca^{2+}, and Cl^- while accepting K^+ and Mg^{2+} using energy (ATP) to pump them. The pH had to be kept at about 7.0. Note that ATP is also used to provide energy for organic synthesis, and it is made from a proton gradient due to redox reactions

energy, so the cytoplasm must have a way of capturing energy. Energy is used in synthesis, uptake and rejection of elements and compounds as required. All environmental energy sources such as light and unstable chemicals also require several inorganic elements in the capture machinery of cells. Many steps in this capture require one-electron redox reactions, which are almost invariably aided by transition metal catalysts. Condensation of phosphate, for example, in adenosine triphosphate (ATP) is also necessary. Energy-containing ATP is made by reaction of adenosine diphosphate (ADP) with phosphate giving a pyrophosphate, which on hydrolysis is used to remove water

$$XOH + YH + ATP = X - Y + ADP + P.$$

Finally, all of this organic chemistry occurs in water and to be reproducible it has to be carried out in a fixed ionic medium.

Table 5.1 Available concentrations in the sea as they changed with time (estimates)

Metal ion	Original conditions (molar)	Aerobic conditions (molar)
Na^+	$>10^{-1}$	$>10^{-1}$
K^+	$\sim10^{-2}$	$\sim10^{-2}$
Mg^{2+}	$\sim10^{-2}$	$>10^{-2}$
Ca^{2+}	$\sim10^{-3}$	$\sim10^{-3}$
V	$\sim10^{-7.5}$	$\sim10^{-7.5}$ (VO_4^{3-})
Mn^{2+}	$\sim10^{-7}$	$\sim10^{-9}$
Fe	$\sim10^{-7}$ (Fe^{II})	$\sim10^{-17}$ (Fe^{III})
Co^{2+}	$<10^{-13}$	$\sim(10^{-11})$
Ni^{2+}	$<10^{-12}$	$<10^{-9}$
Cu	$<10^{-20}$ (very low), Cu^I	$<10^{-10}$, Cu^{II}
Zn^{2+}	$<10^{-12}$	$<10^{-8}$
Mo	$<10^{-10}$ (MoS_4^{2-}, $Mo(OH)_6$)	$\sim10^{-7}$ (MoO_4^{2-})
W	$\sim10^{-9}$ (WS_4^{2-})	$\sim10^{-9}$ (WO_4^{2-})
H^+	pH ~7	pH 8 (sea)
H_2S	$\sim10^{-2}$	10^{-2} (SO_4^{2-})
HPO_4^{2-}	$<10^{-5}$	$<10^{-5}$ (HPO_4^{2-})

Note: The value for the original primitive conditions are estimates based on a pH of 8.0, an H_2S concentration of $\sim10^{-2}$ M, and a CO_2 pressure of one atmosphere. The concentrations in today's aerobic condition are taken from Cox (2)

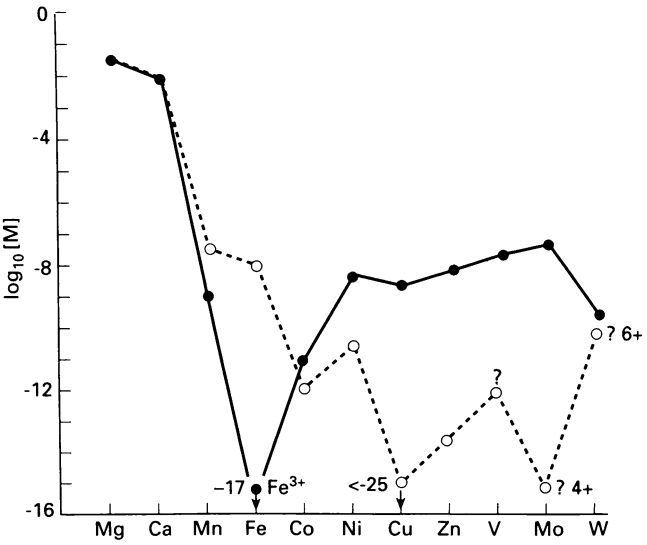


Fig.5.2 Contents of the sea in profile. The *dashed line* gives the molar concentration in the most primitive sea and the *full line* gives the concentration in the sea today

Along with the absolute requirements of the organic chemistry of the cytoplasm of all cells, there is then an absolute need for metal elements. Next the limitations imposed upon the possible selection of these elements will be reviewed as well as the nature of the original inorganic environment of the Earth from which these elements had to be obtained and which allowed life's chemistry to start. Once the system started, certain unique features were unalterable as is seen in the cytoplasm of cells in all life. The most important environments for our consider-action are the primitive sea (Table 5.1 and Fig. 5.2) and the atmosphere from 4×10^9 years ago, because it was in the sea (or at least in water) and the interface with the atmosphere that life began. There will be no attempt to describe the origin of life in this chapter, instead it will try to show the chemical limitations of the primitive sea and atmosphere for any development of a chemical system such as life before going forward to evolutionary considerations as this environment changed.

5.2 An Elementary Introduction to Primitive Earth

The formation of the Earth with its excess of metals and hydrogen gas over non-metals such as oxygen, due to the abundance of elements in the universe and the way Earth was formed, meant that from its beginning the whole planet was a reducing system. Although much of this metal content was isolated in the Earth's core and much hydrogen gas was lost, the different zones as they formed were made from reduced forms of many elements. Thus the atmosphere contained carbon as CO/CO_2 and some CH_4, nitrogen as N_2 and NH_3, oxygen and hydrogen as H_2O, sulfur as H_2S, and chlorine as HCl. The mantle contained lower valent metal oxides such as complex silicates and sulfides. (Note that all the metals could not be made into their preferred thermodynamic states as oxidic minerals so some were formed as sulfides.) It was seen that HCl reacted immediately with the metal oxides to give simple metal chlorides such as NaCl, which is virtually the only form of chlorine on Earth. Its solubility made NaCl the dominant chemical in the sea. The excess of basic metal over acidic non-metal oxides, which is partly due to abundances in the universe and partly to the volatility of non-metal compounds, gave rise to reactions that neutralized the acidity of the non-metal compounds to an ocean pH of about 7–8. At this pH carbonates, phosphates, and hydroxides form a buffered solution. Further solution of many elements was prevented by the insolubility of mixed oxides such as silicates and carbonates and sulfides. Full equilibrium between all elements in the core, the mantle, and the sea was not reached; otherwise there would have been much less of an *inner* metal core and no aqueous layer. However, it is suspected that approximate equilibrium was reached between the sea and its immediate neighbor phases with occasional injections from remote zones.

At that time there would have been and there still is a large temperature gradient from core to atmosphere. It is the high temperature of the isolated core plus energy from the sun that allowed water to remain as a condensed surface *liquid* sea, separate from the core and mantle. Its temperature

range from 0 to 100 °C, a temperature demanded for life to exist, has been maintained fortuitously for about 4×10^9 years. It was these energy gradients and the lack of chemical equilibrium between the sea and the core that also allowed a small flow of unstable reduced materials from the mantle to the sea. This also allowed a flow in cells that needed energy to reduce the oxides of carbon. Earth thus provided all the chemical ingredients as well as the energy for the cytoplasm of the most primitive cells. (Clearly the situation is not stable and a major eruption from the core could return the present atmosphere to its primitive state.)

5.3 Chemical Element Restrictions on Primitive Life

These considerations show that there are clearly strong restrictions upon the elements which could form a reactive system and become living cells. These reactions are due to the limitations of abundance and availability of the elements in the primitive sea. Turning to specific groups of elements, abundance severely restricted the amounts of lithium, beryllium, boron, and fluorine and all elements beyond atomic number 36. Abundance also restricted the amount of chlorine so that it was effectively all in the form of NaCl with some KCl and $MgCl_2$ in the sea or trapped in salt deposits. (There was no sulfate, only sulfide.) The sea was and is not saturated in such soluble salts, because the chloride supply is limited. Other availabilities in the early sea were restricted by the water solubility of oxides—sometimes as mixed oxides such as silicates, phosphates, and carbonates—and sulfides. The obvious restrictions due to oxide solubilities, apart from the mixed oxides of Si and P, were of the elements of Group 3 (Al group), 4 (Ti group), 13 (Ga group), and part of Group 14 (Sn). Restrictions due to sulfide solubilities generally applied to transition metals (Groups 3–12) and later elements in Groups 14 (Pb) and 15 (Sb, Bi), though some remained as insoluble metals (Au). The concentration of hydrogen as H^+ was controlled by the variety of buffering systems described above. Of great interest is the restriction on the availability of the reduced states M^{2+} and M^+ of the first transition metal series, but molybdenum and tungsten will also be looked at among metals and also selenium of the non-metals.

The precipitation of their sulfides limited concentrations [M], of the reduced (divalent) ions of the first transition metal series as follows:

Mn^{2+}	Fe^{2+}	Co^{2+}	Ni^{2+}	$Cu^+(Cu^{2+})$	Zn^{2+}
$<10^{-6}$	10^{-7}	$\sim 10^{-12}$	$\sim 10^{-12}$	$<10^{-20}$	$<10^{-15}$

The order is close to the order of solubility of many other of the salts of these elements and close to the inverse stability and extractability of their complexes. This order is called the Irving-Williams series. Outside this series the most interesting elements are from Group 5, 6, and 16. Molybdenum was reduced by H_2S to Mo(IV) and precipitated as MoS_2, whereas tungsten remained somewhat soluble as WO_4^{2-} or WS_4^{2-}. Note the geological states of these elements. Vanadium from Group 5 would also have been reduced, but its VO^{2+} ion remained somewhat soluble. Very few metallic elements other than Mn^{2+} and Fe^{2+} could have been present at greater than 10^{-8} M due to abundance and solubility restrictions. (The dominance of iron is in part due to its universal abundance.)

Among non-metals—other than carbon, nitrogen, hydrogen, oxygen, chlorine, and sulfur—which were all available as lower oxides or hydrides in the reducing atmosphere and/or in the sea, phosphorus and silicon remained as somewhat soluble elements HPO_4^{2-} and $Si(OH)_4$ in the presence of divalent metal ions in the sea. All halogens remained as halides usually seen in the sea. Selenium as H_2Se was the only other non-metal element of significant abundance and availability, because all others are rare or precipitated as sulfides.

Finally it must observed that these concentrations of elements in the sea, (Table 5.1) are close to equilibrium values of solubility and redox potential and complex ion formation among themselves because on the time scales of interest the elements in different compounds are in fast exchange.

5.4 Elements Presumed (Known) to Be Required by Primitive Cells

In the light of these geochemical data about the composition of the primitive sea and atmosphere, it must be asked which elements were not only available but have been found to be of significance for the required primitive chemistry of cells described above. The most primitive cells had a single compartment, the cytoplasm, encircled by a membrane. These cells are called prokaryotes. Rather than make further remarks about the environment, attention is now turned to the composition of the anaerobic prokaryote cells in sulfide media that are known today and assumed to be similar to the earliest forms of life. The elements in such cells are listed together with their free ion concentrations in Table 5.2. Note that quantitative knowledge of *concentrations* of particular compounds, e.g., *free* ions, is necessary as we wish to appreciate the nature of a *reaction system* that cannot be understood by looking at qualitative considerations such as the

Table 5.2 Composition of cytoplasmic fluid of all cells

Element	Form	Concentration (M)
H	H^+	10^{-7}
Na	Na^+	10^{-2}
K	K^+	10^{-1}
Cl	Cl^-	10^{-2}
Mg	Mg^{2+}	10^{-3}
Ca	Ca^{2+}	$<10^{-6}$
Mn	Mn^{2+}	$\sim10^{-6}$
Fe	Fe^{2+}	$\sim10^{-7}$
Co	Co^{2+}	$<10^{-9}$
Ni	Ni^{2+}	$<10^{-10}$
Cu	Cu^+	$<10^{-15}$
Zn	Zn^{2+}	$<10^{-11}$
Mo	MoO_4^{2-}	$\sim10^{-7}$

Table 5.3 Exchange rates of water from metal ions

Ions	Exchange rate (s^{-1})
Na^+, K^+ many anions	10^9
H^+	10^{11}
Ca^{2+}, Cu^+, Mn^{2+}, Zn^{2+}	10^8
Fe^{2+}, Co^{2+}	$>10^7$
Mg^{2+}	10^6
Ni^{2+}	10^4
Fe^{3+}, Al^{3+}	$<10^3$

total amounts of elements or of such molecules as DNA. The profile of metal elements in each compartment of a cell (or the sea; Fig. 5.2) is called a metallome, which has two concentration parts: the free metallome and the bound metallome. It is also important to distinguish the elements that are in fast exchange within the lifetime of a cell (see Table 5.3). It is immediately noted that there is a general agreement between the concentrations of the available elements free in the primitive sea and those free in the cytoplasm of cells (Tables 5.1 and 5.2). This is still true despite the changes in the sea. Just as the essential organic chemicals of the cytoplasm have remained fixed, so the free cytoplasmic inorganic element concentrations have not changed. In fact the two may well have initially formed an *inevitable*, *locally energized*, *system of reactivity* in some restricted compartment based on the composition of the primitive sea and atmosphere before there were coded molecules such as DNA. The suggestion is that long before there was Darwinian pressure on reproductive evolution there was an energized chemical drive to a particular system. This has not been proven, but it is very difficult to see how life began from any other standpoint.

At the quantitative level there are still some striking divergences between the cellular and the environmental sea free ion concentrations. Most notable are the differences in free sodium, chlorine, and especially calcium. Why is the system energized in these gradients much as organic chemistry is energized by covalent binding?

Before further consideration of the free and combined forms of the elements in the environment and in primitive cells and then the ways in which evolution has generated new cellular systems, and given that the cell concentrations under discussion are not all under thermodynamic control but can have strong kinetic management, a brief outline of the methods of study of the element concentrations and controls will be given in the next section. It is believed that a thorough knowledge of the element content of every compartment of living organisms is equally essential to the knowledge of its protein content, the proteome. In consideration of evolution, it is necessary to follow the analysis of elements in free and combined forms in cellular compartments with that of those available in the environment with time. This may give new insight into what is taken here to be the inevitability of the development of life, although its timing is quite uncertain. It may also help us to set up criteria for possible life forms on other planets.

5.5 Methods of Study

The knowledge that is required to understand the ways in which different elements function in organisms can be divided into:

1. Determination of their total analytical concentration
2. Appreciating the precise location of each element
3. Deciphering the differently combined forms and concentrations including the free ions in different places
4. Determining the kinetics of exchange, including uptake and rejection, and the associated activity of free and combined forms, including restrictions on diffusion in chemical and compartmental traps
5. The link to organic syntheses

This chapter cannot handle these extensive topics in any depth and the interested reader is directed to a variety of methods described in the literature. Under (1) see above destructive micro-analysis is sufficient. To appreciate location (2) the best methods employ sectional slice analysis using dark-field EM methods and low temperatures where possible to determine atomic masses. Note that modern tomography and image reconstruction procedures can give high resolution below 10 Å in all three-space dimensions for certain materials. The free ion concentrations locally are determined using optical or fluorescent binding reagents, and total local element concentrations are determined by staining or SIMS in which a focused laser beam volatilizes a small local volume into a mass spectrometer. Proton bean

microscopy is useful for many elements too. Often the only methods available under (3) depend upon careful extraction because the molecules are unstable. Fractionation into units such as vesicles or of molecules is then required. Traditional analytical methods as in (1) can then be used to quantify element composition. The determination of uptake or rejection (4) can use radio-isotopes or mass spectrometry (isotope mass fractionation) or of course the traditional methods of enzymology. Using all the methods together generates knowledge of the concentrations of selected elements, free and combined, in many separate compartments. However, the collection of analytical knowledge is far from complete.

It is difficult to understand the link under (5) between the concentration of the combined elements and organic synthesis, but we especially need to know about the organic synthesis of the reagents with which metals are combined. There are two very different final situations: thermodynamic equilibration of metal ion, M, and ligand, L; and irreversible insertion of a metal ion into a ligand. Note that the ligand can be a small chelating agent, a protein, or even DNA. In equilibrated binding the uptake of free metal ion and the synthesis of the ligand must be limited by feedback controls over uptake and synthesis, i.e., to expression via DNA/RNA, so that a complex ML is present in a fixed concentration. In the second situation formation of [ML] is not related to free [M] but to control the kinetics of insertion and the accessibility of L feedback controls rest upon free [ML] as well as upon [M] and [L]. The methods of study now depend not on analysis alone but on knowledge of organic synthetic pathways and their controls all the way back to the gene. Genetic control of L synthesis resides in feedback to transcriptional and translational regulation by [M], [L] and [ML] (see Fig. 5.6). Study is now needed of DNA and its regulated response to inorganic elements. Note again that study of evolution through DNA can lead to evolutionary connections, but it says nothing about concentrations and causation of systems. It is not possible to appreciate the developments of chemical systems through the analysis of single molecules. Systems are to be seen through quantitative intensity factors such as temperature and concentrations of components in local zones. Finally, all the processes of uptake, rejection, and synthesis require energy (see Table 5.4), which have changed with time.

Table 5.4 Sources of energy

Period (years ago)	Energy sources
Initial (4.5×10^9)	(a) the sun
	(b) basic unstable chemicals in the crust
	(c) chemicals stored at high temperature in the core
After say 1 billion years (3.5×10^9)	(a) as above
	(b) some oxidised materials, some SO_4^{2-}; very little O_2, H_2O_2
After say 2 billion years (2.5×10^9)	(a) as above
	(b) further oxidised materials; modest O_2, H_2O_2
After say 3 billion years (1.5×10^9)	(a) as above
	(b) further oxidised materials, almost 1% of final O_2 pressure
Today	(a) as above
	(b) man's fuels
	(c) atomic energy

5.6 Equilibrium Binding and Exchange

As indicated above, equilibrium is established rapidly between most organic ligands or inorganic anions with lower valent cations even when binding is quite strong. This feature of inorganic chemistry contrasts strongly with the kinetic control over the reactions of organic elements in compounds. Although there are exceptions, the rule is that thermodynamic constants (see Fig. 5.3) are limiting factors over the binding of many metal elements in cell compartments. These include at least Na^+, K^+, Mg^{2+}, Ca^{2+}, Mn^{2+}, and much of Fe^{2+}, but only some of the bindings in biological cells of other elements found in cells. Molybdenum, cobalt, nickel, copper, and zinc can be considered in this way too. The pattern of anion reactions is often different because they can be linked covalently in organic molecules, but there are frequent occasions when it is necessary to refer to the binding equilibrium constants of Cl^-, HCO_3^-, NO_3^-, SO_4^{2-}, HPO_4^{2-}, MoO_4^{2-}, WO_4^{2-}, and $Si(OH)_4$ in cells. Here our concern is with the metals, referring only now and then to non-metals.

It has already been seen that the equilibrated binding in oxides and sulfides restricted the primitive environmental availability of the metal elements. The plot of insolubility of the sulfides in the Irving-Williams series is very steep, rising to beyond 10^{-20} M at copper at pH = 8. We have to consider these insolubility data against the ability of the organic molecules, which cells could produce from H, C, N, O, P, and S, in considerable variety, to bind reversibly to the metal ions at pH = 7 (Fig. 5.4). The resulting organic molecules have frameworks with pendant side-chains that can bind to metal ions. Single side-chains such as —OH, —NH_2, —SH, —COOH, and —OPO_3H_2 and their anions bind weakly. The vast majority of bindings therefore involve from two to six of such groups in the form of looped structures called chelates. These binding ligands are divided according to the binding (donor) atom into oxygen, nitrogen, and sulfur donors. Model studies have given the optimal conditional binding to various such ligands, and are therefore to be expected of proteins at pH = 7 which form the majority of the frameworks. Proteins have sets of such O-, N-, or S-donor binding centers. The binding constant profiles for the

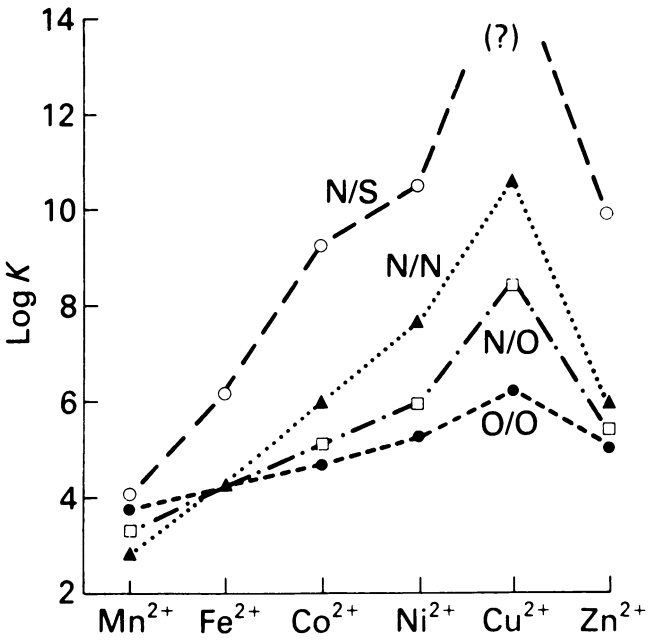


Fig. 5.3 A plot of the stability constants of some metal ion complexes. The constant is for the equilibrium $M + L \rightleftarrows ML$ Mg^{2+} (not shown) binds less to all ligands except those labeled O/O such as ATP (see Fig. 5.1). The ligands here donate electrons to the metal ions. Typically O/O is a pyrophosphate, N/O is an amino acid, N/N is a di-amine, and N/S contains an amine and a thiolate (see Fig. 5.4). Note that the inverse of these plots are the profiles of the free metal concentrations in the presence of certain ligands

different combinations in small model complexes (and proteins) are shown in Fig. 5.3. It is clear that not only did and does availability restrict the possibility of a cell obtaining an element from the environment, but binding constants indicate the limits at which metal ions could have been captured by ligands in primitive cells against the competition from the environmental oxides and sulfides even though uptake could be energized. Certain general statements can be made based upon these effective equilibria and knowledge of cellular organic chemicals, largely proteins. (Effective refers to the binding under the conditions of temperature and competitive reactions in the cell.) These statements include:

1. Na^+ and K^+ could only be retained at the concentration of the free ions in cells, 10^{-1} to 10^{-2} M, by physical barriers, because binding constants are very small. The binding by special cyclic or rigid frame O-donor ligands can retain only a small percentage of the ions.
2. Mg^{2+} and Ca^{2+} could be retained considerably in bound forms at the observed concentration, 10^{-3} and $<10^{-6}$ M respectively, of these free ions, by a variety of O-donor ligands but not by S- or N-donors.
3. In marked contrast with (1) and (2), it was hard to retain almost any transition metal ion by O-donor ligands alone, but the increased binding by N- and S-donors did allow binding at the ion concentrations seen in cells for Mn^{2+} and Fe^{2+} ($\sim 10^{-6}$ M) even in the presence of H_2S, which gave rise to external precipitation of sulfides.
4. There were very few ways of retaining Co^{2+}, Ni^{2+}, Cu^+ (Cu^{2+}), and Zn^{2+} because of the removal of these ions by the sulfide of the primitive sea.
5. Molybdenum was more strongly held in MoS_2 than were tungsten and vanadium in their sulfides so that tungstate and VO^{2+} would have been captured in primitive cells by organic chemical ligands to a greater degree than molybdenum.
6. Among anions that were not covalently retained, chloride could only be retained at the concentration found in cells (10^{-2} M) by physical membrane constraints because it is very weakly bound. Sulfate did not exist in primitive times but some phosphate could be retained at equilibrium binding at 10^{-4} M and also kinetically by covalent binding. Bicarbonate (carboxylate of small ligands) is weakly held but organic phosphates are more strongly retained.

   It is very important to understand that all equilibria in all systems in an isolated compartment are restrictions on the variables of the system that are not biological but purely chemical, contrast a coded synthesis of a protein, and each unavoidable effective equilibrium constant is then an enforced characteristic of life or of any other organization. The profile of free exchanging metal ions at effective equilibrium in the cell cytoplasm is a characteristic of all life and is fundamental to it (Fig. 5.5). Any code had to work with this profile just as it had to produce only certain organic molecules.
7. Because all ligands bind all the metal ions when both are in sufficient concentration, it is important to devise the ligands such that their binding constants are as discriminating as possible opposite the free ion concentrations allowed in cells. This permits each metal to be used in a selected functional way (see below). It is necessary therefore to limit the ligand concentration so that there is little excess free ligand after binding its selected element. The latter avoids binding of the wrong metal to a site. This was achieved in cells by limiting the production of the binding ligand through feedback inhibition by [M], [L], or [ML] via DNA of the synthesis of the ligands. This possibility could hardly have been present in the most primitive system but is a dominant feature in coded cells (see Fig. 5.6).

Before continuing this analysis, it is necessary to be careful not to overgeneralize. Equilibrium considerations are important, but in two ways cellular chemistry escapes their limitations:

1. Elements can be energized as they pass through physical, membrane barriers so that free concentrations can be increased or reduced by up to 10^4 relative to environmental

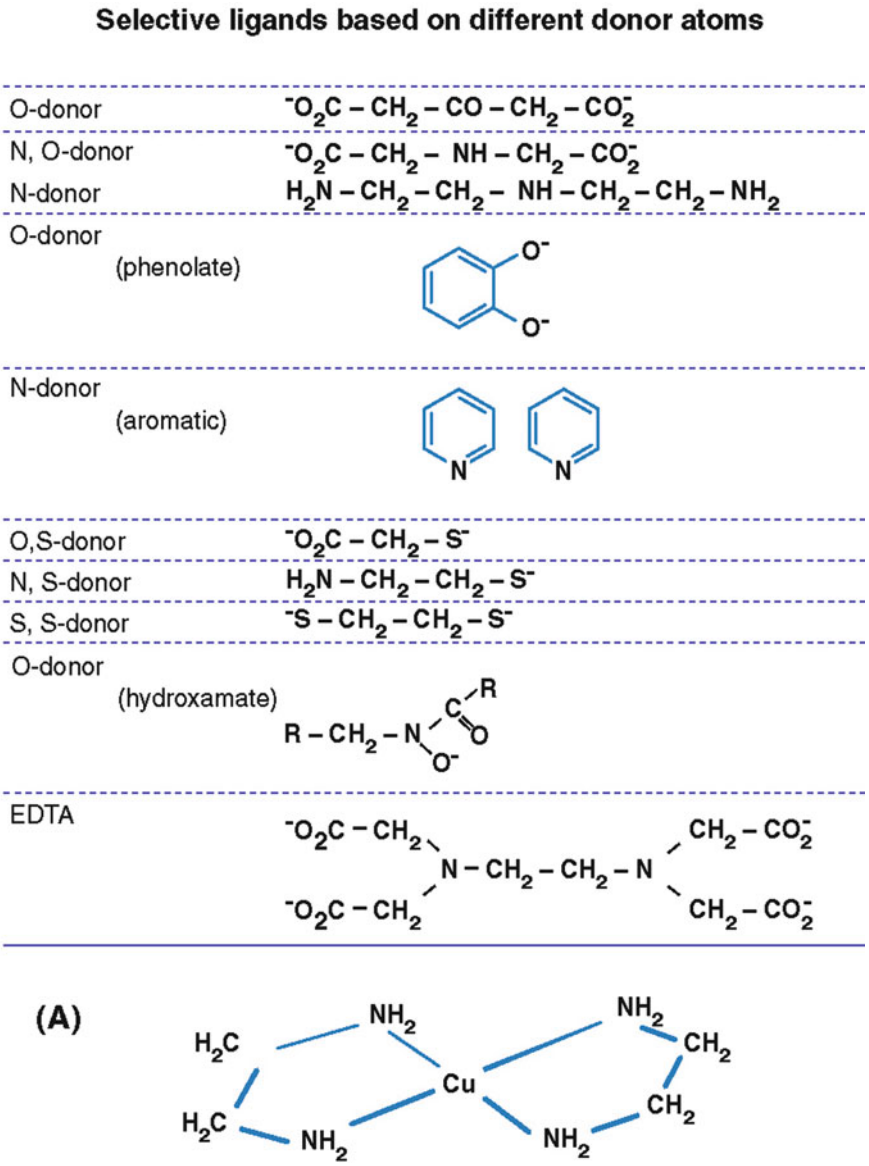


Fig. 5.4 The formulae of some organic ligands and a complex ion with organic ligands (A)

values while to some degree all ions will leak into cells. This prevents equilibrium with the environment but not internally to a compartment.

2. Binding can be made as irreversible inside cells as in the uptake of non-metals.

Only those elements that exchange reasonably readily inside cells can be examined. A factor of 10^4 to 10^{-4} has been added in Fig. 5.5 as the variation away from equilibrium based upon known values of free ions in the sea and the known limitations of uptake or rejection into or out of cells. This gives a picture of free ion levels based on model chemical and real biological complex ion studies (shown in Fig. 5.5), which applies to the cytoplasm of all cells. These profiles have been referred to as the free metallome because there are parallels in all cells. Moreover, the elements that were bound by selected organic ligands, given the variety of O-, N-, and S-donors which the cell could make, are close to expectation (Figs. 5.3 and 5.4). This leads again to the suggestion that life started as a system of energized chemical reactions in one compartment which required the presence of an environmental zone on Earth containing about 20–30 available elements. Of these, about 20 are utilized in cells and a few are rejected. The zone that became a cell was then self-sustaining but unable to reproduce. These are general features of life which are not due to competitive Darwinian pressure but arise from geological and chemical limitations. In this light it is possible to see how cells use elements in an energized way.

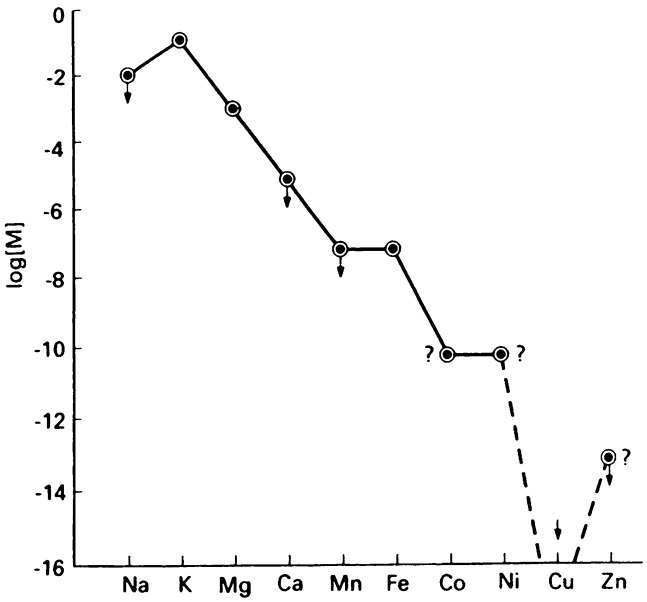


Fig. 5.5 The profile of free metal ion concentrations in primitive cells. The values are not very different in the present-day cytoplasm of all cells

5.7 The Functional Value of Elements in Cells

5.7.1 Osmotic and Electrical Balance

The most basic requirement of the inside of a cell is to avoid excess concentration and charge, which could destabilize a cell by osmotic or electric pressures (see Fig. 5.1). By necessity a cell is rich in organic molecules and, hence, it must have had a lower content of inorganic ions than the sea. The organic molecules of the cell are mostly anions so the cell must have also rejected the major anions of the sea, but even then it had to maintain approximate charge neutrality. These objectives are achieved by pumping out Na^+ and Cl^- and allowing K^+ to come in. Another essential feature is that Ca^{2+} must be strongly rejected because it forms precipitates and aggregations with organic anions. Most other free and smaller divalent ions, with the possible exception of the larger Mn^{2+}, can be retained up to the limits seen in the primitive sea. Mn^{2+} forms some insoluble salts such as oxalates. The resultant energized gradients of Na^+, Cl^-, and Ca^{2+}, the main ions rejected, allowed the favorable uptake of K^+ (to neutralize charge) and an internally equilibrated intake of all other cations (and anions) in small free amounts but with a high percentage bound. Although this description of the primitive cell system is certainly a necessity, it is still unknown how this energized system evolved. (This puzzle parallels that of the origin of synthesis of the particular organic compounds, which came to form a coordinated set in a cell and no solution to the problems is offered.) However, given fast exchange it is clear that the two surfaces of the energized pumps, in and out, must have binding constants for each selected ion close to the inverse of concentrations of the ion in the cytoplasm and in the sea, respectively.

5.7.2 Simple Condensation and Energy-Transfer Reactions

Given our lack of understanding of the basic initial organic chemistry in life, we will assume that all cells can select the organic chemicals, L, they produce and control their amounts. This step needs to be taken because the next metal elements, M, for consideration bind strongly to these organic chemicals. It is essential that [M] and [L] are fixed. [ML] is then fixed selectively by equilibrium. Hence, it should be taken into account that a coded cell can limit production of an organic chemical that binds metal ions by a feedback control to the production center, DNA/RNA (Fig. 5.6). In feedback control the production of [L] occurs when [M] is high, but [M] itself is limited by feedback to the pumps which stop entry at certain values. For elements in fast exchange, *e.g.*, Mg^{2+} and Ca^{2+}, this implies that the feedback control must have a binding constant close to the inverse of the free metal concentration. This is true for all the metal ion complexes that operate under free metal ion control including the pumps. The equilibrated metal ion, therefore, binds equally to all proteins that are active: enzymes, some small molecules, the inner sites of membrane pumps, any carrier molecules, and transcription factors which control the production of all these molecules either directly or ultimately. The case of Mg^{2+} is typical. Its binding constant is close to 10^3 M^{-1} to all its sites (Fig. 5.7), which implies that all binding sites are roughly half-saturated because free [Mg^{2+}] is close to 10^{-3} M. This is ideal for response to variations in [Mg^{2+}] because a system acts most sensitively as a dynamic buffer at half-saturation. Thus the activity of the cytoplasm of cells is closely dependent on the value of free [Mg^{2+}] in the sea (10^{-2} M). This includes all reactions of nucleotide tri- and diphosphates (ATP)—the major energy distributing entity of cells for metabolism including pumping and synthesis. The only available sites that can bind Mg^{2+} because of its chemical nature (see Sect. 5.6) are O-donor centers. None of these Mg^{2+} O-donor sites can bind any other metal ion at the concentrations available inside a cell, which again are limited by those in the sea. Many of the condensation reactions of the organic syntheses depend on O-donor groups brought together for synthesis with energy-giving ATP which also has only O-donor groups and also requires activation by metal ions. Limitations are imposed on the availability of other elements by solubility or energized rejection (Ca^{2+}). Hence, it is clear that Mg^{2+} was the only

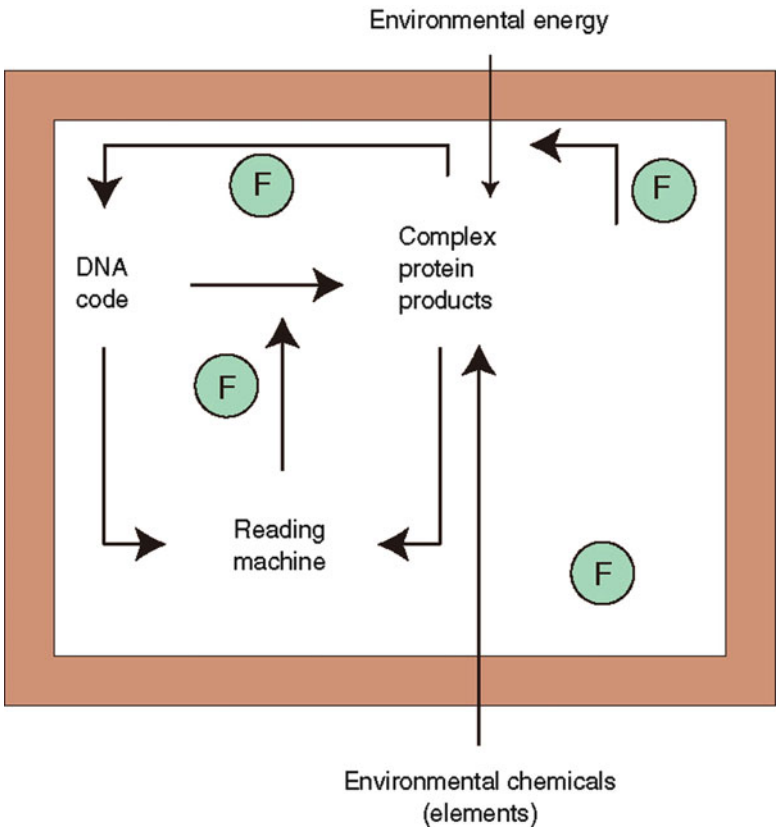


Fig. 5.6 A simple picture of feedback, *F*. In each case a molecule or ion in a cell builds up to a certain level by pumping (Fig. 5.1) or by metabolism. At this level the molecule or ion feeds back to the pump or the metabolic path and blocks it. This ensures that cells work at a fixed concentration of many small components, including both free metal ions, the metallome, and free mobile coenzymes, which are part of the metabollome

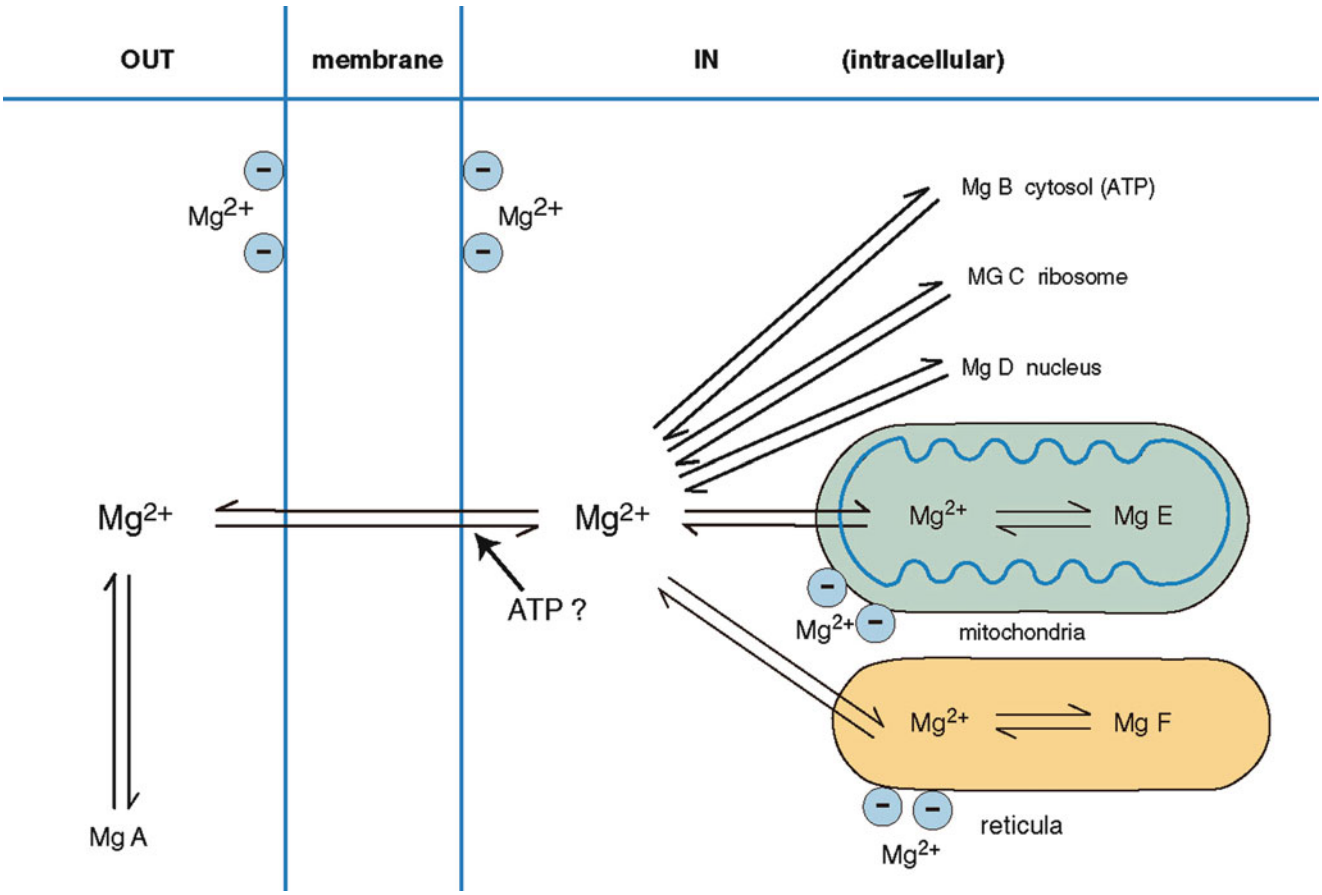


Fig. 5.7 The general way in which Mg^{2+} interacts in equilibrium with many sites in a cell. Note that the free Mg^{2+} concentration in the cytoplasm of all cells is close to 10^{-3} M, and this is controlled by feedback and energy input at pumps (Figs. 5.1 and 5.6)

possible initial acid catalyst for much cytoplasmic activity. The $[Mg^{2+}]$ concentration is actively maintained by pumps, but this may not have been initially necessary. Pumps have to operate to ensure 10^{-3} M cytoplasmic Mg^{2+} against any external concentration. It is seen that some cellular chemistry arises from chemical (thermodynamic) necessity, not from coded chance.

The corresponding necessary binding constants for K^+, Na^+, and Cl^- in the cell are 10^1 M^{-1} (K^+) and $>10^2$ M^{-1} (Na^+ and Cl^-) opposite their free ion concentrations in the cytoplasm of 10^{-1} M (K^+) and $<10^{-2}$ M (Na^+, Cl^-). These binding factors also arise from the nature of the primitive sea and not from any coded imposition. In fact the code molecules, DNA and RNA, are still absolutely dependent for activity upon binding constants of about 10^{-3} M^{-1} for Mg^{2+} and 10^{-1} M for K^+.

Next the free Ca^{2+} ion levels in the cytoplasm are considered. Here outward pumping was and is essential and reduces free $[Ca^{2+}]$ in the cytoplasm of all cells generally to $<10^{-6}$ M, although advanced cells have $[Ca^{2+}]$ at 10^{-8} M. This means that the range of useful Ca^{2+} equilibrated binding sites inside cells must be around 10^6–10^8 M^{-1}. Later in eukaryotic cells, which have a huge variety of Ca^{2+} binding proteins (Fig. 5.8), Ca^{2+} became the major messenger giving the cell, upon fast Ca^{2+} pulsed entry and exit, information about its external world. On entry Ca^{2+} rises to 10^{-7} or 10^{-6} M and binds to all the response proteins of about ten different kinds. These have known binding constants of about 10^7 M^{-1} and this includes the exit pump which rejects the pulsed Ca^{2+} immediately. $[Ca^{2+}]$ at 10^{-7} M does not usually connect directly to DNA to give transcription; instead it activates Mg^{2+}-dependent sensors such as kinases, which are proteins dependent on ATP. In a passing note none of these Ca^{2+} selective proteins bind Mg^{2+} with a constant $>10^3$ M^{-1} so that Ca^{2+} and Mg^{2+} do not compete. These effective binding constants are necessities even though the syntheses of the molecules are coded. Note that Ca^{2+} pumps bind Ca^{2+} reversibly on the outside at $<10^3$ M^{-1} and on the inside at $>10^6$ M. It is also found that the reversible binding of Ca^{2+} to CO_3^{2-} (carbonate) and PO_4^{3-} (phosphate) gave rise to precipitates at pH = 7 at 10^{-3} M Ca^{2+}, close to the value in the sea.

5.7.3 Electron Transfer

When the binding constants and the availability of elements for the catalysis of the other essential reactions of primitive cells such as reduction and energy capture by electron flow

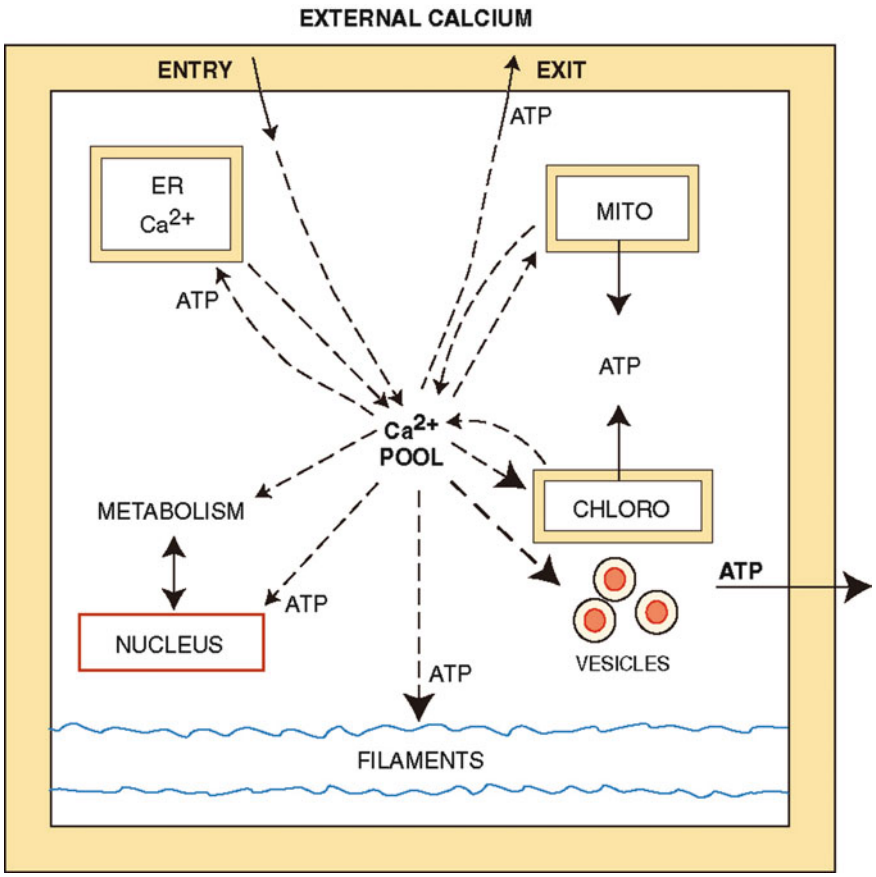


Fig. 5.8 The general way in which Ca^{2+} can act as a messenger in eukaryotes. The initial effect is to allow Ca^{2+} entry to increase due to an environmental change. This input raises $[Ca^{2+}]$ above 10^{-7} M so that it activates a multitude of cellular responses including energy available (ATP), synthesis, and import/export of chemicals. Subsequently and rapidly the $[Ca^{2+}]$ returns to 10^{-7} M

are considered, it is seen from solubility data (Table 5.1) that only Fe^{2+} could have been used initially. It could not have been retained selectively by O-donors but only by S- or N-donors of the organic molecules available (Fig. 5.3). The binding constants to all those of its proteins, which are in fast exchange, must be about 10^7 M^{-1} if all are to operate together in a controlled fashion. This limitation, built into feedback controls, is due to the nature of Fe^{2+} and organic ligands of the types found in proteins.

5.7.4 Oxygen-Atom Transfer

Iron was essential for one-electron reactions $Fe^{2+} \rightleftarrows Fe^{3+}$ especially in a sulfur matrix. The major primitive iron proteins are small $[FeS]_n$ units. Such units cannot do another required oxidation/reduction reaction, the transfer of oxygen (a two-electron reaction), such as M + RCOOH $\rightleftarrows$ RCHO + MO. This reaction is catalyzed readily at low redox potential by molybdenum and tungsten. In primitive conditions it could only have been that tungsten was used in two-electron, O-atom transfer because it was the only available element with a low two-electron oxidation-reduction potential for such transfer. Molybdenum was precipitated as a sulfide. It does look as if life's system in the cytoplasm arose as an inevitable consequence of the nature of the primitive environment, and it was repeatable in form but not reproducible. The basis of this assertion is that much of the inorganic chemistry was in fast exchange unlike the organic chemistry, which could only occur in the presence of these elements. The use of tungsten, a very rare element in essential reactions, was a necessity not an accident. Free tungstate, later molybdate, is approximately 10^{-7} M in the cytoplasm of cells (see Tables 5.1 and 5.2).

5.8 Evolution

5.8.1 New Forms of Old Elements

There are three possible modes for the evolution of such a chemical system. These are logistical requirements, independent from coding, and all three depend upon isolating the initial *cytoplasmic, partly equilibrated, organic/inorganic system* from alteration, while evolving in other ways. The central chemistry has to be so maintained almost unchanged in order to keep the basic reducing and acid/base metabolism, which generates all the essential organic polymers common to all known life. The first possible mode of evolution is to develop new organic chemical pathways in the cytoplasm, but now using irreversible chemical traps for the already used or any other newly available elements. This is a change of the bound not the free metallome. The developments of protein-independent ring chelates for iron, cobalt, nickel, and magnesium, all related in synthesis to uroporphyrin (Fig. 5.9), and of the organic part of the coenzymes for molybdenum (or tungsten or vanadium) are clear examples. In effect these chelates, which do not exchange their central metal ions, are new "elements," i.e., new independent variables in the chemical system of cells. They avoid thermodynamic competition with the environment or cell's internal equilibria, as does the cell's organic chemistry. However, their internal concentrations are controlled by feedback like all other cell components.

5.8.2 New Compartments

The second possible evolutionary mode is the creation of new physical compartments, *e.g.*, vesicles (Fig. 5.10a), which can have the same or different elements as the cytoplasm but at different concentrations because they are independently energized and controlled (Fig. 5.10b). For bacteria the development was the introduction of the periplasm, a simple compartment surrounding the cytoplasm, which is richer in the elements Na^+, Ca^{2+}, Mn^{2+}, and Cl^-, because the periplasm was open to diffusion of ions from the outside but did not allow proteins to escape. This isolation with new external binding sites and the gradients of ions produced novel catalytic energy-producing and communication possibilities which have evolved over 3×10^9 years. The rejected Mn^{2+} (plus Ca^{2+} and Cl^-) produced a periplasmic facing enzyme in the cytoplasmic bacterial membrane capable of generating external O_2, and in turn, this has enabled all future evolution of chemicals and compartments as seen below. It also generated increased reducing capacity internally. Both features were aided by the new chlorophyll (Mg) and electron transfer (heme iron) complexes mainly in membranes (Fig. 5.9). At the same time the binding of Ca^{2+} on the outside of the cell became the basis for the later receptor-based messages giving inward Ca^{2+} pulses (Fig. 5.8). There was, also later, the creation of internal vesicles, membrane-contained space separated from the cytoplasm. Cells containing such vesicles are called *eukaryotes*.

Kinetically controlled vesicular compartments created the second opportunity to multiply the variables of the system, in contrast with a phase at equilibrium which is a restriction on a system. The compartments could contain different organic molecules including proteins and enzymes. In this way the new compartments create new activities, the most obvious of which is the synthesis of biological minerals such as carbonate, silica, and phosphate shells. These are the basis of many sedimentary rocks. It was the development of such vesicle compartments that caused the explosion of

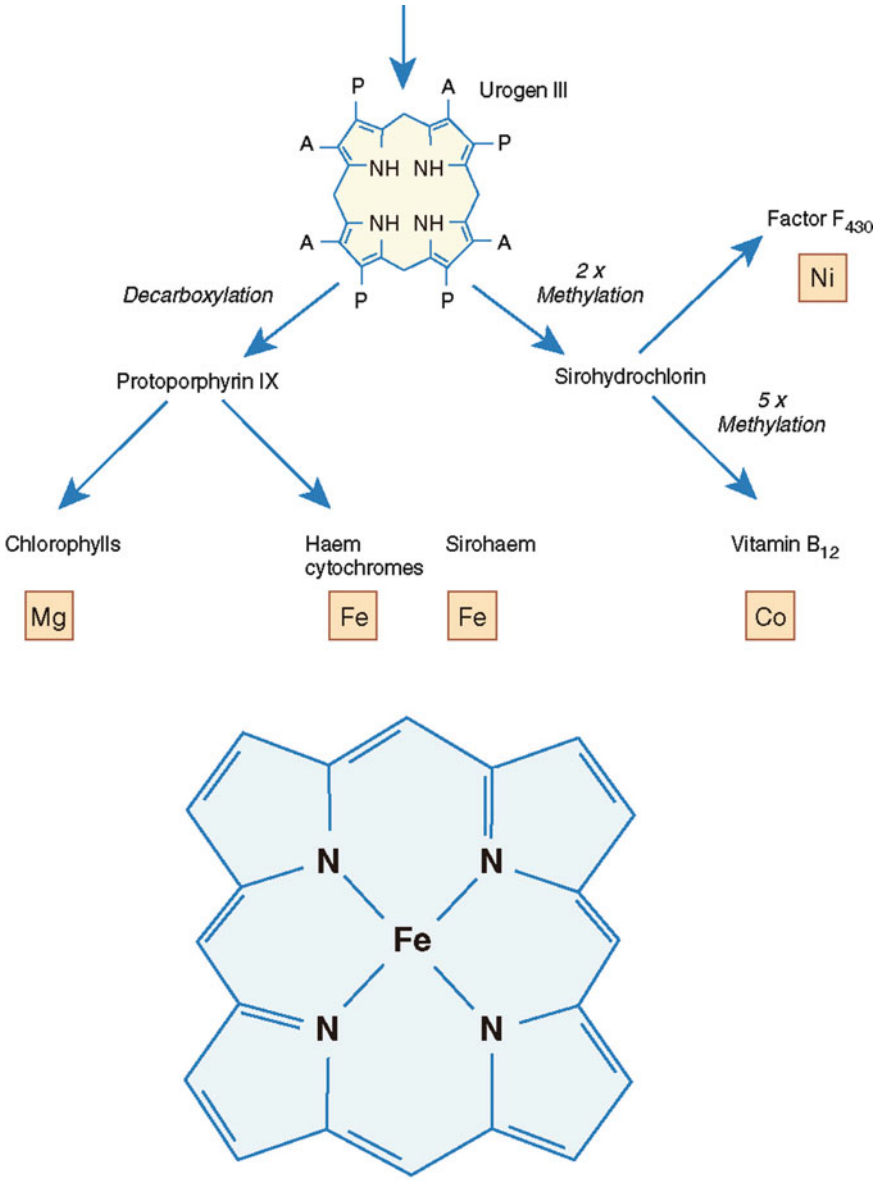


Fig. 5.9 The development of porphyrin-like chelating agents to generate new variables in the cytoplasm of primitive cells

separate species in eukaryotes. The origin of vesicular compartments, separate but kinetically linked to the cytoplasm, lies with the coming of dioxygen as a new chemical species. Note that the equilibrium for Ca^{2+} outside the cytoplasm, now also in vesicles where it binds to proteins, has a binding constant of 10^3 M^{-1} opposite the free (environmental) Ca^{2+} of approximately 10^{-3} M. The binding constant for external Mn^{2+} may well be around 10^7 M^{-1} opposite the level of the primitive sea of 10^{-7} M free Mn^{2+}. The production of dioxygen linked to energized Mn chemistry using light made it possible first to make the new membrane structures found in eukaryotes and then to do much more novel chemistry in ever extending organization. This is the third evolutionary step.

5.9 Dioxygen and Evolution: Single Cell Eukaryotes

5.9.1 New Compartments and New Communication Networks

Elsewhere we have described the slow progression of the increase in the oxygen pressure in the atmosphere over 4×10^9 years (Fig. 5.11), which is a new kinetic compartment inside and outside cells. Its final effects are not yet complete, but Fig. 5.12 shows that with time different oxidation states of elements have been introduced into the Earth's waters by its activity. This is seen as an inevitable

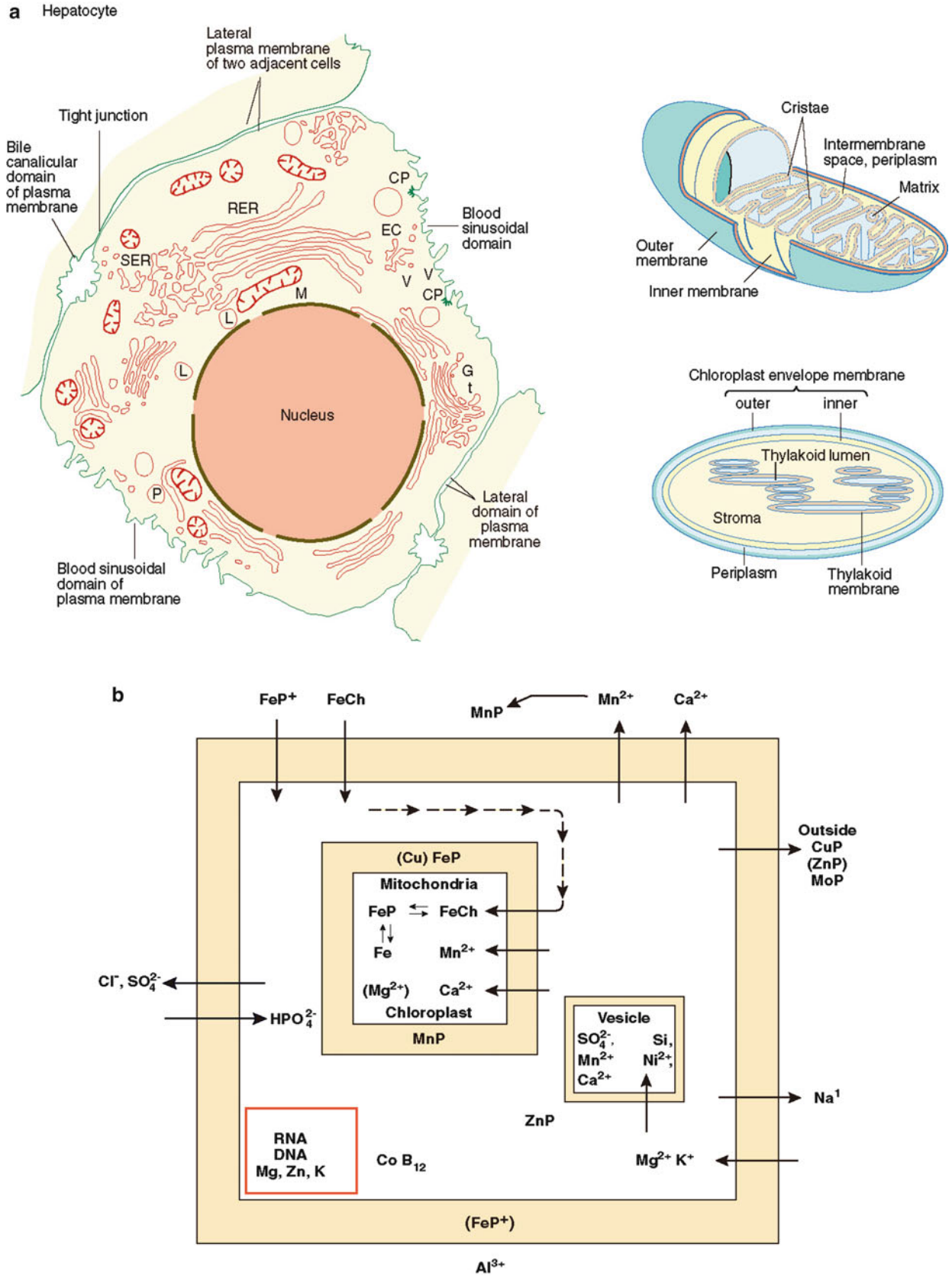


Fig. 5.10 (**a**) The nature of a modern cell with compartments. *M* mitochondria (see *side Figure*), *RER* rough endoplasmic reticulum, *L* lysozome, *V* vesicle, *SER* smooth endoplasmic reticulum, *G* Golgi apparatus, *CP* chloroplast (see *side Figure*). (**b**): The distribution of elements in the new compartments; compare with Fig. 5.1

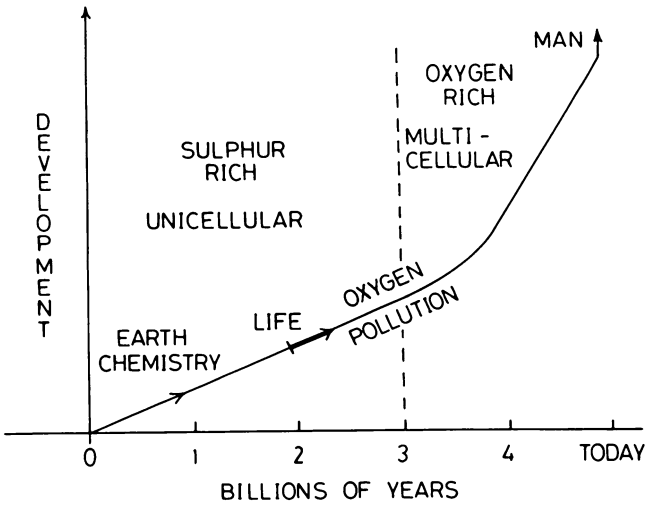


Fig. 5.11 The rise of dioxygen with time and some of the enforced cellular changes

timed thermodynamic succession following water splitting to O_2^+ bound H by cells and produces new kinetically controlled possibilities. The consequences for cellular evolution, which created the oxygen in the first place, are also inevitable: only certain new reaction paths could be created which give new chemicals in cells, and they had to be largely confined in the new compartments and controlled by novel proteins. The timing of change is unknown, but once oxygen was produced the general direction of use of higher redox potential chemicals is unavoidable. Note that it is mainly acid-base (Zn^{2+}) and redox catalysis (copper, molybdenum, selenium) which appear, while iron and manganese are reduced in availability and do not change functional value significantly in the cytoplasm.

Subsequent to the appearance of ring chelates and the coenzymes of molybdenum, tungsten, and vanadium, the first step in evolution due to oxygen was the creation of a flexible outer membrane through the synthesis of an oxidized

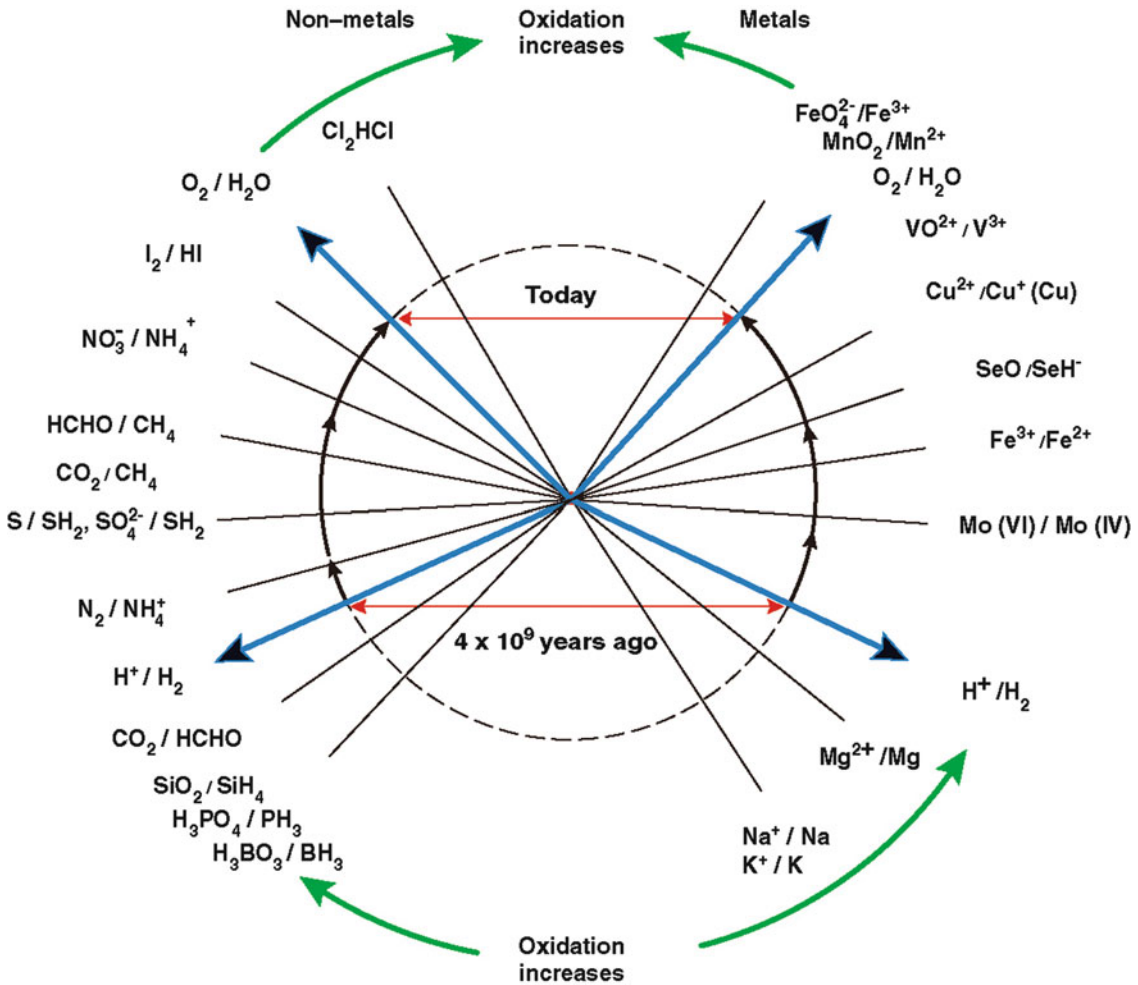


Fig. 5.12 The enforced changes in the oxidation states of elements in the environment due to the equilibration with the gradual rise in the oxygen pressure

squalene derivative, cholesterol. Simultaneously, to maintain the stability of a cell, filamentous internal structure was essential. Together these changes allowed the construction of much larger cells or eukaryotes. Once this structure appeared, new compartments, vesicles, and organelles could be permanently developed within the cell. The organelles, mitochondria, and chloroplasts are, in fact, captured prokaryote cells with their own DNA. Within the structures new reactions pathways could evolve and some pathways in the cytoplasm could be safely dropped because the organelles could also manage them. The evolution of the organic chemistry should not be stressed here, but it must be remembered that any organization that develops as a system of compartments, which is the general way to develop, needs a central command unit and a communication network. The communication network to and from the cytoplasm, to and from vesicles, and the external environment together with its relationship to the central command unit of the cell, DNA/RNA, is described next.

DNA remains the central command unit in the cytoplasm of all cells. Not only is it linked to the activity of cells through proteins and the input of inorganic elements by feedback, but it is also limited by membrane activities. To maintain activity in a systematic fashion the command structure, DNA, receives information from the cytoplasm by sensing its contents, and this very sensing sends back instructions to maintain the system by synthesis. The system also sends signals to and from the cytoplasm and to and from the membrane to adjust the input or output of chemical elements to or from its contents, including those from the environment. The major communication for the organic chemistry connected to the DNA directly in primitive cells were free ions, the metallome, substrates, and coenzymes acting on transcription factors, while similar chemicals acting on receptors and pumps are the major communication to the membrane. Here particular attention is again drawn to the cytoplasmic network principally of the inorganic elements hydrogen, phosphorus, sulfur, sodium, potassium, chlorine, magnesium, and iron in primitive cells. These must be closely involved with protein synthesis. This network could not be greatly changed throughout evolution because it maintains the central cytoplasmic system including the membrane uptake. (Much of it is derivative from the nature of the sea, as explained above.) Thus new vesicle compartments, after O_2 rose, clearly needed communication modes, which would not be confusing to this network but would connect to it, linking to the central command of the DNA of each cell. One almost dominant new messenger in the evolution of eukaryotes mentioned above was the Ca^{2+} ion, which was previously just rejected from the prokaryote cytoplasm to the external environment and now also into the vesicles of eukaryotes. This gave the necessary energized large gradient using the binding capability of pulsed Ca^{2+} to activate cytoplasmic systems (Fig. 5.8). Once again this is not the place for details, but it is now known that the new Ca^{2+} message from outside the cytoplasm interacts directly with the old phosphate messenger system in the cytoplasm. The phosphate systems required iron for energy transformation and magnesium to catalyze many reactions. It is hard to see how eukaryote organization could have arisen without the utilization of the calcium gradient originally formed by prokaryotes as a way of sensing. These developments needed little use of oxygen. Note that even the earlier eukaryote cells were informed about the environment in this way.

5.9.2 New Oxidized Elements

One of the first chemicals to be oxidized by molecular oxygen was sulfide to sulfate (see Fig. 5.12). Sulfur isotope studies indicate the sulfate produced by organisms appeared some 3×10^9 years ago. Sulfate has value like oxygen itself. Although both new components can be introduced into organic compounds, they are also sources of energy. Do not forget that light produced oxygen (and then sulfate) simultaneously with reduced carbon compounds. The new source of energy for cells was then the reaction of the reduced organic molecules with oxygen or sulfate. Sulfate bacteria are well-known and are virtually anaerobes.

The element oxidized at about the same time was molybdenum from its lower oxidation state MoS_2 (see Fig. 5.12). In solution molybdate replaced tungstate forming Moco, and assisting O-transfer reaction, while possibly it replaced reduced vanadium in FeMoco, the coenzyme for nitrogen fixation. Molybdate became established at about 10^{-7} M in the cytoplasm. Subsequently zinc was released from sulfide and became a major signaling element and acid-base catalyst, but it is kept at or below 10^{-11} M as a free ion in the cytoplasm. Other elements with new uses mainly outside cells were selenium in the detoxification of peroxides and in assisting metabolism of the most easily oxidized halide, iodide. Iodide as iodine was incorporated into tyrosine derivatives and found value as the hormone, thyroxine. Vanadium as vanadate, which was able to bind peroxide, was involved in halide oxidation and halogen (bromium, chlorine) incorporation in organic molecules, for example, in toxins. These changes in element use in cells, especially in their vesicles and extracellular spaces, follow in time the succession of availabilities in the sea which is the succession of redox potentials (Fig. 5.12). The sea is closely at equilibrium at a given time with a redox potential increasing toward 0.8 V. One non-metal, which is now oxidized toward nitrate, is nitrogen. Today O_2 is above the NO_3^-/N_2 potential. Nitrate is already an energy source and a nitrogen source for many organisms, but it is dangerous in man.

Some oxidations of metal ions have been saved until the last. The first, which took place some 2×10^9 years ago, was the virtual removal of ferrous to be replaced by ferric

Fe^{3+} iron. Fe^{3+} is extremely insoluble at pH = 7.0, and this oxidation is a great problem for the required 10^{-7} M Fe^{2+} in the cell cytoplasm. All cells have devised intricate methods for scavenging of iron from the environment. A second oxidation from an increase in availability was the release of copper as Cu^{2+} from cuprous sulfide. Copper as Cu^{2+} was and is an extreme poison in the cytoplasm of cells and is always maintained at $<10^{-15}$ M. Its use is outside the cytoplasm.

Finally, although nickel (Ni^{2+}) and cobalt (Co^{2+}) became more available in the sea, the substrates for their enzymes largely disappeared, *e.g.*, CH_4, H_2, CO. Note that plants require no cobalt and all higher life requires very little of it as well as very little nickel.

Although all this basic chemistry is largely to be found in the periplasm of prokaryotes (bacteria) and in eukaryotes in their vesicles, the biggest change which oxidation finally generated some 1×10^9 years ago was of multicellular animals with a new large dependence on copper. The further developments of oxidized organic and inorganic chemistry is therefore best described in more detail under the heading of multicellular eukaryotes, although much of the new chemistry described is common to the prokaryote periplasm and the vesicles of single cell eukaryotes.

5.10 Multicellular Organisms

5.10.1 Cell/Cell Organization

Further development of internal organization from simple cells (prokaryotes) and then single cells with internal compartments (eukaryotes) must clearly be the introduction of novel chemistry, mainly oxidative, in new *external compartments,* as cells came together. Once cells became kept in a spatial organization they could then have separate functions in one organism. Such a development needed fixed external filaments and, because communication is central to organization, messengers external to and between the cell (Fig. 5.13). The filament connective tissue had to be open to adjustment with growth making growth signals essential. Hence new chemistry external to a cell had to provide the ability to make strong filaments, ways to make and break the filaments, and new messengers for communication, hormones, and transmission factors (Fig. 5.13). The parallel with the earlier evolution of internal compartments and filaments plus new internal messengers (Ca^{2+}) is obvious. Again the old chemical systems of the cytoplasm could not be changed much because the basic primitive metabolism in

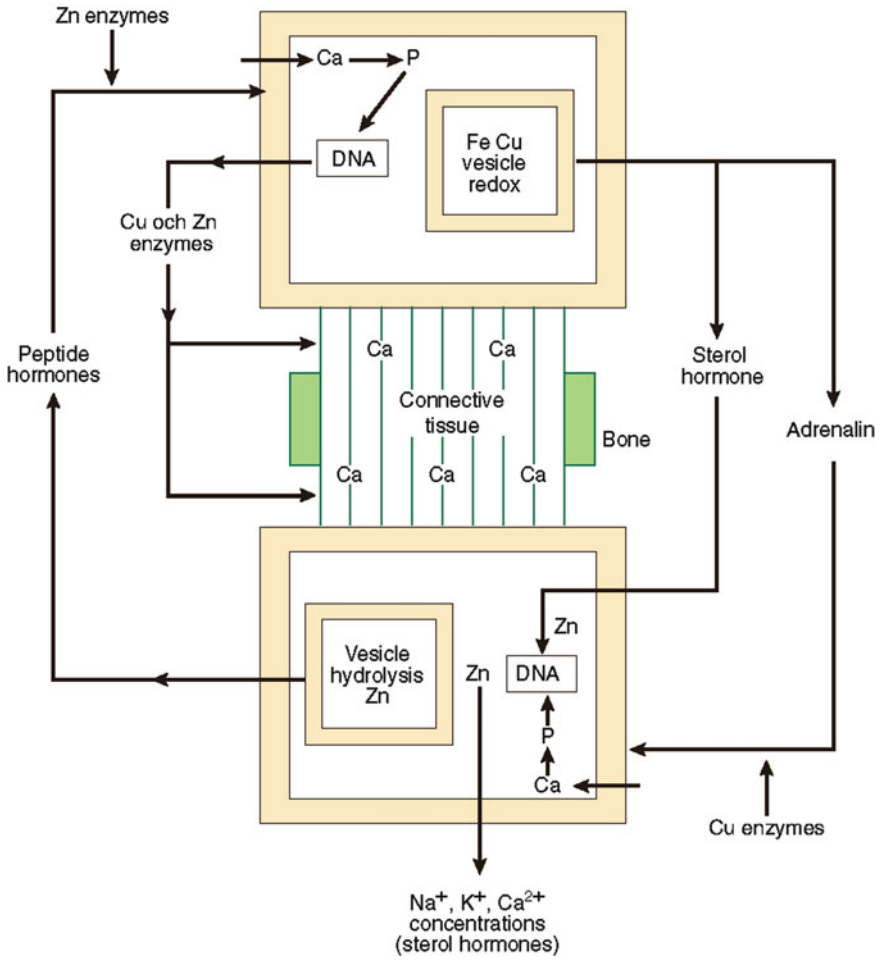


Fig. 5.13 The interlocking of cellular activities of differentiated cells as in higher organisms utilizing a greater variety of elements than in primitive cells

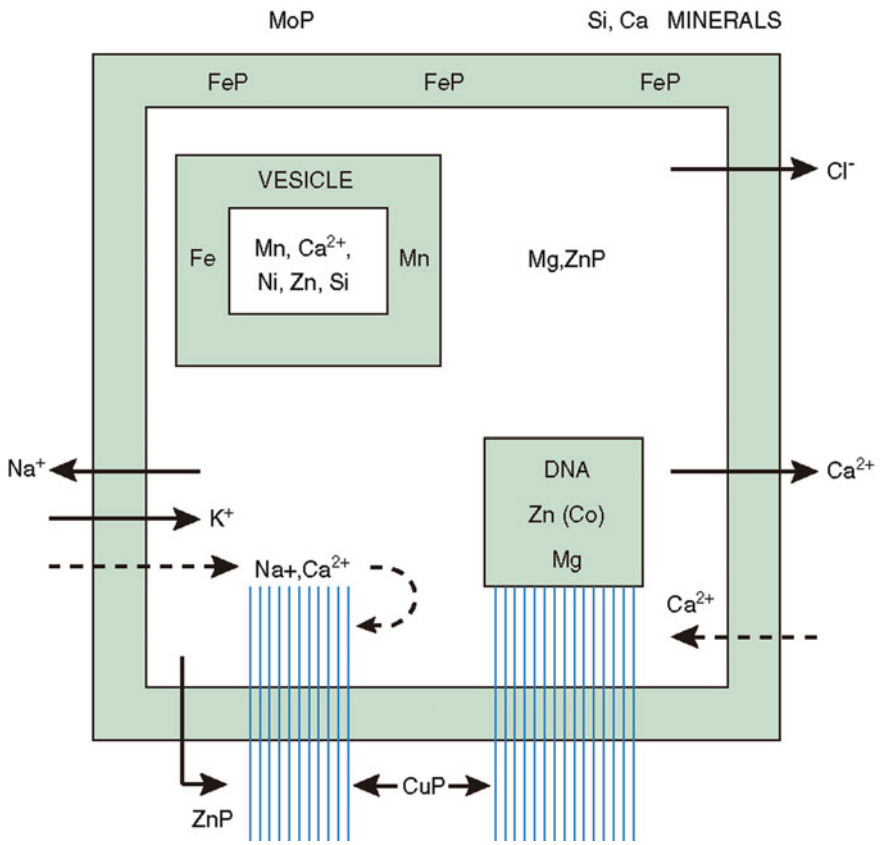


Fig. 5.14 An increase in cell-cell organization due to control by zinc and copper enzymes over the extracellular filaments and the use of messengers (see Fig. 5.13 and Table 5.5)

all cells had to be left as it was. It is not a coincidence that the appearance of this multicellular life coincided with a higher oxygen content of the atmosphere and the extensive use not only of higher oxidation states of iron (especially heme), molybdenum, manganese, and selenium, but also the use of zinc and copper. Zinc was released from its sulfide before copper but, by one billion years ago, both were freely available in the sea.

Let us consider the new filaments external to cells (Fig. 5.14). In order to gain rigidity they had to be cross-linked by oxidation. At first heme iron and manganese lignases, using peroxide, were employed in plants, but use of peroxide would have been dangerous in animals where the risk from incorrect oxidation is greater. Because external ferrous iron was not available, the obvious metal to use was copper, and it was found that copper became the major oxidase for cross-linked external filament production. Does the appearance of copper herald the coming of animals? Use was also made of sulfate in the sulfation of polysaccharides to keep an open network of connective tissue. Individual steps in the creation of this new external network and its activity cannot be timed, but a major feature became cell differentiation and then organ development in multicellular constructs. Zinc became the major hydrolase center for external breakdown of filaments to allow growth (Fig. 5.14).

5.10.2 Cell/Cell Communication: New Messengers

It has been seen that communication was based first inside the cytoplasm upon organic substrates and coenzymes and a few non-metal and metal ions such as phosphate, ferrous iron, and magnesium. This prokaryote organization was augmented by the use of the calcium gradient between outside and inside when larger cells (eukaryotes) with inner vesicles evolved. The possibility of triggering reactions due to simple ion messenger signals was then exhausted until oxidizing conditions arose. Oxidation allowed two new kinds of messenger: oxidized organic molecules such as adrenaline, sterols, thyroxine, and amidated peptides as seen in Table 5.5. The role of copper, iron, and selenium must not be missed both in their production and destruction. A messenger has to be removed as soon as its effect has been attained. An old way of making and destroying a messenger quickly was by hydrolysis. Hydrolysis allowed preparation of small peptides from larger units in vesicles, but it also allowed the destruction of the molecules from the vesicles once their message task had been completed outside cells. It was the increased availability of zinc that allowed this activity. The two elements dominating, respectively, hydrolysis (Mg^{2+}) and oxidation/reduction (Fe^{2+}) in the cytoplasm throughout evolution

Table 5.5 Organic messengers produced by oxidation (6)

Messenger	Production	Reception	Destruction
NO	Arginine oxidation (heme)	G-protein	? Heme oxidation
Sterols	Cholesterol oxidation (heme)	Zn-fingers	Heme oxidation
Amidated peptides	Cu oxides	(Ca^{2+} release)	Zn peptidases
Adrenaline	Fe/Cu oxidase	(Ca^{2+} release)	Ca enzyme?
δ-OH Tryptamine	Fe/Cu oxidase	(Ca^{2+} release)	Cu enzyme?
Thyroxine	Heme (Fe) peroxidase	Zn finger?[a]	Se enzyme
Retinoic acid	Retinol (vitamin A) oxidation	Zn finger?[a]	?

[a]In the nuclear receptor super family of transcriptional receptors

became augmented by the two elements Zn^{2+} for hydrolysis and Cu^{+}/Cu^{2+} for oxidation/reduction in vesicles, periplasm, and outside cells. The metals cobalt and nickel almost dropped out of use in higher organisms because their redox chemistry is much less valuable in oxidative as opposed to reductive chemistry. Of course the redox chemistry of non-metals also changed, and now the value of selenium in peroxide chemistry generally and opposite de-iodination developed with the ability to use the iodinated organic hormone thyroxine as a messenger in animal life. In plants, molybdenum liberated from sulfide became valuable in oxidation as well as in the creation of the hormone absiscic acid from its aldehyde.

Just as the very early use of kinetically isolated new compounds was possible in the cytoplasm of prokaryotes, *e.g.*, heme and chlorophyll, such a development was now possible in eukaryotes though it could not easily relate to oxidative chemistry. The major change seen is of zinc in transcription factors linked to messengers such as sterols at very low free concentrations of $<10^{-10}$ M. It is our belief that in long-lasting cell systems zinc acts like Mg^{2+}, but without interference due to its high binding constants to very different binding sites and much slower exchange. Zinc could be said to have become an internal connector of cell growth and development.

Thus oxidative chemistry outside the cytoplasm added on to reductive chemistry inside the cytoplasm was forced by the "accidental" conversion of the chemical nature of sea and atmosphere by molecular oxygen. There was no other way organization (evolution) could arise except by utilizing novel oxidative chemistry in new compartments.

5.11 The Non-metal Balance: The Metabollome

The handling of non-metals by organisms is more difficult to describe than that of metals for the very obvious reason that much of the establishment of their steady state levels is through kinetic feedback controls as seen in Fig. 5.6. However, they also have the common problem of the sharing of energy and chemicals between pathways so that different pathways must have almost equal access to the non-metals. Thus there is a network comparable but more extensive than those for magnesium and calcium, for example, Figs. 5.7 and 5.8. Consider the case of phosphate in nucleotide triphosphates like ATP. These coenzymes transfer phosphate and energy in many chemical steps leading to nucleotides, proteins, saccharides, and lipids. Their networks have to communicate with one another. As a consequence binding constants for them, usually with Mg^{2+} attached, are all close to 10^{3} M and their concentrations, in the locality where they are used, are also close to 10^{-3} M. The suggestion is that the non-metal concentrations in certain compounds are relatively fixed. The profile of these compounds is called the metabollome.

Now the other free non-metal compounds which must be of relatively fixed concentration are the other mobile coenzymes that distribute carbon, hydrogen, nitrogen, and sulfur. They are among the common vitamins. Of especial interest here and associated with energy distribution, much as ATP is thus associated, are the hydrogen/electron carriers. There are two distinct cytoplasmic networks to consider just as there were two kinetically distinct networks for free Fe^{2+} and iron in heme. Here the two networks are at a potential close to −0.45 V NAD(H), (NADP(H)) and at around 0.0 V, glutathione. Note that the running of these two H(*e*) carriers about 0.5 V apart is close to the running of the two primitive prebiotic non-metal reactants, the H_2/H^+ and the H_2S/S_n redox couples. It was these elementary couples and then the corresponding coenzymes that supplied the initial feedstock of cellular life. These coenzyme pods connect not just to an energy supply but also to the distribution of hydrogen. They are all held close to millimolar in concentration in the cytoplasm of all active cells.

It would appear that the feedstock of non-metals from the environment may be variable, unlike that of metals, but for certain components of the non-metal system there is a fixed profile. This profile is the most intriguing part of the metabollome. Again this is suggestive of a selective

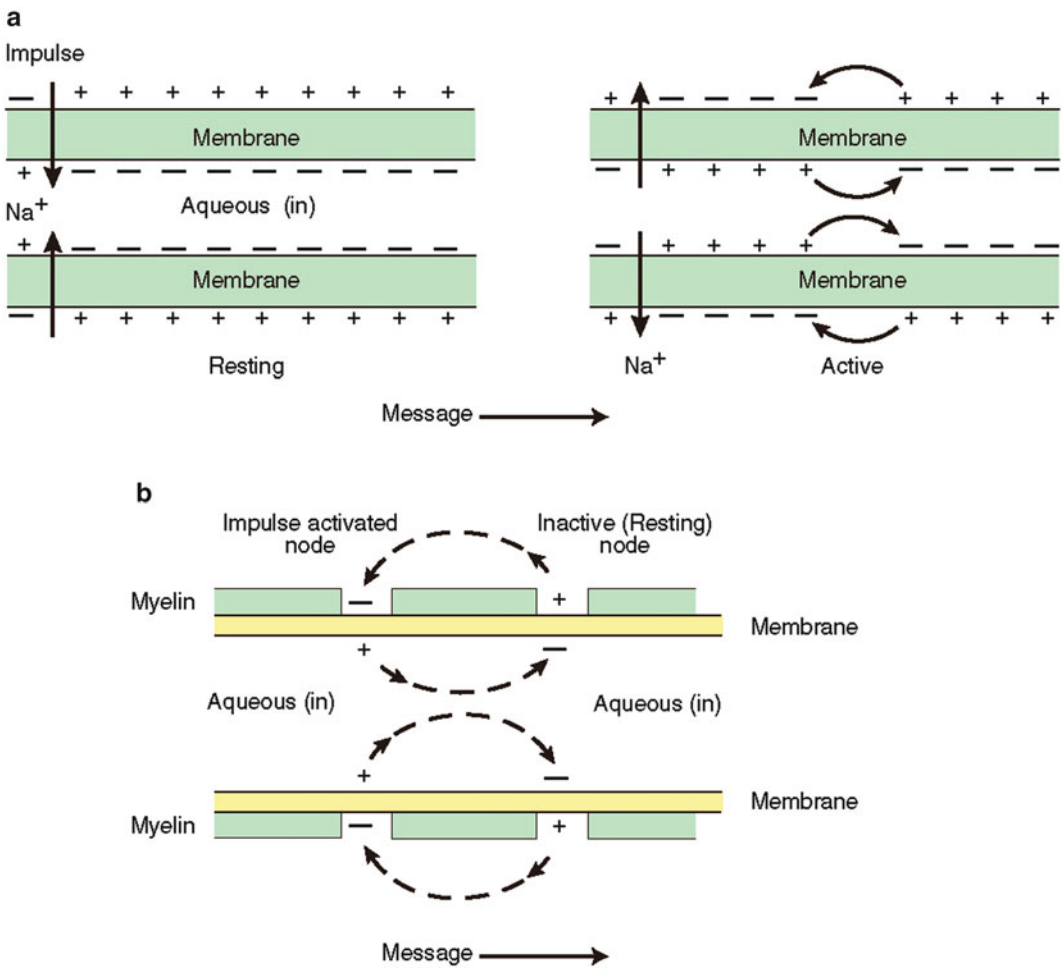


Fig. 5.15 A nerve cell in schematic representation

quantitative requirement for a chemical system of reactions on which a code was superimposed to make reproductive organisms. Maybe the metaphoric original "watchmaker" of Dawkins was blind in some sense, but he must have had some other chemical sense which drove the beginnings and even evolution as if in a tunnel.

Once the code was in place and the cytoplasmic activity established for all time, new chemistry of the non-metals appeared in the vesicles. Examples are the handling and use of sulfate in connective tissue, of the dinitrification processes for nitrate metabolism in the periplasm, peroxide chemistry in defense mechanisms inside cells, the oxidation of halides to create a range of halogenated organic molecules in vesicles, and the development of organic messengers (Table 5.5). How did these changes (see Fig. 5.10) drive evolution? Are there ways by which geological changes interact more directly with the genetic apparatus than through blind mutation of the whole genome? All novel chemicals are poisons, but they affect localized genes more than others. Do these genes mutate most rapidly to give rise to useful novel chemical features?

5.12 The New Command Center: The Brain

Once a multicellular structure and then organs appeared, increasing the size of organisms enormously, there were more problems than just producing extracellular matrices and their management and a message system between cells. With gross differentiation there was the logistical need to evolve a central organization of the parts, which were now organs. This need was the greatest in animals, which are scavengers rather than synthesizers such as are plants. Strikingly, the answer found was the brain as an organ. Initially there was development from a new physical type of compartment or cell, a nerve, which grew as a fast communicating compartment linking, in long cylindrical structures (Fig. 5.15), one set of organs to another. The message system pressed into service in this construct was the Na^+/K^+ current. Although it utilizes a new Na^+/K^+ ATPase pump, this message system is derived directly from the required rejection of Na^+ and Cl^- by primitive cells. However, $Na^+/K^+/Cl^+$ flows are not a complete answer

because when these ions cross membranes they produce electric potential switches but no chemical trace as they bind so poorly. The solution found was to connect the potential switch to a calcium message and the calcium message to a release of new organic chemical messenger molecules at nerve cell junctions. Note that the use of Na^+, K^+, and Cl^- to bring about fast electric potential switches of Ca^{2+} over long distances is the only possible mechanism open to cell systems. Cells cannot use electronic circuits. The new command center, the brain, is baced upon a coordinated set of such nerve cells.

It did not matter that this new command center, the brain, relies on DNA genes for its basic units, proteins. After its creation an animal was no longer dominated by genes but by the interaction with the environment. In fact because it can sense the environment, the machinery of the brain is capable of responding to control it. The further appearance of organization outside the most complex biological organism is seen in man's society and his industry, which can capitalize on the whole 92 elements of the periodic table in new external compartments using new reducing conditions to produce metals in what is the logical conclusion to evolution on Earth. The metals have new functional value in electronics. Further advances of materials allow the use of light as a messenger. Clearly the computer becomes the next new electronic command center. All these are but logical developments of organization. However, while rejoicing in this end product of chemical evolution, it must be remembered that the cytoplasm of all cells is a remnant of the primitive environment and cannot tolerate many of the new chemicals man can generate. The chemical system called living is robust but it needs protection. Many new protective devices have been introduced during the course of evolution, but there is no space to discuss them here. Note, however, that rare elements—platinum, gold, mercury, and bismuth—never used by cells become, in man's hands, the basis of medicines, though they are, in fact, poisons.

5.13 Survival: Ecology and Symbiosis

Now we have dealt with the progression of compartments from single cell prokaryotes, to single cell eukaryotes, to multicell organisms, to organisms with organs and brains. We must pause now to ask if these systems are in competition, or are they really a natural way for cooperation to evolve in which one kind of cell system is best for one kind of task, with the ultimate objective of maximal use of the environmental chemicals in an overall ecosystem. In such an ecosystem symbiosis reigns supreme. Can we look at these developments of the chemistry and physics of cells (Fig. 5.16) and see them as an inevitable progression, dependent on the changes of environmental chemistry and the logistics of organization? Maybe there was no way to evolve novelty from an organic chemical cell except to utilize the possible oxidative chemistry and increase in compartments. However, there is a problem of the economy of such a system. The ability of one organism to evolve in an environment becomes increasingly difficult the more complicated the chemistry and physical constructs it uses, because it must keep all that came before and all that is new in harmony. Intrinsically a prokaryote is a better simple synthesis unit and a better stripped-down energy transducer than a eukaryote, and in turn better than a multicellular organism, because it does little else except synthesize. A eukaryote is a better consumer, even of prokaryotes, and has greater potential. Moreover, the eukaryote has a longer lifetime and needs greater protection. It follows that cooperative life-forms will have a greater survival chance than competitive styles, and this is seen in the successive development of symbiosis and of ecosystems. We see prokaryotes (organelles) in all eukaryotes and a gross dependence of eukaryotes on prokaryotes for N_2 fixation. Man as the last product (thus far) of evolution is not able to synthesize many of the basic requirements of the cytoplasm and requires essential amino acids, fats, coenzymes (vitamins), and saccharides from lower organisms. All minerals are collected by plants for animals. This again is a logical outcome of cooperative organization, not of competition. In the context of this chapter lies the requirement for the supply of 20 essential elements, and while we may concentrate upon the dependence of one phyla upon another for their supply, we must not miss the dependence of the biological ecosystem upon the geosphere. The environment is not separable from the overall chemical systems of life and its evolution.

5.14 Human Health and the Environment

Why is this account of evolution important in the context of human health and our environment? It clearly shows that there has been at all times including today a close interaction between changes of elements in the environment and in organisms. This interaction has led to human life in today's environmental context. The well-being of humans is also dependent on other organisms which interact with them, internally, some as parasites, and externally. Many other factors are part of our food chain including minerals, dependent on element availability. Perhaps a way to illustrate the various collectivities between organisms themselves and their environment is to replace the Darwinian picture of simple linear inheritance along an isolated branch of a tree with a complex evolutionary development of a network. The network involves the environment as well as cross-connections between other organisms as shown in Fig. 5.17(a), (Williams and Rickaby 2012), see also Fig. 5.16.

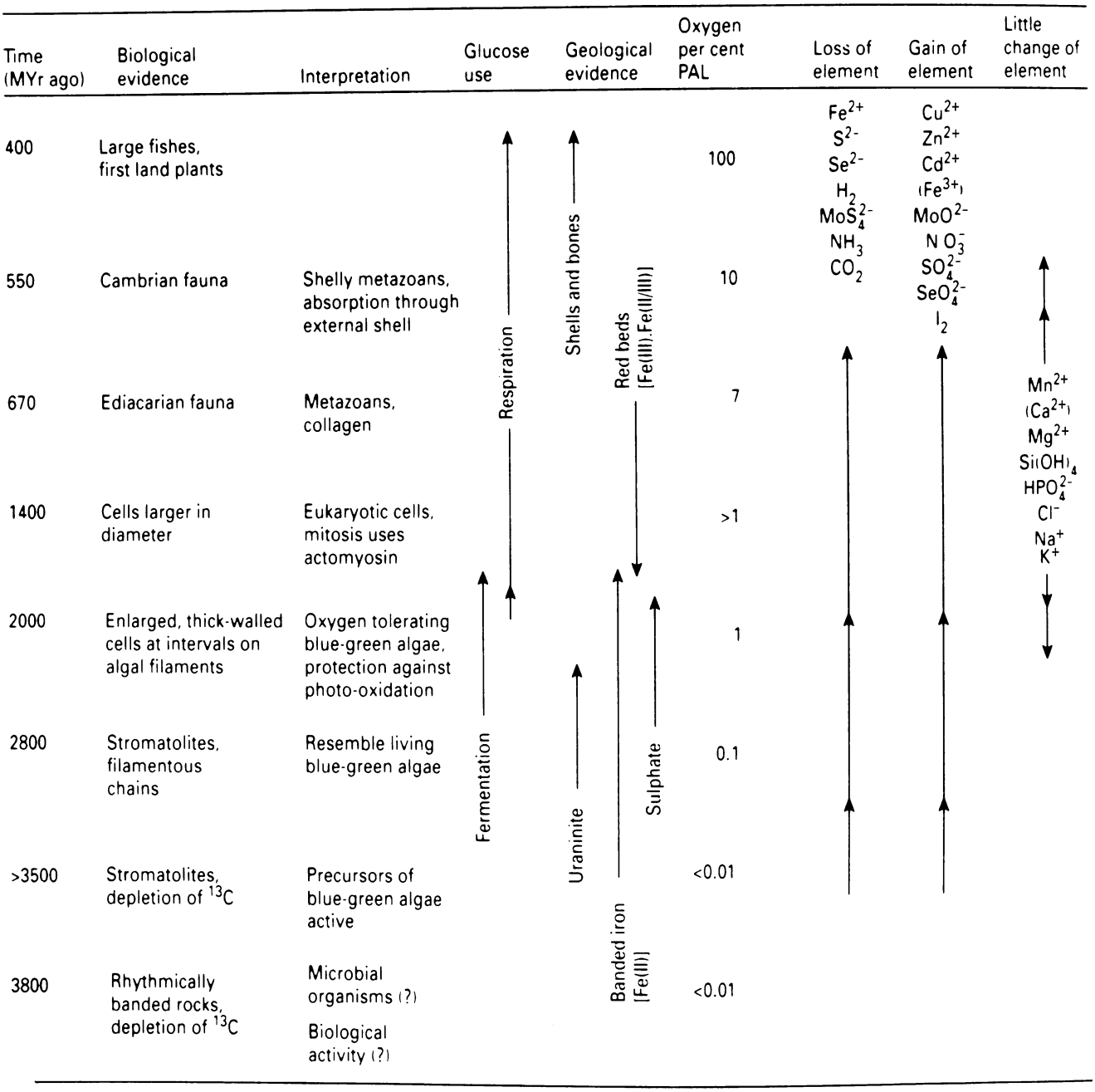


Fig. 5.16 The simultaneous sequence of the evolution of Earth and life

In Fig. 5.17 Darwin's own picture of a tree of evolution is given as an aside. The main figure clearly illustrates the importance of symbiotic development, Margulis (1999). Human health is then a problem of our well-being and that of external and internal symbionts while it must resist certain organisms and parasites. It is also necessary to be mindful of the effects of any changes in environmental chemistry.

This view of evolution poses an age-old question. Is it the case that evolution is just a matter of random selection of advantageous characteristics of an environment including selection of supportive organisms? Here the conventional view is that slowly gene mutations lead to better adapted organisms in an environment. The mutations are considered to be random. Now recently it has become apparent through studies of the immune system of man and of his resistance to drugs that this cannot simply be true of all evolutionary events. The immune studies show that the response to antigens in humans can be localised in the DNA to selected sensitive DNA regions, but we know that response is limited to non-reproductive cells. The second example shows that

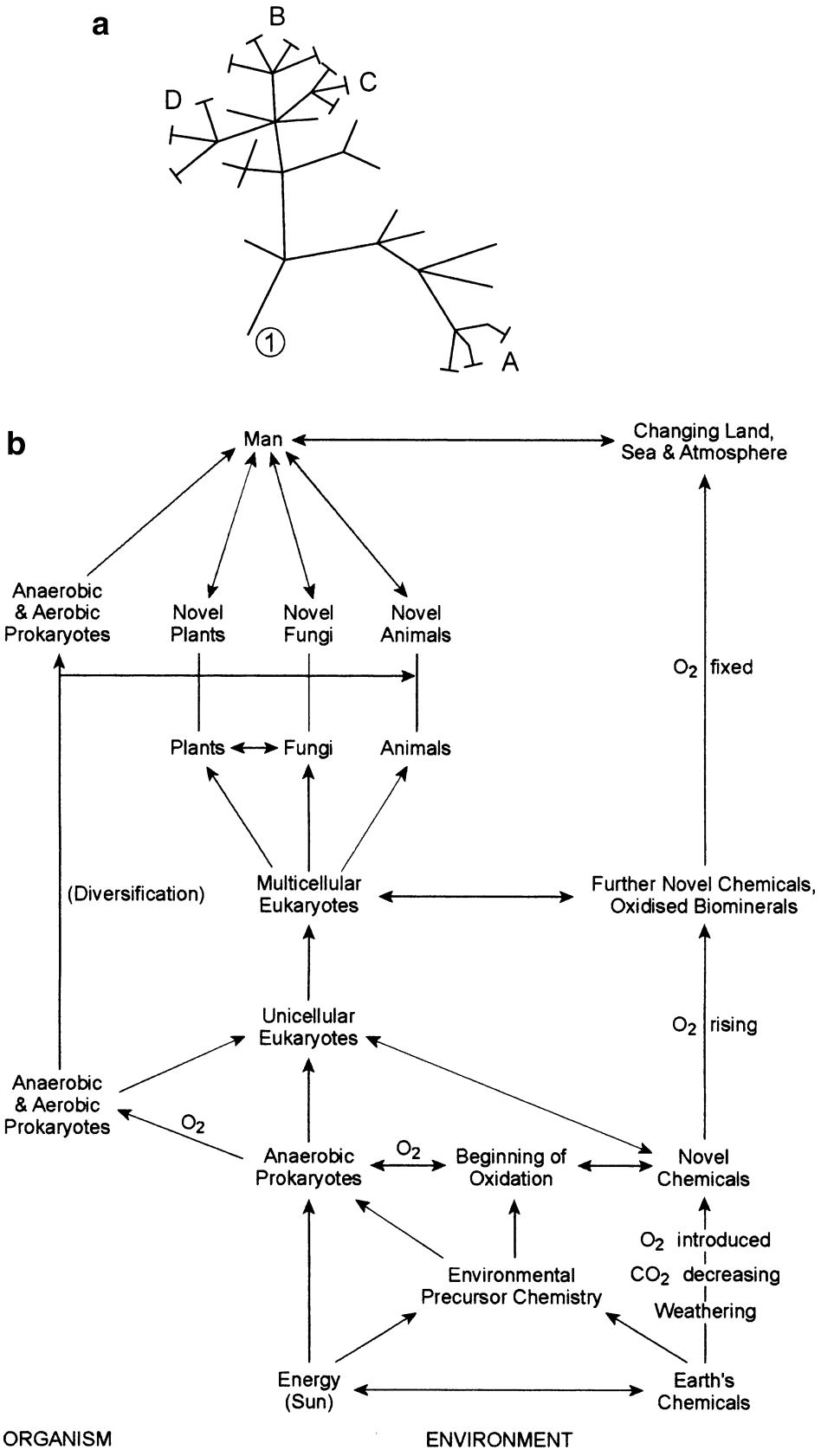


Fig. 5.17 **a** and **b**: A novel way of considering evolution which includes the changes in environmental and organism chemistry and increasing symbiosis, Williams and Rickaby (2012). This outline is to be contrasted with that of Darwin who sketched out the first idea of a tree of noninteracting branches, see inset. The random Darwinian approach has been greatly elaborated by modem biologists and geneticists. The chemical changes shown stopped shortly after 0.5 billion years ago. Subsequent evolution has been largely by random mutation along Darwinian branches. However man is now introducing chemical changes in his environment which could cause evolutionary changes like those of the period between 3.5 and 0.5 billion years ago

Table 5.6 Local DNA (gene) duplication

Organism	Numbers of very similar proteins					
	Peroxidase	P450	Cu oxidases	Fe oxidases	Zn fingers	Calmodulin
		Heme	EC: 1	2 OG Fe(II)		
Eubacteria	1	1	3	1	5	0
Yeast	1	3	9	1	3.73	4
Arabidopsis	95	268	111	116	1,631	80
Homo sapiens	0	70	21	9	2,984	160

Note: proteome size from top 3,500, 6,000, 32,615, 37,742

DNA can be adjusted locally by resistance such as that to a drug is rapidly inheritable. A particular study is that of resistance to hydroxyurea, Arit et al. (2011). Hydroxyurea acts on the DNA to produce locally a multiplicity of highly selected sections of DNA. This multiplicity allows relatively fast subsequent mutation of the local DNA fragment. In the context of Fig. 5.17 we consider that the action of a drug such as hydroxyurea is likely to be very similar to the effect of the introduction of oxygen which was at first a poison, or its products of novel molecules in the environment. We have tested this hypothesis by examining multiplication of the DNA of metalloenzyrnes which remove these molecules. Here we had the benefit of the bio-informatic approaches of Andreini et al. (2009) and of Dupont et al. (2010). Table 5.6 illustrates the result that as organisms increased in complexity the numbers of local gene (DNA) duplicates increased very greatly, Williams and Rickaby (2012). The duplicates represent members of the same family of enzymes with a variety of closely related functions which are different in different phyla. For example the number of oxidases in four classes of iron and copper enzymes for removing of small molecules is very much greater in advanced plants (*arabidopsis*) than in animals (*homo sapiens*). This is very much in accord with the need for protection in a plant from external molecules which has no immune response. On the other hand advanced animals have much the greater number of signalling proteins, Table 5.6. The greater number of zinc fingers has the role of transcription factors responding to hormones while the calcium proteins, calmodulins, trigger the animal cells' response to external threats. Note all the large numbers of localised multiplicity are related to stress in the environment where animals have to respond to external dangers and to search for food. Today the benefit of a rapid response is important for the ability to be able to respond to the impact of changes in the environment. It is the case that as well as localised multiplicity the DNA is now open to localised gene multiplication by fast mutation. This could be a way of evolving in the face of environmental change. We must also add symbiosis as a way of carrying some functions of one DNA to assist another organism with a separate DNA, Fig. 5.17. Each separate organism gains from reduction in complexity.

Can mining on Earth by man and the subsequent release of new elements of the Periodic Table into the environment cause rapid evolution by multiplication and mutation of DNA? The effects will be seen first in the simplest organisms but do remember that they may cross-interact with higher organisms as in Fig. 5.17.

5.15 Conclusion

The central theme of this chapter is that the primitive sea greatly influenced the way the chemical elements could be used by a chemical system which became primitive life (Fig. 5.5). We consider that the profile of free elements in that sea self-generated a chemical system of reactions that became the one we see in the cytoplasm of all cells even today. Hence, fundamental to life before there was a code was an organized set of reactions that can be characterized by the profile of its free ion exchanging inorganic content—its cytoplasmic free metallome and metabollome. Metallomes and metabollomes, free and bound, have evolved largely due to new compartments and in oxidative possibilities. The way in which this was done is then the major story of cell evolution even though it relied on codes and cooperation to secure survival. The appearance of species is a decoration on this logical progression. The final twist is the turn around to extreme reductive chemistry which has allowed man to evolve in the new industrial world around us. Much as the geochemical changes in evolution changed the environment they made possible the changes of living forms so today man-made chemical changes could affect human health.

See Also the Following Chapters. Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 7 (Biological Functions of the Elements).

Further Reading

Anderson GM, Crerar DA (1993) Thermodynamics and geochemistry: the equilibrium model. Oxford University Press, New York

Berkner IV, Marshall LC (1965) On the origin and rise of oxygen concentration in the Earth's atmosphere. J Atm Sci 22:225–261

Caporale LH (ed) (1999) Molecular strategies in biological evolution. Ann NY Acad Sci 870: 1–21
Cox PA (1995) The elements on earth. Oxford University Press, New York
Depew DJ, Weber BH (1995) Darwinism evolving. MIT Press, Cambridge, MA
Fraústo da Silva JJR, Williams RJP (2001) The biological chemistry of the elements, 2nd edn. Oxford University Press, Oxford
Margulis L (1999) Symbiotic planet. Basic Books, New York
Martell AE, Hancock RD (1996) Metal complexes in aqueous solution. Plenum Press, New York
Neidhardt FC (ed) (1996) *Escherichia* and *Salmonella*, 2nd edn. American Society for Microbiology Press, Washington, DC (This is a detailed compilation of the properties of prokaryotes)
Stryer L (1995) Biochemistry, 4th edn. W. H. Freeman, New York
Turner D, Whitfield M, Dickson AG (1981) Inorganic speciation of trace metals in the sea. Geochim Cosmochim Acta 45:855–890
Williams RJP (1970) The biochemistry of sodium, potassium, magnesium and calcium. Q Rev Chem Soc Lond 24:331–360
Williams RJP, Fraústo da Silva JJR (1996) The natural selection of the chemical elements. Oxford University Press, Oxford
Williams RJP, Fraústo da Silva JJR (1999) Bringing chemistry to life. Oxford University Press, Oxford

Recent New References

Andreini C, Bertini I, Cavallero G, Holliday GL, Thornton JM (2009) Metal-MACiE: a database of metals involved in biological catalysis. Bioinformatics 25:2088–2089
Arit MF, Ozdemir AC, Birkeland SR, Wilson TE, Glover TW (2011) Hydroxyurea induces de novo copy number variants in human cells. Proc Natl Acad Sci (USA) 108:17360–17365
Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anolles G (2010) History of biological metal utilisation inferred through phylogenomic analysis of protein structures. Proc Natl Acad Sci (USA) 107:10567–10572
Williams RJP, Rickaby REM (2012) Evolution's destiny: co-evolving chemistry of the environment and life. Royal Chemical Society, Cambridge

Uptake of Elements from a Biological Point of View

6

Ulf Lindh

Contents

The utilization of an element in biology is intimately dependent on its uptake into the living organism. A lot is known of the qualitative aspects of uptake; for example, common sense tells us that what originates in the geological background has to be transported through the soils and presented to plants in a convenient form for uptake. These processes are affected by physicochemical factors due to Nature itself and the increasing pressures from human activities.

From a biologist's perspective, the uptake process is extremely fascinating. Current knowledge tells us that organisms need about 20 of the naturally occurring elements found in the periodic table. Over the course of evolution, mechanisms have developed for the uptake and utilization of elements in organisms that are more or less specific for each. Additionally, those elements that are not essential or even detrimental to the organism are excluded, usually in an efficient manner. This chapter deals with the general aspects of element uptake as well as principles of exclusion. In addition, examples will be given pertaining to some major, minor, and trace elements collectively described as essential. This chapter begins with a discussion of the periodic table and what is meant by *essentiality*.

6.1 Essentiality of Elements

A usually ignored fact in biology is that explanations for the behavior of elements can be found in the periodic table itself. Among the elements known to be involved in biology, 11 appear to be approximately constant and predominant in all biological systems. The human body is comprised of about 99.9% of the 11 elements, but surprisingly only four of them – hydrogen, carbon, nitrogen, and oxygen – account for 99% of the total. These four elements, the major elements, comprise the bulk of living organisms. In addition to these elements, there are the minor elements – sodium, magnesium, phosphorus, sulfur, chlorine, potassium, and calcium (also called *electrolytes*). The minor elements

U. Lindh (✉)
Biomedical Radiation Sciences, Rudbeck Laboratory, Research in Metal Biology, Uppsala University, Uppsala, Sweden
e-mail: Ulf.Lindh@ibg.uu.se

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_6,

appear in much lower concentration than major elements. One group of elements has still to be defined: the trace elements. From an analytical chemistry standpoint, trace elements would be described as elements appearing in low concentrations in living systems (i.e., <100 mg/kg). In biology, however, trace elements would be defined by exclusion; for example, a biological approach begins by excluding the major elements as well as the minor elements. Furthermore, group 18 elements (the noble gases) are excluded due to their disinclination for chemical reactions, a property that makes these elements less likely to be a factor in biological functions. Depending on how many elements are considered naturally occurring, the trace elements thus constitute the remainder of the periodic table (i.e., 73 or 75 elements). Most of the elements of the periodic table, then, are trace elements in the eyes of a biologist. Surprisingly, this exclusion definition coincides closely with the one of analytical chemistry. Most of the trace elements appear in biology at concentrations below or well below 100 mg/kg.

Essentiality is usually defined in an operational way, based on early protein chemistry. More stringent criteria have evolved as our knowledge has improved. A trace element can be considered essential if it meets the following criteria: (1) it is present in all healthy tissues of all living things, (2) its concentration from one animal to the next is fairly constant, (3) its withdrawal from the systems induces reproducibly the same physiological and structural abnormalities regardless of the species studied, (4) its addition either reverses or prevents these abnormalities, (5) the abnormalities induced by deficiencies are always accompanied by specific biochemical changes, and (6) these biological changes can be prevented or cured when the deficiency is prevented or cured. It is obvious that the number of elements recognized as essential depends on the sophistication of experimental procedures and that proof of essentiality is technically easier for those elements that occur in reasonably high concentrations than for those with very low requirements and concentrations. Thus, it can be expected that, with further improvement of our experimental techniques, additional elements may be deemed essential.

These six criteria might appear to be too stringent, and in some cases they must be modified. Most of the trace elements essential to both plants and animals are found in the first row of the transition (redox) metals. Zinc is not included in the transition metals, and it does not take part in redox reactions, which is an important property in biology; however, zinc is a good Lewis acid. All of the bulk elements are non-metals. The minor elements include metals as well as non-metals, with only one oxidation state available. Metals are the dominant components of the essential trace elements, but some very important trace elements are non-metals, such as selenium and iodine. Boron and silicon are non-redox non-metals, and both of them are acknowledged as being essential. Boron, in fact, has been shown to be essential to plants, although it is found in appreciable concentrations in animals as well. The functions of these elements will be discussed in greater detail in Chap. 7.

6.2 General Aspects of Element Uptake

The uptake of elements is a process that may vary considerably depending on the complexity of the living system being considered. Unicellular organisms account for the simplest processes, but in complex organisms several aspects of the uptake process must be considered. In humans, for example, the primary uptake process takes place in the gastrointestinal tract, predominantly in the duodenum and first part of the jejunum. Elements taken up have to be transported across the mucosal cells of the intestines to reach the bloodstream, from which they are transported to the liver, where the elements are isolated and delivered into the main bloodstream. After being transported to the organs that will utilize them, these elements must then enter the cells of these organs. If the final target is not found inside the cell, then further transport across additional membranes may be required. Let us review the general principles of transport across cell membranes. A cell or an organelle cannot be entirely open or entirely closed to its surroundings. Although the cell interior must be protected from certain toxic compounds, metabolites must be taken in and waste products removed. Because the cell must contend with thousands of substances, it is not surprising that much of the complex structure of membranes is devoted to the regulation of transport.

6.3 The Thermodynamics of Transport

Before considering specific mechanisms of transport, it is useful to review some general ideas. The free energy change, ΔG, for transporting one mole of a substance from one place with concentration C_1 to another place with concentration C_2 is

$$\nabla G = RTln\frac{C_1}{C_2} \tag{6.1}$$

where R is the gas constant and T the temperature. According to Eq. (6.1), if C_2 is less than C_1, then ΔG is negative and the process is thermodynamically favorable. When more substance is transferred (between two finite compartments), C_1 decreases and C_2 increases, until $C_2 = C_1$. Now, the system is at equilibrium, and $\Delta G = 0$. This equilibrium is the ultimate state approached by

transport across any membrane. The concentration of any substance traversing the membrane will end up the same on both sides. In kinetic terms, if the molecules enter the membrane randomly, the number entering from each side will be proportional to the concentration on that side. Once the concentrations are equal, the rates of transport in the two directions will be the same; consequently, no net transport occurs. This equalization can be sidestepped under three conditions, each of which is important in the behavior of membranes.

A substance may be bound by macromolecules restricted to one side of the membrane. It could also be chemically modified once it has crossed the membrane. Compound M could be more concentrated inside a cell than outside, but much of M could be bound to cellular macromolecules or could have been modified. This part of M is not included in Eq. (6.1), which only deals with the free M, implying, then, that the concentration would be equal on both sides at equilibrium. We can use oxygen in red blood cells to illustrate this principle. Measurements would indicate that the total concentration of molecular oxygen is lower in the blood plasma than in the red cells. Included in the total concentration of the red cells, however, is the portion bound to hemoglobin. The concentration of free dioxygen is still the same in the red cells and plasma at equilibrium.

A membrane is often characterized by an electrical potential governing ion distribution. The well-known principle saying that equal charges repel each other and unequal charges attract each other may now be used to show a simple example. The negatively charged inside of a cell tends to attract cations and to drive out anions. Mathematically, the free energy change for transport over a membrane is

$$\Delta G = RT \ln \frac{C_1}{C_2} + ZF\Delta\Psi \tag{6.2}$$

where Z is the charge of the ion, F is the Faraday constant (96,485Jmol^{-1} V^{-1}) and $\Delta\Psi$ is the transmembrane electric potential (in V). If Z is positive and $\Delta\Psi$ negative (with the inside negative relative to the outside), transport of cations into the cell is favored. The concentration difference of ions across the membrane in most cells is kept at more than ten times, implying that active transport is a major energy-requiring process in biology.

Usually Eq. (6.1) does not reflect the real situation well enough and must be modified accordingly:

$$\Delta G = RT \ln \frac{C_1}{C_2} + \Delta G' \tag{6.3}$$

where $\Delta G'$ may correspond to a thermodynamically favored reaction. Adenosine triphosphate (ATP) hydrolysis coupled to the transport might be such a situation. Equation (6.3) is clearly a generalization of Eq. (6.2) that now allows a variety of processes to participate in the transport.

Equations (6.2) and (6.3) convey the message that transport across membranes (in and out of cells) can take place against unfavorable concentration gradients. The sodium–potassium pump provides continuous import of potassium and export of sodium, thereby maintaining the concentration difference between inside and outside. Following is a review of the mechanisms by which substances are passed through membranes.

6.3.1 Passive Transport: Diffusion

Passive transport or passive diffusion occurs due to the random walk of molecules through membranes. The process is the same as the Brownian motion of molecules in any fluid and is called *molecular diffusion*. During passive transport, the diffusing substance ultimately reaches the same free concentration inside and outside the membrane. The net rate of transport, J (molcm^{-2} s^{-1}), is proportional to the concentration difference ($C_2 - C_1$) over the membrane:

$$J = -\frac{KD_1(C_2 - C_1)}{l} \tag{6.4}$$

where l is the thickness of the membrane, D_1 is the diffusion coefficient of the diffusing substance in the membrane, and K is the partition coefficient for the diffusing material between lipid and water (the ratio of solubilities of the material in lipid and water). For ions and other hydrophilic substances, K is a very small number. Diffusion of such substances through lipid membranes is thus extremely slow. In agreement with Eqs. (6.1) and (6.4) says that net transport stops when $C_2 = C_1$. If C_1 and C_2 are expressed in molcm^{-3} and l in centimeters, then D_1 has the dimension cm^{2}s^{-1}. Note that D_1 is not the same as the diffusion coefficient (D) that the same molecule would have in aqueous solution. D_1 depends not only on the size and shape of the molecule but also on the viscosity of the membrane lipid.

K, D_1, or the exact thickness of the membranes involved are not usually known, so the rate of passive transport is often described in terms of the permeability coefficient, P, which can be measured by direct experiment:

$$J = -P(C_2 - C_1) \tag{6.5}$$

By comparing Eqs. (6.5) and (6.4), we see that P is expressed by:

$$P = \frac{KD_1}{l} \tag{6.6}$$

with the dimensions cms^{-1}.

Table 6.1 Permeability coefficients from some ions and molecules through membranes

P (cms^{-1})	Membrane phosphatidylserine	Human erythrocyte
K^+	$<9 \times 10^{-13}$	2.4×10^{-10}
Na^+	$<1.6 \times 10^{-13}$	10^{-10}
Cl^-	1.5×10^{-11}	1.4×10^{-4a}
Glucose	4×10^{-10}	2×10^{-5a}
Water	5×10^{-3}	5×10^{-3}

[a]Facilitated transport. Note that whenever facilitated transport is encountered, the permeability coefficient rises dramatically

Table 6.1 shows the permeability coefficients for a number of small molecules and ions in membranes. The low *P* values of the ions are expected, because ions, as already mentioned, have low values of *K*; however, the relatively large permeability value for water is conspicuous. Biological membranes are not, in fact, very good barriers against water, the reasons of which are not entirely clear, but they have the obvious advantage of allowing the ready exchange of water with their surroundings.

6.3.2 Facilitated Transport: Accelerated Diffusion

The functional and metabolic needs of cells often require transport that is more efficient than passive diffusion. The adequate handling of catabolically produced CO_2 is coupled to the red blood cell exchange of HCO_3^- and Cl^-. In respiring tissues, CO_2 enters the red cell and is converted to HCO_3^- by carbonic anhydrase, an enzyme that is zinc dependent. HCO_3^- leaves the red cell, and Cl^- enters. The HCO_3^- is transported to the lungs in the plasma where it enters the red cell, and Cl^- is driven out. Inside the red cell, HCO_3^- is again converted to CO_2 by carbonic anhydrase, leaves the cell, and is exhaled. The permeability coefficients for Cl^- and HCO_3^- in red cell membranes are about $10^{-4}cms^{-1}$, or about ten million times greater than the permeability coefficients for such ions in pure lipid bilayers such as the artificial phosphatidylserine membrane described in Table 6.1. Some special mechanism, then, is required to account for this difference. Two general types of facilitated transport, or facilitated diffusion, are known. One type, which is responsible for the rapid transport of Cl^- and HCO_3^- through red cell membranes, involves pores formed by transmembrane proteins (Fig. 6.1a). The other type is mediated by transmembrane carrier molecules (Fig. 6.1b).

Pore-facilitated transport is an important process. An example is the band 3 integral protein that exists as a dimer in the red cell membrane and serves as an anion channel that exchanges Cl^- and HCO_3^-. This protein probably spans the membrane 12 times. Exit of the HCO_3^- is balanced by an influx of Cl^-, which means that, in the absence of Cl^- ions, the transport of HCO_3^- stops. The band 3 protein does not simply form a hole in the membrane for the passage of ions; rather, the pore is very selective and exchanges HCO_3^- and Cl^- on a 1:1 basis. Such facilitated transport, however, is not necessary for O_2 or CO_2. These small, nonpolar molecules are allowed to move rapidly through the membrane by passive diffusion.

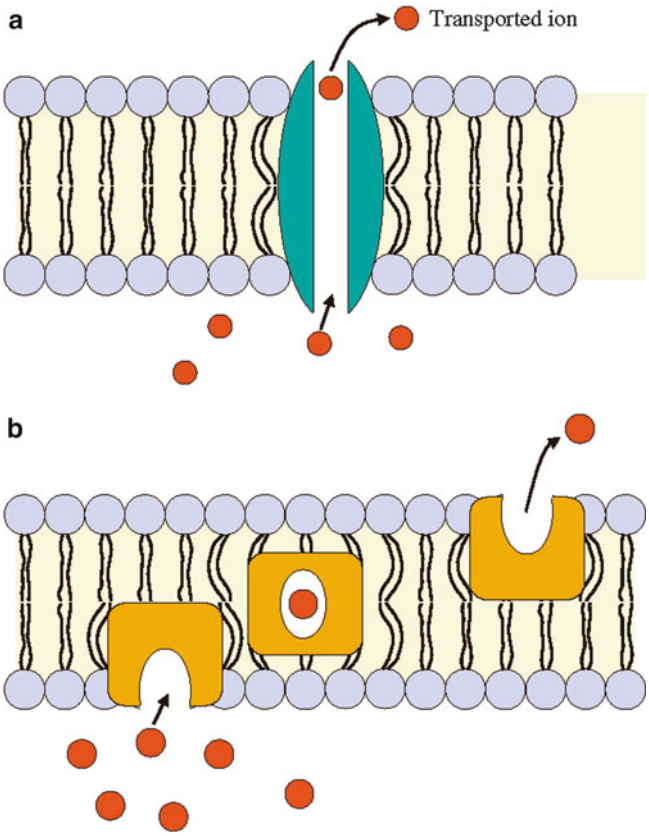


Fig. 6.1 (**a**) Facilitated transport by pores. (**b**) Facilitated transport by carrier molecules

A common example of carrier-facilitated transport is the antibiotic valinomycin (from *Streptomyces*), which is a polymer with an approximately spherical shape. Its outer layer has numerous methyl groups; thus it is hydrophobic. Inside the sphere are collections of nitrogen and oxygen that can bind (chelate) a potassium ion. This structure, however, cannot chelate other cations, so it is specific for potassium. Due to its hydrophobic exterior, valinomycin easily passes through a membrane, in contrast to the ion itself. In mathematical terms, the valinomycin increases the factor *K* in Eq. (6.6).

The measurable difference between passive and facilitated diffusion is, of course, the transport rate. Another measure is the phenomenon of saturation, which is a characteristic feature of facilitated transport. There is a finite number of carriers or pores in a membrane, and each of them

can handle only one ion at a time. Saturation will occur when all carriers are occupied (Fig. 6.2). Equations (6.4) and (6.5) imply that the rate of passive diffusion increases linearly with the concentration difference because there are no sites to saturate. The carrier-facilitated transport can be described by:

$$V_0 = \frac{V_{\max}[S]_{\text{out}}}{K_{tr} + [S]_{\text{out}}} \tag{6.7}$$

Equation (6.7) looks familiar and, with some knowledge of enzyme kinetics, it is recognized as being similar to the Michaelis–Menten equation. V_0 represents the initial rate of transport into the membrane at an external concentration of $[S]_{\text{out}}$. $V_{\max}$ is the maximum transport rate of the substrate, and K_{tr} is analogous to the Michaelis constant K_{m}. This means that K_{tr} is the substrate concentration when the transporter is half-saturated. The transport rate approaches the maximum value at a high substrate concentration. In Fig. 6.2, the straight line illustrates passive diffusion with, theoretically, no saturation.

It still has to be remembered that passive and facilitated transport are diffusion processes, and as such they do not require energy. Pores are more effective because they offer open gates. Carriers increase the solubility in membranes by offering a hydrophobic outer surface. Irrespective of the kind of diffusion transport, the final free concentration on both sides of the membrane will be equal at equilibrium.

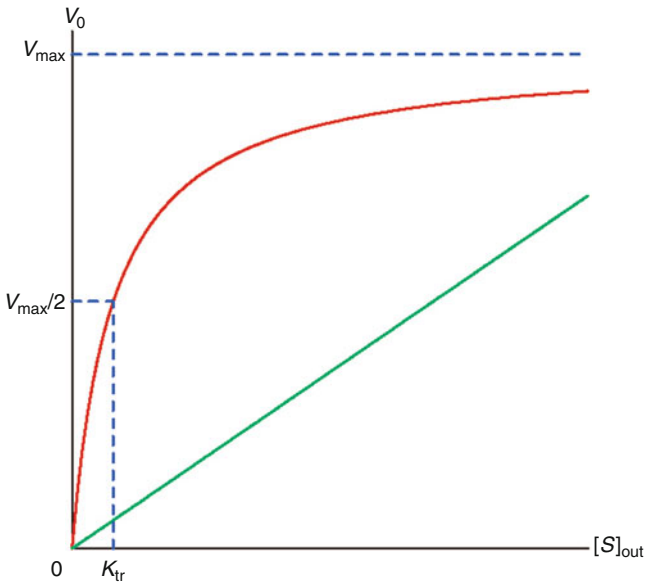


Fig. 6.2 Facilitated and passive transport

6.3.3 Three General Classes of Transport

With regard to transport across membranes, three different types have been identified, all of which depend on the number of substances and the direction in which each is transported. When only one substrate is transported, it is referred to as a *uniport transport process* – for example, the transport of glucose into red blood cells. The band 3 protein (the anion-exchange protein) is an example of an *antiport process*, in which one ion is transported out of and another into the cell. *Symport* is the transfer of two substrates in the same direction. Glucose and certain amino acids are transported via symport with Na^+. In this case, use is made of the gradient caused by the Na^+, K^+-ATPase in the plasma membrane. Figure 6.3 illustrates these general classes of transport. This characterization, however, takes into account only the direction of transport and does not show whether or not energy is required.

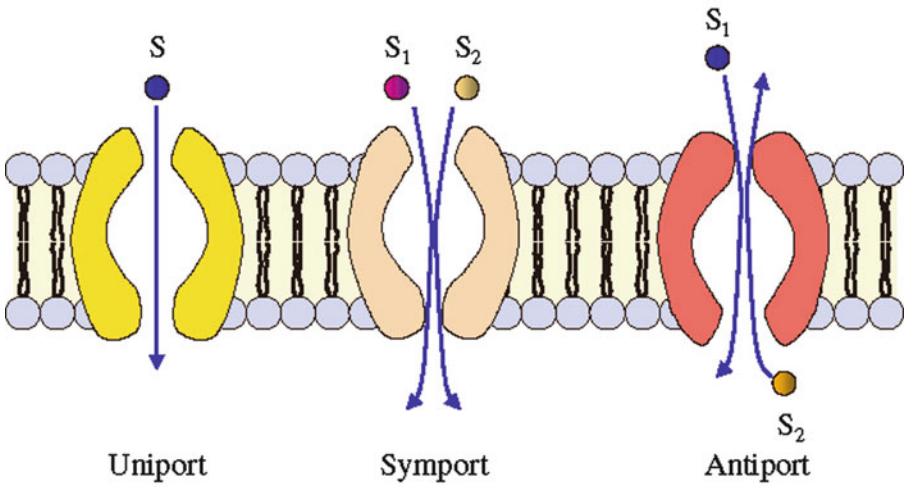


Fig. 6.3 The three general classes of transport systems differ in the number of substrates transported and the direction in which each is transported. In uniport systems, only one substrate is transported. Symport and antiport systems are characterized by the transport of two substrates in the same and in opposite directions, respectively

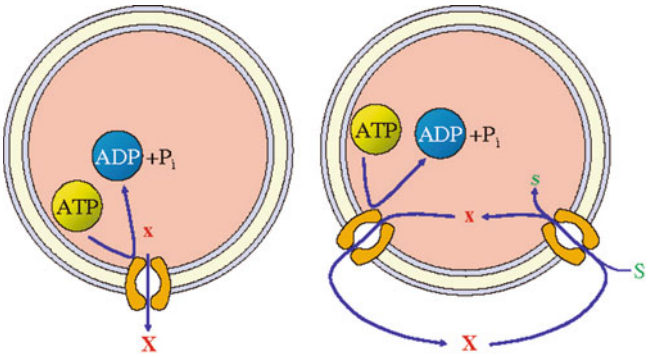


Fig. 6.4 Active transport. In primary active transport (*left*), the energy released by ATP hydrolysis drives solute transport against an electrochemical gradient. In secondary active transport (*right*), a gradient of ion X (often Na^+) has been secured by primary active transport. Transport of X down its electrochemical gradient accordingly provides the energy to drive cotransport of a second solute (*S*) against its electrochemical gradient

6.3.4 Active Transport: Transport Against a Concentration Gradient

In many situations, transport must be carried out even against concentration gradients. This requires a type of transport other than facilitated or passive diffusion. A calcium ion ratio of 30,000 must be maintained across membranes of the sarcoplasmic reticulum in muscle fibers. Using Eq. (6.1), such a ratio corresponds to $\Delta G = 25.6\ \text{kJ mol}^{-1}$, which indicates an insurmountable impediment. An active transport is necessary in such a scenario, but such a process is thermodynamically unfavorable and can only take place when coupled to a thermodynamically favorable process, such as absorption of light, an oxidation reaction, the breakdown of ATP, or an accompanying flow of some other chemical species down its electrochemical gradient. We can differentiate between *primary* and *secondary active transport* (Fig. 6.4). In the former process, accumulation is coupled directly to a thermodynamically favorable chemical reaction, such as the conversion of ATP to ADP + P_i. When uphill transport of one solute is coupled to the downhill flow of a different solute that has originally been transported uphill by primary active transport, the process is secondary active transport.

The energy required to export one Ca^{2+} ion from the inside of the cell is 9.1 kJ. The energetic cost of moving an ion depends on the electrochemical potential or the sum of the chemical and electrical gradients (see, for example, Eq. 6.2). Most cells maintain more than tenfold differences in ion concentrations across their plasma or intracellular membranes; therefore, for many cells and tissues active transport is a major energy-consuming process.

6.3.5 Types of Transport ATPases

The four known types of ATPases are P-type, V-type, F-type, and multidrug transporter. The last type will not be dealt with in this chapter because it is not involved in the transport of elements. P-type ATPases are the most versatile, at least from an elemental point of view. They all transport cations and are reversibly phosphorylated by ATP in the transport cycle. They are all integral proteins with multiple membrane-spanning regions, although they are only single polypeptides. This type of transporter is very widely distributed. The Na^+, K^+-ATPase, an antiporter for Na^+ and K^+, and the Ca^{2+}-ATPase, a uniporter for Ca^{2+}, are ubiquitous, well-understood P-type ATPases in animal tissues. They maintain disequilibrium in the ionic composition between the cytosol and the extracellular media. P-type ATPases are responsible for pumping H^+ and K^+ over the plasma membrane in parietal cells lining the mammalian stomach, thereby acidifying the contents. Bacteria use P-type ATPases to export toxic metal ions such as Cu^{2+}, Cd^{2+}, and Hg^{2+}. Table 6.2 provides a summary of the properties of the transport ATPases.

V-type ATPases act as proton pumps and are not structurally similar to the F-type ATPases. The name V-type derives from their role of keeping the pH of vacuoles of fungi and higher plants at 3–6. In addition to acidification of vacuoles, the same occurs for lysosomes, endosomes, the Golgi complex, and secretory vesicles in animal cells. The complex structures of V-type ATPases are similar throughout the family and possess an integral transmembrane domain that comprises the proton channel and a peripheral domain containing the ATP-binding site and the ATPase activity.

The F-type ATPase is so called because it has been identified as an energy-coupling factor. F-type ATPases catalyze uphill as well as downhill transport of protons. The uphill process is propelled by ATP hydrolysis, whereas the downhill reaction drives ATP synthesis. In this case, we may call them ATP synthases rather than ATPases. The F-type ATPases are multi-subunit complexes. They provide a transmembrane pore for protons and a molecular machine using the energy release by downhill proton flow to form the phosphoanhydride bonds of ATP.

6.3.6 Ion Pumps

Anyone having been in the slightest contact with biochemistry or physiology has not been able to avoid the sodium–potassium pump (Fig. 6.5). This remarkable ion pump maintains concentration gradients for sodium from

Table 6.2 Four classes of ATPases

P-type ATPases	Organism or tissue	Type of membrane
Na^+, K^+	Animal tissues	Plasma
H^+, K^+	Acid-secreting (parietal) cells of mammals	Plasma
H^+	Fungi (*Neurospora*)	Plasma
H^+	Higher plants	Plasma
Ca^{2+}	Animal tissues	Plasma
Ca^{2+}	Myocytes of animals	Sarcoplasmic reticulum (ER)
Cd^{2+}, Hg^{2+}, Cu^{2+}	Bacteria	Plasma
V-type ATPases		
H^+	Animals	Lysosomal, endosomal, secretory vesicles
H^+	Higher plants	Vacuolar
H^+	Fungi	Vacuolar
F-type ATPases		
H^+	Eukaryotes	Inner mitochondrial
H^+	Higher plants	Thylakoid
H^+	Prokaryotes	Plasma

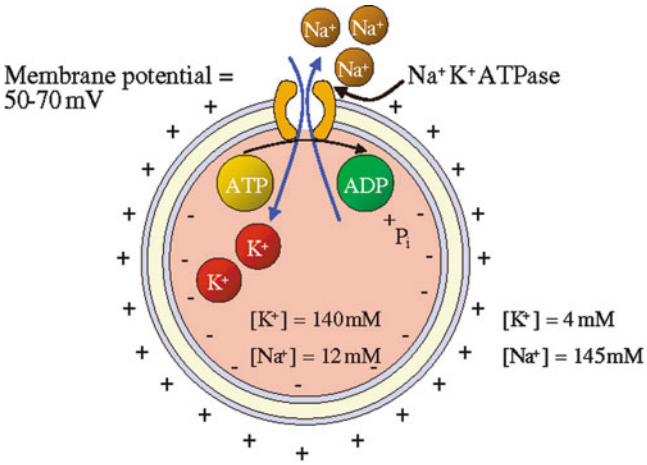


Fig. 6.5 In animal cells, this active transport system is primarily responsible for establishing and maintaining the intracellular concentrations of Na^+ and K^+ and for generating the transmembrane electrical potential

the inside of cells to the outside of 12 and 145 m*M*, respectively. At the same time, the concentration of potassium is kept at 140 m*M* intracellularly, in contrast to 4 m*M* outside. This situation pertains to almost all animal cells and would not be possible to maintain based solely on passive diffusion. The motor of this pump is the Na^+, K^+-ATPase that couples ATP breakdown to the concomitant movement of both Na^+ and K^+ against their electrochemical gradients.

The sodium–potassium pump transports two K^+ ions into the cell and exports three Na^+ ions at the cost of one molecule of ATP converted into ADP and P_i. Na^+, K^+-ATPase is an integral protein comprised of two subunits of approximate molecular weight 50,000 and 110,000, both of which span the membrane. This transporter is a P-type ATPase. The mechanism seems simple; however, the import of two potassium ions and the simultaneous export of three sodium ions are still not fully understood. It is generally assumed, though, that the ATPase cycles between two forms, one of which is phosphorylated with a high affinity for K^+ and a low affinity for Na^+, as well as one that is dephosphorylated with a high affinity for Na^+ and low affinity for K^+. The breakdown of ATP to ADP and P_i by hydrolysis takes place in two steps catalyzed by the enzyme: formation of the phosphoenzyme and hydrolysis of the phosphoenzyme with the overall net reaction:

$$ATP + H_2O \rightarrow ADP + P_i$$

In this way, energy is supplied to cover the expenditure of the pump.

Calculating the cost of exporting three moles of sodium from 12 to 145 m*M* at 37 °C, we arrive at 39.5 kJ. Correspondingly, the cost of importing two moles of potassium is 4.8 kJ. The net energy necessary to perform the transport is 44.3 kJ. Hydrolyzing one mole of ATP under physiological conditions to ADP gives 31 kJ. According to this calculation, more than 44 kJ is required for the transport, which does not seem reasonable. The trick is that in most cells the concentration of ATP is much higher than the concentrations of ADP and P_i. The energy available in real life is thus enough to pay for the transport.

Free calcium ions in cytosol are usually kept at a concentration of about 100 n*M*, which is far below what is found outside cells; thus, it is a significant finding that the total concentration of calcium in cells is much higher. One reason is that inorganic phosphates such as P_i and PP_i occur at millimolar concentrations. The concentration of free calcium ions must be kept low because inorganic phosphates readily combine with calcium, and relatively insoluble calcium phosphate precipitates will form. Maintaining the

concentration of free calcium ions requires effective pumping out of the cytosol. This is accomplished by a P-type ATPase, which is the plasma membrane Ca^{2+} pump. Another calcium ion pump of the P-type resides in the endoplasmic reticulum (ER) and moves Ca^{2+} into the ER lumen, which is separated from the cytosol. In myocytes (muscle cells), Ca^{2+} is usually sequestered in a specialized form of ER called the *sarcoplasmic reticulum*. These pumps are closely related in structure and mechanism and are collectively called *sarcoplasmic and endoplasmic reticulum calcium* (SERCA) pumps. In contrast to the plasma membrane Ca^{2+} pump, the SERCA pumps are inhibited by the tumor-promoting agent thapsigargin.

These different pumps – the plasma membrane Ca^{2+} pump and SERCA pumps – share similarities in that both are integral proteins cycling between two conformations in a mechanism not very different from that for Na^+, K^+-ATPase. The calcium-ion pump of the sarcoplasmic reticulum has been thoroughly characterized and has been identified as a prototype for Ca^{2+} pumps of the P-type. It is built from a single polypeptide (M– 100,000) spanning the membrane ten times. In a large cytosolic domain, there is a site for ATP binding, as well as an Asp residue undergoing reversible phosphorylation by ATP. This process favors a conformation with a high-affinity Ca^{2+}-binding site exposed at the cytosolic side, and the opposite process, dephosphorylation, favors a conformation with a low-affinity Ca^{2+}-binding site at the luminal side. One consequence of these cyclic changes of conformation is that the transporter binds Ca^{2+} on the side of the membrane where the calcium ion concentration is low and releases it on the side where the concentration is high. The energy released by hydrolysis of ATP to ADP and P_i during one cycle of phosphorylation and dephosphorylation drives Ca^{2+} across the membrane against the steep electrochemical gradient.

6.4 Uptake and Regulation of Iron

Iron is vital for all living organisms because it is essential for many metabolic processes, the most well-known being oxygen transport; however, further examples include DNA synthesis and electron transport. Although iron is abundant in nature, the metal is most commonly found as the virtually insoluble Fe^{3+} hydroxide $Fe(OH)_3$. Thus, iron-uptake systems require strategies to solubilize Fe^{3+}. Many organisms use siderophores (low-molecular-weight molecules secreted by bacteria, some fungi, and plants), which can solubilize Fe^{3+} for uptake by siderophore-specific transport systems. Genetic and biochemical evidence has demonstrated the presence of multiple pathways for iron uptake by eukaryotic cells. In mammals, changes in iron absorption are the major control point for altering the iron content of the body and of individual cells.

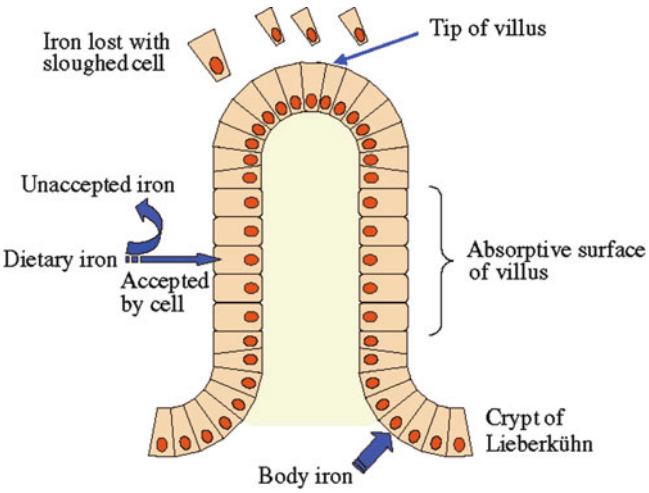


Fig. 6.6 Iron is taken up into the enterocytes of the proximal small intestine from both the diet and blood plasma. The enterocytes are born in the crypts of LieberkŸhn and move toward the villous tip to be discarded into the intestinal lumen at the end of a 2- to 3-day life span (Adapted from Conrad and Umbreit 2000)

The intestine is the major site of iron regulation with regard to controlling the uptake of dietary iron across the brush border and the release of absorbed iron across the basolateral membrane to the circulation. Cells responsive to iron uptake are born in the crypt of Lieberkühn, located in the duodenum and jejunum. These cells differentiate and move toward the absorptive surface of the villus, where they are referred to as *enterocytes*. Gradually, mature enterocytes move toward the tip of the villus and are sloughed into the intestinal lumen (Fig. 6.6).

In the intestinal lumen, iron exists in the forms of ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron salts. Because ferric iron becomes insoluble at pH values above 3, ferric ions must be reduced or chelated by amino acids or sugars to be efficiently absorbed. Most ferrous iron remains soluble even at pH 7, so absorption of ferrous iron salts is more efficient than absorption of ferric iron salts; however, most dietary inorganic iron is in the form of ferric iron. In most industrialized countries, two-thirds of the iron in the diet is present as ferric iron and one-third as heme iron (Carpenter and Mahorey 1992). Reduction of ferric irons becomes necessary for efficient dietary iron absorption and is mediated by a mucosal ferrireductase that is present in the intestines. Inhibition of ferrireductase activity in intestinal cells reduces iron absorption, which demonstrates the importance of ferric iron reduction in dietary iron import. Alternatively, uptake of ferric irons might be mediated by the paraferritin pathway, though less efficiently. In addition to ferrireductase activity, the presence of dietary ascorbate provides a reduction of ferric iron to ferrous, whereby absorption is enhanced. Figure 6.7 illustrates the intestinal absorption and balance of iron.

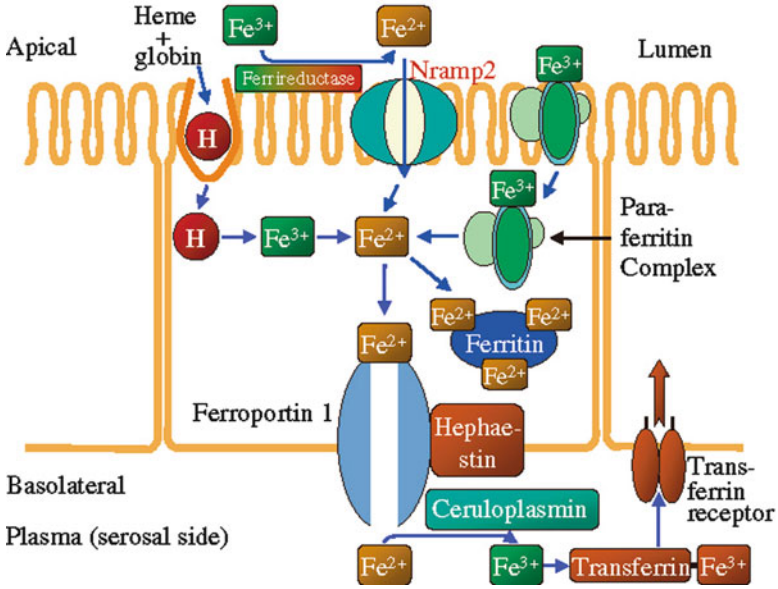


Fig. 6.7 Pathways of heme and non-heme uptake of iron and transport in the intestine. The majority of the dietary ferric iron is reduced to ferrous iron or solubilized by mucin, ascorbic acid, or other reducing agents. The ferric iron in the lumen of the intestine is reduced by ferrireductase or in the cytoplasm by monooxygenase. An iron transporter, Nramp2, situated at the apical cell surface, transports most of the ferrous iron into the enterocyte. Yet another pathway for ferric and ferrous iron into the cell is the paraferritin complex, comprised of β integrin, mobilferrin, and flavin monooxygenase. Heme is taken up in the enterocytes as an intact metalloporphyrin (H). The uptake process is probably mediated by endocytosis. In the cytoplasm, heme is degraded by heme oxidase to release its inorganic iron. Two possibilities exist for the intracellular iron; either it is stored in ferritin or it is transported over the basolateral membrane by ferroportin 1 into plasma. Iron export seems to be facilitated by hephaestin concomitant with ferroportin 1. Ferroportin 1 could also load iron onto transferrin assisted by the plasma ferroxidase, ceruloplasmin (Adapted from Lieu et al. 2001)

6.4.1 Non-Heme Iron Uptake

Absorption of both heme and non-heme iron occurs predominantly in the crypt cells of the duodenum and jejunum (Wood and Han 1998). Enterocytes, the specialized cells located on the intestinal villus, control the passage of dietary iron in the lumen of the intestine and the transfer of iron into the circulation of the body. To enter the circulation, dietary iron has to cross three cellular barriers: iron absorption over the apical membrane, intracellular iron transport through the cell, and iron export over the basolateral membrane and into the circulation. The enterocytes, however, have no transferrin receptors on the surface exposed to the lumen (Pietrangelo et al. 1992); thus enterocytes differ from other nuclear-bearing cells. Consequently, there has to be a mechanism of absorption other than the usual transferrin–transferrin receptor pathway. Absorption of iron across the apical membrane of the enterocytes is mediated by a divalent cation transporter, which is called *Nramp2* or *divalent cation transporter 1* (DCT1) (Fleming et al. 1997), or more recently *divalent metal transporter 1* (DMT1). Nramp is an acronym for *natural resistance-associated macrophage protein*; Nramp2 is highly homologous to Nramp1, a molecule that is important in host defense against pathogen infection. Evidence speaks in favor of Nramp2 being responsible for iron transport from the duodenum lumen into the cytoplasm of enterocytes.

6.4.2 The Iron Importer: DMT1 (Nramp 2)

The gene coding for MDT (Nramp2) in humans contains more than 36,000 bases. At least two different forms of mRNA are coded (Lee et al. 1998), as this is the first step in the expression of the gene. One region of the DMT1 (Nramp2) isoform I is the 3′ untranslated region, which contains an iron-responsive element similar to the iron-responsive element present in the 3′ untranslated region of the mRNA transferrin receptor 1 (Tandy et al. 2000). The isoform II, however, lacks the iron-responsive element. Expression of DMT1 (Nramp2) isoform I is, therefore, upregulated in iron-deficient animals and human intestinal cells, whereas expression of DMT1 (Nramp2) isoform II is not due to the lack of the iron-responsive element (Fleming et al. 1999). The importance of the presence of a functional iron-responsive element in the DMT1 (Nramp2) isoform I is that its expression probably is controlled by intracellular iron concentration. Such a control should not be functional in the case of DMT1 (Nramp2) isoform II.

As might be expected, DMT1 (Nramp2) is highly expressed at the duodenum brush border, which corroborates its important role in intestinal iron absorption (Cannone-Hergeaux et al. 1999), because enterocytes lack the transferrin receptor system at the absorptive surface. DMT1 (Nramp2) is located on the plasma membrane as well as on

subcellular vesicular compartments. The DMT1 (Nramp2) protein is thought to consist of 12 transmembrane domains, and studies show that DMT1 (Nramp2) acts as a proton-coupled divalent cation transporter (Gunshin et al. 1997). DMT1 (Nramp2) is, therefore, capable of transporting not only ferrous iron but also a number of divalent cations such as Zn^{2+}, Mn^{2+}, Co^{2+}, Cd^{2+}, Ni^{2+}, and Pb^{2+} (Gunshin et al. 1997). In addition, DMT1 (Nramp2) function is pH dependent, being optimal at a pH of <6 (Gunshin et al. 1997). DMT1 (Nramp2) has been shown to be colocalized at the subcellular level with transferrin, and DMT1 (Nramp2) might be involved in transporting transferrin-bound iron across the membrane of endosomes into the cytoplasm (Gruenheid et al. 1999).

Developments in molecular genetics have recently provided interesting information about a possible link between variability in the *DMT1* gene and resistance to recombinant human erythropoietin therapy in chronic kidney disease patients. This finding may explain why a comparison of responder patients and non-responders showed no significant differences for iron, transferrin saturation and ferritin values. However, non-responders showed significantly lower hemoglobin concentration and mean cell hemoglobin (Costa et al. 2008).

6.4.3 Heme Iron Uptake

Hemoglobin iron from food is absorbed more efficiently than inorganic iron; therefore, absorption of iron from myoglobin and hemoglobin is different from the way in which inorganic iron is absorbed. Hemoglobin is enzymatically digested in the intestinal lumen, and the heme molecule is internalized by the enterocytes as an intact metalloporphyrin (Majuri and Grasbeck 1987). It might be that the heme molecule enters the cell through a receptor-mediated process. Once inside the enterocyte, heme is metabolized by heme oxygenase, and inorganic iron is released. This is either stored as ferritin or transported across the basolateral membrane to enter the bloodstream (Fig. 6.7). When the enterocyte ends its life cycle, iron in the form of ferritin will be sloughed with the aged cells and leave the body through the gastrointestinal tract. Humans have a limited means of eliminating iron; therefore, this process is an important mechanism of iron loss (Lieu et al. 2001).

6.4.4 Paraferritin-Mediated Iron Uptake

Nramp2 is a much better transport agent for ferrous iron than ferric iron. In addition, ferrous and ferric iron can also be internalized by enterocytes in different pathways. Paraferritin is a membrane complex with a molecular weight of 520 kDa that contains β-integrin, mobilferrin (a homolog of calreticulin, which is a lectin-like chaperone promoting efficient folding of proteins in the ER), and flavin monooxygenase. It participates in the mucin-mediated iron uptake in the intestinal lumen (Fig. 6.7) (Umbreit et al. 1998). Experiments with erythroleukemia cells show that an anti-β_2-integrin monoclonal antibody blocks 90% of ferric citrate uptake. Little effect, however, was observed on the uptake of ferrous iron. Consequently, it seems that ferric iron is absorbed via the paraferritin-mediated pathway (Conrad et al. 1999). A possible mechanism is that ferric iron is solubilized by mucin in the intestinal lumen, transferred to the mobilferrin- and β-integrin-containing paraferritin complexes, and then internalized (Conrad et al. 1999). Having been internalized, flavin monooxygenase is associated with the complexes and ferric iron is reduced to ferrous iron in parallel with the activity of NADPH. The β-integrin- and mobilferrin-containing paraferritin complex also interacts with β_2 microglobulin. In addition, mobilferrin and β_2 microglobulin have been shown to play critical roles in the development of iron overload in hemochromatosis in animals (Rothenberg and Voland 1996).

6.4.5 The Iron Exporter: Ferroportin1 and Hepcidin

A novel iron transporter gene, ferroportin1, has recently been identified (Donovan et al. 2000). Sequence analysis of ferroportin1 shows that it has a stem-loop structure, typical of iron-responsive elements, in the 5′ untranslated region (Donovan et al. 2000). It has been shown that the iron-responsive element binds to iron regulatory proteins 1 and 2. This indicates that expression of ferroportin1 is regulated by intracellular iron levels (McKie et al. 2000). Studies of ferroportin1 demonstrate that ferroportin1 mediates iron efflux across membranes by a mechanism requiring an auxiliary ferroxidase activity (Donovan et al. 2000; McKie et al. 2000). Ferroportin1 is expressed highly in the placenta, liver, spleen, macrophages, and kidneys. In the cell, ferroportin1 is located on the basolateral membrane of duodenal enterocytes (McKie et al. 2000). This suggests that ferroportin1 probably functions as an iron exporter in the enterocytes (Fig. 6.7). Ferroportin1 is located on the basal surface of placental syncytiotrophoblasts, which probably suggests a role for ferroportin1 in iron transport into the embryonic circulation (Donovan et al. 2000). The function of ferroportin1 is believed to be in parallel with the membrane-resident ferroxidase hephaestin and serum ceruloplasmin (McKie et al. 2000). Hephaestin has a high degree of similarity to ceruloplasmin, which is a multi-copper oxidase possessing ferroxidase activity, which is required for the release of iron into blood and the binding to transferrin.

Hephaestin does not transport iron, which is similar to ceruloplasmin; however, it facilitates the transport of iron from enterocytes into the body's circulation (Harris et al. 1998). Sex-linked anemic mice with defective hephaestin show normal dietary iron absorption into the enterocytes, but they suffer from a defect in the transport of iron from duodenum to the blood (Vulpe et al. 1999). Unfortunately, the mechanism by which ferroportin1 mediates the transport of iron across the basolateral membrane and by which it interacts with hephaestin and ceruloplasmin is still unknown (Lieu et al. 2001).

Hepcidin is a 25-amino acid peptid hormone synthesized in hepatocytes, which are the predominant producers. It inhibits iron entry into the plasma compartment from the three main sources of iron. The production of hepcidin is regulated by iron and so that more hepcidin is produced when iron is abundant thereby limiting iron absorption and release from stores (Ganz and Nemeth 2008, 2012). The only known receptor for hepcidin is ferroportin. Hepcidin is, as is iron, regulated by the erythropoietic requirement for iron. During active erythropoiesis hepcidin production is suppressed and more iron will be available for hemoglobin synthesis. The interaction between hepcidin and ferroportin is attracting increasing interest and the system is also considered as a therapeutic target in anemias and iron overload disorders (Ganz and Nemeth 2011).

6.4.6 Regulation of Dietary Iron Absorption

The regulation of iron absorption by enterocytes is exercised in various ways. In the first place, it may be modulated by the amount of iron in recently consumed food. This mechanism is referred to as the *dietary regulator* (Andrews 1999). Enterocytes are resistant to acquiring additional iron for several days after consumption, a phenomenon referred to as *mucosal block* (Andrews 1999). A second regulatory mechanism, which monitors the iron levels stored in the body rather than the dietary iron status, is referred to as the *stores regulator*. When there is iron deficiency, the stores regulator can modify the amount of iron uptake by a factor of approximately two to three (Finch 1994). Saturation of plasma transferrin with iron is also thought to influence the dietary absorption of iron, at least indirectly. A detailed mechanism of the stores regulator remains to be defined. The *erythropoietic regulator* is a third regulatory mechanism that has a greater capacity to increase iron absorption than the stores regulator. The erythropoietic regulator does not respond to the cellular iron levels; however, it does modulate iron absorption in response to the requirements for erythropoiesis. Further studies are required to increase our knowledge of the molecular mechanisms of intestinal iron absorption (Lieu et al. 2001).

6.4.7 Transferrin Receptor-Mediated Iron Uptake

In the blood, iron is transported by the plasma glycoprotein transferrin, which has a molecular weight of about 80 kDa and a high affinity for ferric iron. Most cells in the body, except enterocytes in the intestine, get iron from transferrin. The uptake of iron in cells begins with the binding of transferrin to a receptor on the cell surface known as the *transferrin receptor*. It binds transferrin only when it carries iron. The maximum capacity of transferrin is two iron ions. The transferrin–receptor complex is then internalized through the endocytic pathway. Transferrin receptors do not interact directly with iron, yet they control iron uptake and storage by most cells in the organism. There are at least two types of transferrin receptors. Transferrin receptor 1 is a membrane-resident glycoprotein that is expressed in all cells, with the exception of mature erythrocytes. The other type, transferrin receptor 2, is a homolog of transferrin receptor 1. It is specifically expressed in the liver, particularly in the hepatocytes. Following internalization, the endosome is covered with clathrin. This is a protein complex of three large and three small polypeptide chains that is thought to help bend the membrane in the internalization process. The endosome is then uncoated by an uncoating ATPase. Then protons are pumped into the endosome, causing iron to be released from the transferrin. Iron then passes through the endosomal membrane and enters the intracellular labile pool. Intracellular iron in the labile pool can then be utilized for the synthesis of heme- and non-heme-containing proteins or stored within ferritin with a molecular weight of 474 kDa. The storage capacity of ferritin is as high as 4,500 iron atoms. The receptor-bound transferrin is recycled back to the cell surface for reuse after completing a cycle of highly specific and efficient cellular iron uptake (Fig. 6.8).

6.4.8 The Iron-Binding and Iron-Transport Protein: Transferrin

Transferrin is a single polypeptide chain of a glycoprotein that consists of two globular domains. Both domains offer a high-affinity binding site for one iron ion. Iron affinity of transferrin is pH dependent, and iron is released from transferrin when the pH is below 6.5. Transferrin might also be involved in the transport of a number of metals, such as aluminum, manganese, copper, and cadmium (Davidsson et al. 1989; Moos et al. 2000), but iron has the highest affinity to transferrin and will drive other metals out.

The liver is the primary site of synthesis of transferrin (Takeda et al. 1998); however, it is synthesized in significant amounts in the brain, testis, lactating mammary gland, and

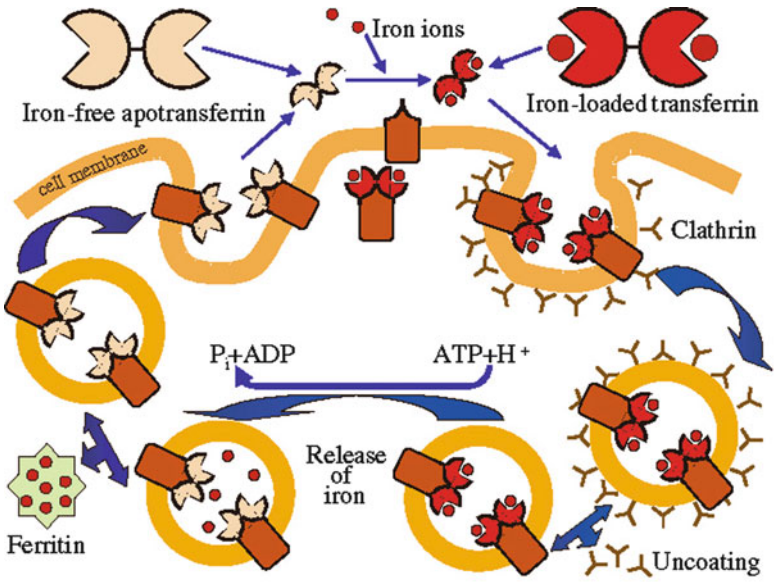


Fig. 6.8 A schematic view of the extracellular binding of iron to apotransferrin, receptor-mediated endocytosis, release of iron into the endosome (driven by ATP), and the loading of iron into ferritin

some fetal tissues during development. The three different forms of transferrin, existing as a mixture, are iron-free (apotransferrin), one-iron (monoferric), and two-iron (transferrin diferric). The ratio between these forms depends on the concentration of iron and transferrin in blood plasma. Under normal conditions, most of the iron molecules in blood plasma are bound to transferrin (Lieu et al. 2001). The main function of transferrin is to capture iron from plasma and to transport it to various cells and tissues in the organism.

6.4.8.1 The Transferrin-Binding and Transferrin-Transport Protein: Transferrin Receptor 1

Transferrin receptor 1 is a dimer comprised of two identical subunits and having a molecular weight of approximately 90 kDa. The receptor crosses the plasma membrane. The monomers are joined by two disulfide bonds and consist of three domains: a 61-residue amino-terminal domain, a 28-residue transmembrane region that helps to anchor the receptor into the membrane, and a large extracellular carboxyl terminus of 671 amino acid residues (McClelland et al. 1984). As a type II membrane protein, the carboxyl terminal ectodomain of the transferrin receptor 1 is critical for transferrin binding. Indeed, replacement of the carboxyl-terminal, 192-amino-acid residues of the human transferrin receptor 1 with the corresponding region of the chicken transferrin receptor dramatically reduces or completely abolishes its binding affinity for transferrin. Because each ectodomain contains a binding site for the transferrin molecule, a homodimer of transferrin receptor 1 can bind up to two molecules of transferrin simultaneously.

Transferrin receptor 1 is synthesized intracellularly in the ER. Additionally, it undergoes a number of posttranslational modifications. Its ectodomain is comprised of three nitrogen-linked glycosylation sites and one oxygen-linked glycosylation site (Kohgo et al. 2002). Correct folding is strongly dependent on the nitrogen-linked glycosylation sites of transferrin receptor 1. If the oxygen-linked glycosylation at threonine 104 is eliminated, the cleavage of transferrin is enhanced, which, in turn, promotes the release of its ectodomain (Rutledge and Enns 1996). The segment of transferrin receptor 1 that crosses the membrane consists of 18 hydrophobic amino acids and also undergoes posttranslational modifications. The hydrophobic membrane-crossing segment is covalently bound to fatty acids and is subjected to acylation with palmitate. This probably helps to fasten the receptor to the plasma membrane (Kohgo et al. 2002).

The part of transferrin receptor 1 that is resident in the cytoplasm is important for the clustering of the receptor into the chlatrin-coated pits of the plasma membrane and, subsequently, for endocytosis (Iacopetta et al. 1988). A conserved internalization signal (YTRF; tyrosine, threonine, arginine, phenylalanine) within the 61-amino-acid residues in the cytoplasmic part of transferrin receptor 1 is critical for efficient endocytosis of the receptor (Collawn et al. 1993).

Human transferrin receptor 1 is a tightly associated homodimer. Each transferrin receptor 1 monomer consists of three distinct globular domains (Rolfs and Hediger 1999). The general form of the homodimer suggests that transferrin could bind to either side with no contact between the two transferrin molecules. The extracellular part of transferrin receptor 1 is separated from the membrane by a stalk, which presumably includes residues involved in disulfide bond formation and oxygen-linked glycosylation (Rolfs and Hediger 1999; Lieu et al. 2001).

Transferrin receptor 1 is the general mechanism for cellular uptake of iron from plasma transferrin. The current model of iron uptake from transferrin via receptor-mediated

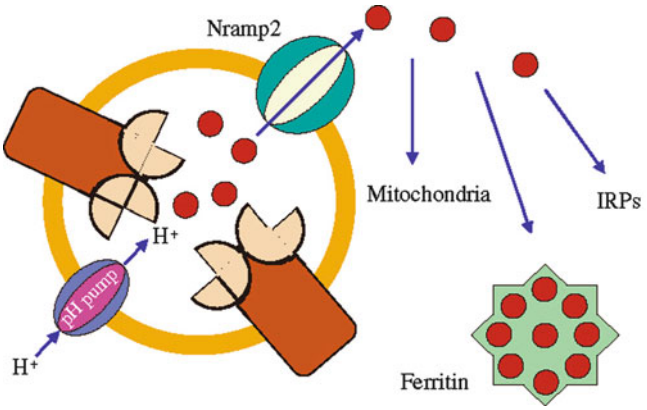


Fig. 6.9 Release of iron from transferrin in the endosome and following events. A proton pump (energy-requiring) decreases the pH within the endosome with the consequence that iron is released from transferrin. The iron transporter Nramp2 carries out the subsequent transport of iron over the endosome membrane into the cytoplasm. *IRP* stands for iron-responsive protein

endocytosis in mammals is shown in Fig. 6.8 and in greater detail in Fig. 6.9. The first step, the binding of transferrin to transferrin receptor 1, is accomplished by a physical interaction that does not require an increase in temperature or energy (Conrad and Umbreit 2000). Transferrin can bind one or two ferric ions, and the iron status of transferrin affects its affinity for its receptor. Diferric transferrin has the highest affinity, followed by monoferric transferrin; apotransferrin (without iron) has the lowest affinity. An estimated dissociation constant for diferric transferrin is about 2–7 n*M* (Lieu et al. 2001). The plasma concentration of diferric transferrin is about 5 μ*M* under physiological conditions; consequently, most surface transferrin receptors become saturated with transferrin. Thus, the homodimeric transferrin receptor can mediate a maximum uptake of four atoms of iron at the same time.

The complexes consisting of transferrin receptor–-transferrin–iron interact with adaptor proteins in the clathrin-coated pit and are then internalized by the cells through an endocytic pathway mediated by the receptor. The tyrosine internalization motifs located on the parts of the transferrin receptors that reside in the cytoplasm seem to be necessary for a high-affinity binding to the adaptor protein complexes on the plasma membrane. Importantly, but not surprisingly, this process is temperature and energy dependent (Lieu et al. 2001). Inside the endosome, an ATPase proton pump causes acidification of the endosome and results in the release of iron from transferrin. The apotransferrins remain attached to the transferrin receptors and return to the cell surface, where they are released from the cells. The binding between transferrin and the transferrin receptor is dependent on the pH, which is critical to membrane uptake and the release of transferrin. The release of apotransferrin from its receptor occurs at neutral pH at the cell surface. Both the ligand and receptor, in this way, become available for recycling the absorption of iron. After its release from the transferrin, iron passes through the endosomal membrane into the cytoplasm via the iron transporter Nramp2 (also known as the DCT1). Iron that enters the cell can be utilized in the synthesis of heme or incorporated in iron-containing molecules. Intracellular iron can also be stored in the ferritin complexes or can modulate the activity of iron regulatory proteins (Lieu et al. 2001).

6.4.8.2 Second Transferrin-Binding and Transferrin-Transport Protein: Transferrin Receptor 2

Transferrin receptor 2, which is a homolog of transferrin receptor 1, was recently identified. Its gene is located on chromosome 7q22 and gives rise to two transcripts approximately 2,900 and 2,500 bases in length. Like transferrin receptor 1, transferrin receptor 2 is a type II transmembrane glycoprotein, and it shares 66% similarity in its ectodomain with transferrin receptor 1. Although the cytoplasmic portion of transferrin receptor 2 is very different from transferrin receptor 1, transferrin receptor 2 also contains an internalization motif (YQRV; tyrosine, glutamine, arginine, valine), which is similar to the YTRF motif in transferrin receptor 1. Transferrin receptor 2 does not possess iron-responsive elements. It seems that expression of transferrin receptor 2 is not regulated by an iron-regulatory, protein-mediated feedback regulatory mechanism in response to cellular iron status (Kawabata et al. 1999).

In contrast to transferrin receptor 1, transferrin receptor 2 is primarily expressed in the liver. Like transferrin receptor 1, binding of transferrin receptor 2 to transferrin is also pH dependent. The binding of apotransferrin to transferrin receptors 1 and 2 only takes place at acidic pH. Expression levels of both transferrin receptors 1 and 2 correlate with stages of the cell cycle which, in turn, are related to requirements for iron during DNA synthesis. Nevertheless, transferrin receptor 2 differs from transferrin receptor 1 in its binding properties with transferrin and regulation of expression. Holotransferrin has a lower affinity for transferrin receptor 2 than for transferrin receptor 1.

Transferrin receptors 1 and 2 are likely not only to be regulated through distinct pathways but also to mediate iron uptake and storage by a different, yet unidentified, mechanism. Transferrin receptor 1 seems to play a general role in cellular iron uptake; however, transferrin receptor 2 appears to play a specific role in iron uptake and storage in the liver, due to its high expression in hepatocytes (Lieu et al. 2001).

6.4.9 Control of Iron Metabolism

Not surprisingly, animal cells differ somewhat from plants and lower eukaryotes with regard to the way in which they control iron metabolism. Transcription is the preferred method by which plants and lower eukaryotes maintain iron homeostasis. In the yeast *Saccharomyces cerevisiae*, for example, the iron-regulated transcription factor AFT1 controls production of multiple gene products that are needed to make up the high-affinity iron transport systems. Similarly, in plants, the iron storage protein ferritin is transcriptionally regulated by iron, which differs from the way in which it is regulated in animal cells, as animal cells utilize posttranscriptional control of iron metabolism in most cell types in the body. Apparently, regulation of gene transcription has a more important role in cell-type-specific modulation of iron homeostasis. Tissue-specific regulation of the expression of H- and L-ferritin, an erythroid-specific isoform of 5-aminolevulinate synthase, is one example, as well as control of the relative expression of iron regulatory protein 1 (IRP1) and IRP2 between tissues. This means that mammalian iron homeostasis is maintained through integrated use of sensory and regulatory systems operating at multiple levels of gene regulation (Eisenstein 2000).

Although iron is an essential trace element, it might be detrimental if it is available as a free ion; consequently, besides providing storage, it is also necessary to prevent toxicity. Alteration of ferritin gene transcription provides an important means by which the relative abundance of the ferritin subunits can be modified to meet the unique iron storage and/or detoxification needs of specific tissues. The ratio of the abundance of the heavy-chain (H) and light-chain (L) subunits varies among tissues, and this variation is probably due to tissue-specific differences in the rates by which the ferritin genes are transcribed (Tsuji et al. 1999; Eisenstein 2000). Gene transcription of ferritin can be modulated by both iron-dependent and iron-independent factors. Experiments show that an excess of iron can cause a selective increase in L-ferritin gene transcription in the liver. In other systems, though, H- and L-ferritin transcription is altered in parallel to iron. Transcription of the ferritin genes is also modulated by a number of iron-independent signaling pathways (Eisenstein 2000). Synthesis of transferrin and transferrin receptor takes place in fewer tissues than is the case for the ferritins. Transcription thus seems to dictate their expression in specific tissues. Iron deficiency causes induction of the transcription of both genes (Testa et al. 1989). In many cases, however, transferrin receptor expression in iron deficiency is controlled by the regulation of mRNA stability. Seemingly contradictory is the fact that transcription appears to be a more critical factor in the iron regulation of transferrin expression in a deficiency, although translational regulation may be important in reducing transferrin synthesis when there is excess iron (Eisenstein 2000). The gene transcription of transferrin and transferrin receptor is enhanced during hypoxia to increase iron delivery to the erythron, which is necessary to advance erythropoiesis and increase oxygen-carrying capacity (Tacchini et al. 1999; Eisenstein 2000). During erythropoiesis, transcription plays a greater role in modulating transferrin receptor expression than in many non-erythroid cells. This makes increases in transferrin receptor expression possible without maximal induction of IRP activity and allows for simultaneous expression of the erythroid isoform of delta-aminolevulinate synthase (ALAS, eALAS). The growth state of cells also influences transferrin receptor gene transcription. Clearly, regulation of the transcription of the H- and L-ferritins, transferrin, and transferrin receptor genes contributes greatly to the maintenance of cell and organ iron homeostasis (Eisenstein 2000).

6.4.10 Iron Regulatory Proteins and the Coordination of Iron Homeostasis

Iron regulatory proteins play a significant role in maintaining iron homeostasis by coordinating many of the mRNAs that encode proteins for the control of uptake or the metabolic fate of iron. Furthermore, changes in rates of transcription help establish the final level of these mRNAs. IRPs are considered central regulators of mammalian iron metabolism because they regulate the synthesis of proteins that is required for the uptake, storage, and use of iron by cells (Eisenstein 2000). Important factors with regard to this regulation include the following:

- It is well accepted that IRPs are critical factors of the posttranscriptional regulation of transferrin receptor expression.
- IRPs play a major role in determining the iron storage capacity of cells by regulating translation of both H- and L-ferritin mRNA.
- Translation of the mRNA for the eALAS also seems to be regulated by IRPs.

In this role, IRPs may coordinate the formation of protoporphyrin IX with the availability of iron. IRPs may thus be important modulators of iron cycling in the body. In addition, the mRNA-encoding divalent metal transporter 1 (DMT1) and ferroportin1/iron-regulated gene 1 (IREG1) contain iron-responsive element (IRE)-like sequences. This suggests that IRPs might possibly affect the use of these mRNAs. DMT1 expression is iron regulated in some but not all situations; however, there are indications that the abundance of ferroportin1/IREG1 mRNA responds to changes in iron status. If the IREs in DMT1 and ferroportin1/IREG1 mRNA are functional, as is the case for transferrin receptor mRNA,

then IRPs probably are major modulators of the transmembrane transport of transferrin and non-transferrin iron (Eisenstein 2000).

Upon changes in the cell concentration of iron, the concentrations of the proteins ferritin, ferroportin, DMT1 and transferrin 1 also change. These changes are mediated by a second iron response mechanism targeted to mRNA. Mature mRNA encoding the IRPs are stored in the cytoplasm in a repressed or inactive form until the iron level changes (Theil 2011). The repressor proteins IRP1 and IRP2 bind specifically to noncoding RNA structures in DMT1, transferrin 1, ferritin and ferroportin 1, which all are called IREs. IRE-RNA/IRP binding has two different effects depending on the structure and location of the IRE-RNA, before or after ribosome binding. The iron regulatory structures are protected by IRP binding to mRNA, which encodes uptake proteins DMT1 and transferrin 1. Degradation of mRNA is prevented through this mechanism as well. IRP binding thus stabilizes the mRNA concentration and increases mRNA translation and protein synthesis. The binding of IRE-RNA/IRP to ferritin and ferroportin mRNA inhibits ribosome binding and protein synthesis and translation (Theil 2011).

6.5 Uptake and Regulation of Zinc

A number of physiologic systems contribute to zinc homeostasis under different conditions. Central to the maintenance of zinc homeostasis, however, is the gastrointestinal systems, especially the small intestine, liver, and pancreas. Specifically, the processes of absorption of exogenous zinc and gastrointestinal secretion and excretion of endogenous zinc are critical to zinc homeostasis throughout the body. During evolution, cells developed efficient uptake systems to allow for the accumulation of zinc even when it is scarce. These uptake systems use integral membrane transport proteins to move zinc across the lipid bilayers of the plasma membrane. Once inside a eukaryotic cell, a portion of the zinc must be transported into intracellular organelles to serve as a cofactor for various zinc-dependent enzymes and processes present in those compartments; therefore, transporter proteins must be present in organelle membranes to facilitate this flux of zinc. Zinc can also be stored in certain intracellular compartments when supplies are high and used later if zinc deficiency ensues. Again, zinc transporters are required to facilitate this transport in and out of organelles.

6.5.1 Families of Zinc Transporters in Eukaryotes

Many types of transporters have been found to be involved in zinc transport. In prokaryotes, transporters of the ATPase binding cassette (ABC) family have been demonstrated to work in zinc uptake. The zinc ABC proteins of *Escherichia coli*, for example, are a major source of zinc incorporation for these cells. A family of P-type ATPases functions as zinc efflux transporters; the ZntA protein in *E. coli* is one such transporter. Interestingly, this protein is important for zinc detoxification by pumping the metal ion out of the cell when intracellular zinc levels get too high (Hantke 2001).

Eukaryotes have been found not to use ABC transporters or P-type ATPases; instead, zinc transport apparently is accomplished by two other families of transporters. The uptake of zinc and transport from the extracellular space to the cytoplasm has been found to be associated with the ZIP (Zrt-, Irt-like proteins) family. Additionally, the mobilization and transport of stored zinc from an organelle to the cytoplasm have been shown to be carried out by ZIP transporters. The CDF (cation diffusion facilitator) family does the opposite of the ZIP proteins; namely, it pumps zinc from the cytoplasm out of the cell or into the lumen of an organelle. All of the known members of these families play roles in metal ion transport, and zinc is often the substrate; consequently, it might be that several other members work in zinc transport (Gaither and Eide 2001).

Two of the first members of the ZIP family to be discovered were Zrt1 of *Saccharomyces cerevisiae* and Irt1 of *Arabidopsis thaliana* (Gaither and Eide 2001); thus the name ZIP transporters. Zrt1 is a zinc uptake transporter in yeast, and Irt1 is an iron transporter in plants. Currently, 86 ZIP members can be found in the protein sequence database at the National Center for Biotechnology Information (NCBI). This list includes proteins from eubacteria, archaea, fungi, protozoa, insects, plants, and mammals.

The degree of sequence conservation can be used to split the ZIP family into subfamilies. Most proteins in the ZIP family are predicted to have eight membrane-crossing domains; however, some may have as few as five. The majority of ZIP proteins share a similar predicted topology where the amino and carboxyl termini are located in the extracellular space (Fig. 6.10a). Parts of this topology have been corroborated for some members of the family. Examples include the amino terminus of Zrt1 and the carboxyl terminus of hZip2, which have been shown to be on the outside surface of the plasma membrane. Many of the ZIP proteins have a long loop region located between membrane-crossing domains III and IV. This region is called the "variable region," because both its length and sequence show little conservation among the family members. Many of the ZIP proteins are characterized by the presence of many histidine residues. In Zrt1, this sequence ishistidine–aspartate–histidine–threonine–histidine–aspartate–glutamate…, and in Irt1 the sequence is histidine–glutamate–histidine–glutamate–histidine–glutamate–histidine….The function of this region is not known; however, it is acknowledged to be a potential metal

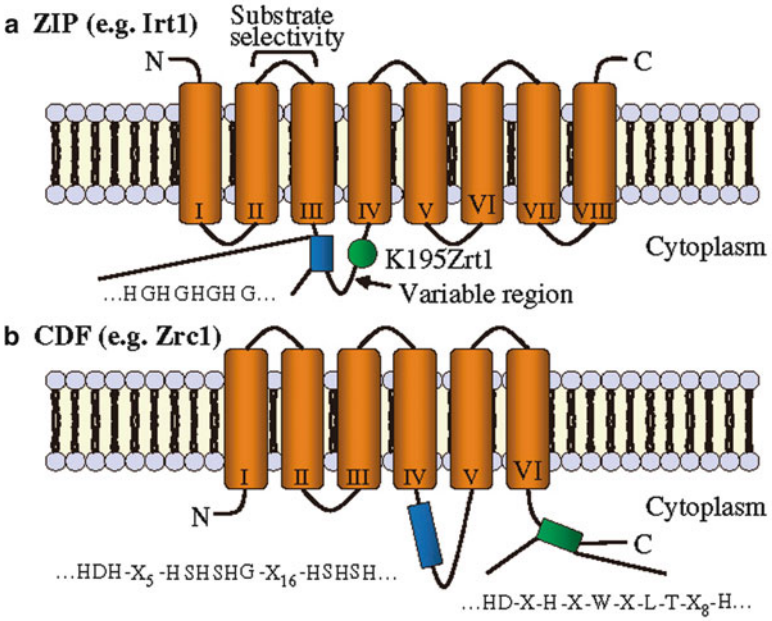


Fig. 6.10 A sketch of the predicted membrane topology of ZIP and CDF proteins. (**a**) ZIP proteins as well as Irt1 are predicted to have eight membrane-crossing domains (I–VIII). Indicated as the variable region are the conserved and functionally important residues in domains VI and V, the ubiquitinated K195 in Zrt1, and the extracellular loop region affecting Irt1 substrate specificity. (**b**) The majority of CDF transporters as well as Zrc1 are predicted to have six membrane-crossing domains (I–VI). Conserved polar or charged residues within the membrane-crossing domains I, II, and V are indicated. *H* histidine, *G* glycine, *D* aspartate, *S* serine, *L* lysine, *T* threonine, *W* tryptophan, and *X* any amino acid (Adapted from Gaither and Eide 2001)

binding domain. Consequently, its conservation in many of the ZIP proteins implies a role in metal ion transport or its regulation (Gaither and Eide 2001).

The mechanism of transport used by the ZIP proteins has yet to be unveiled. Conspicuously, the zinc uptake by human hZip2 zinc transporter has been shown to be energy independent (Gaither and Eide 2000); however, this finding does not correspond with studies of the yeast zinc transporters Zrt1 and Zrt2, which demonstrated strict energy dependence (Zhao and Eide 1996). Fungal and human ZIP proteins may, consequently, use different mechanisms. Zinc uptake by hZip2 was stimulated by HCO_3^- but was not dependent on K^+ or Na^+ gradients (Gaither and Eide 2000); it has been suggested that hZip2 functions *in vivo* by a Zn^{2+}–HCO_3^- symport mechanism. Another possibility is that zinc uptake by these proteins may be driven by the concentration gradient of the metal ion substrate. Although the total level of zinc in cells can be as high as several hundred micromoles (Mantzoros et al. 1998), only small amounts of that zinc are present in a "free" or labile form. Estimates of the labile pool of zinc in cells are in the nanomolar range (Suhy and O'Halloran 1995). A concentration gradient of labile zinc across the plasma membrane may thus be an important driving force for Zn^{2+} uptake. The negative-inside membrane potential existing in cells could also be a driving force for the uptake of zinc (Gaither and Eide 2001; Zhang and Allen 1995).

The CDF (cation diffusion facilitator) proteins were early recognized to often play roles in metal ion transport (Nies and Silver 1995). They are similar to the ZIPs found in organisms at all phylogenetic levels. Many members of this family have been implicated specifically in the transport of zinc from the cytoplasm out of the cell or into organellar compartments (Gaither and Eide 2001). The CDF family was recently said to comprise 13 members; however, more sequence data and better tools have increased the number of members to 101 (Gaither and Eide 2001).

The CDF family can be divided into three different subfamilies (I, II, and III) based on clusters or proteins with greater sequence similarities. CDF subfamily I is found mostly in prokaryota, including both eubacteria and archaea. Subfamilies II and III are comprised of about equal numbers of species of eukaryotic and prokaryotic origin. Six membrane-crossing domains seem to be common in a majority of the members of the CDF family. Their predicted membrane topology is similar to that shown for one such protein, Zrc1, from *Saccharomyces cerevisiae*, shown in Fig. 6.10b (Gaither and Eide 2001).

6.5.2 Zinc Transport and Its Regulation in Plants

Both ZIP and CDF family genes have been discovered in many plant species and have contributed to our increasing understanding of zinc transport and regulation in plants. The number of ZIP family members in plants is remarkable. The *Arabidopsis* genome contains 18 ZIP family genes from three of the four subclasses of ZIP proteins. ZIP subfamily II, however, does not have a plant representative. Plants and animals are multicellular organisms, as reflected by the high number of potential metal ion transport proteins, probably

due to the greater diversity of tissue-specific roles played by these proteins (Gaither and Eide 2001). The first ZIP protein to be discovered in any organism was Irt1 (iron-regulated transporter 1) (Eide et al. 1996). Its gene (*IRT1*) was cloned because its expression in a yeast mutant with an impaired iron uptake suppressed the growth defect of this strain when growth media contained low amounts of iron. Irt1 expression indeed increased iron uptake in this yeast strain, as confirmed by biochemistry (Eide et al. 1996). Later studies showed that Irt1 could also transport Zn^{2+}, Mn^{2+}, and Cd^{2+} (Korshunova et al. 1999). Iron accumulation, rather than the transport of other metals such as zinc, seems to be the main function of Irt1 in plants. In addition, Irt1 is expressed only in the roots of plants for which iron access is restricted. If Irt1 takes part in the accumulation of metals other than iron, such as zinc, this probably occurs only under iron-limiting conditions (Gaither and Eide 2001). It has, indeed, been observed that iron-limited plants accumulate higher levels of other metals such as zinc, manganese, and cadmium (Cohen et al. 1998). These findings corroborate the prediction of Gaither and Eide (2001).

Zip1 through Zip4, ZIP transporters in *Arabidopsis*, may play roles in zinc transport. In *Saccharomyces cerevisiae*, the expression of Zip1, Zip2, or Zip3, each with distinct biochemical properties, results in increased zinc uptake. These proteins, consequently, most probably are zinc transporters. Zip4 expressed in yeast, however, does not result in increased zinc uptake. This may be due to poor expression or mislocalization of the protein in the yeast cell. *ZIP1* is expressed predominantly in roots while *ZIP3* and *ZIP4* mRNA could be found in both roots and shoots. The induction of *ZIP1*, *ZIP3*, and *ZIP4* mRNA takes place under zinc-limiting conditions. A role for these proteins in zinc transport is thus further confirmed. Neither subcellular localization of these proteins nor tissue-specific expression has been determined, so their exact roles cannot be assessed as yet. It is quite clear that some mechanism of regulation exists in plants because there is a zinc-responsive regulation of mRNA levels in response to zinc availability (Gaither and Eide 2001).

If regulation of the expression of zinc transporters was altered in any way, zinc accumulation in plants would probably be greatly impacted. This presumption may in part explain the physiology of an unusual group of plants called *metal hyperaccumulators*. These are plants that take up large quantities of metal ions from the soil. They are of great interest because of their potential to remove metal pollutants from surface soils in a process called *phytoremediation* (Raskin 1995; Gaither and Eide 2001). A well-known hyperaccumulator is *Noccaea caerulescens*, formerly *Thlaspi caerulescens*, a member of the Brassicaceae family that also includes *Arabidopsis*. Certain ecotypes of *N. caerulescens* are capable of accumulating zinc in their shoots at levels up to as much as 30,000 $\mu g g^{-1}$ without evident toxic effects (Gaither and Eide 2001). Plants that are not hyperaccumulators normally accumulate only 0.1% of that level. A salient ability to accumulate and detoxify metal ions should therefore be a significant property of hyperaccumulators. In studies of *N. caerulescens*, it was found that the maximum velocity, V_{max}, was elevated almost fivefold compared to a non-hyperaccumulating ecotype, *Thlaspi arvense*, although there was no difference in the Michaelis-Menten constant K_m (Lasat et al. 2000). Expression of zinc uptake transporters should thus be higher in *N. caerulescens*. *ZNT1*, a ZIP family member, has been cloned from *N. caerulescens* and *T. arvense*. In *T. arvense*, Znt1 is expressed at a low level and regulated by zinc status. In *N. caerulescens* this gene is expressed at a much higher level and is unaffected by zinc availability. The increased zinc accumulation in this and perhaps other metal hyperaccumulating plant species can thus be explained by Znt1 expression.

Many members of the CDF family are also contained in the genomes of plant species. For example, *Arabidopsis* alone encodes ten CDF member genes. The proteins expressed by these genes are likely to function in subcellular zinc compartmentalization as well as in zinc efflux. The Zat protein of *Arabidopsis* is the only plant CDF member to have been studied. This protein seems to be a zinc transporter; however, *ZAT* mRNA expression is not zinc regulated. Transgenic plants overexpressing the *ZAT* gene demonstrate increased zinc resistance. In roots of these transgenic plants, the zinc content was also found to be increased, which suggests that Zat transports zinc into an intracellular compartment (*e.g.*, the vacuole or root cells). In any multicellular organism, zinc transporters are required for both cellular zinc uptake as well as efflux to allow utilization of the metal. Plants, for example, need a zinc efflux transporter to pass zinc from the root tissue into the xylem for distribution to aerial portions of the plant. CDF proteins such as Zat probably perform this function as well (Gaither and Eide 2001).

A relatively novel aspect of the mechanism that allows hyperaccumulators to survive in spite of extremely high concentrations of zinc is the defensins. They are small cysteine-rich peptides with antimicrobial activity. Plant defensins have a characteristic three-dimensional folding pattern that is stabilized by eight disulfide-linked cysteins. Plant defensins are generally considered as secreted proteins and no relationship between defensins and metals was described earlier (Marquès and Oomen 2011). It seems that the particular phenotype of the hyperaccumulating plant is due to the high expression level of defensins rather than to a specific action of the protein.

6.5.3 Zinc Transporters and Their Regulation in Mammals

Both the ZIP and CDF families are represented by several zinc transporters found in mammalian organisms. Fourteen ZIP genes have been identified in humans, and three have been found in the mouse. Functional data are available for three of the human genes (*hZIP1*, *hZIP2*, and *hZIP4*). Recently, a subfamily of mouse zinc transporter genes was characterized (Dufner-Beattie et al. 2003). The proteins hZip1 and hZip2 appear to play roles in zinc uptake across the plasma membrane. Expression of *hZIP2* mRNA has been detected in prostate and uterine tissue as well as monocytes, indicating restricted tissue specificity (Kambe et al. 2004). Overexpressed *hZIP2* in cultured K562 erythroleukemia cells resulted in an increased accumulation of zinc compared to control cells. Furthermore, the hZip2 protein was localized to the plasma membrane of these cells (Kambe et al. 2004). These results indicated that *hZIP2* might serve in zinc uptake in the few tissues where it is expressed (Gaither and Eide 2001).

Endogenous uptake of zinc was shown to be biochemically different from uptake mediated by hZip2 in the K562 cell line in a number of ways. For example, HCO_3^- treatment stimulated zinc uptake mediated by hZip2, whereas the endogenous system did not react. Moreover, several other metal ions (*e.g.*, Co^{2+}, Fe^{2+}, and Mn^{2+}) significantly inhibited zinc uptake by hZip2; however, the endogenous uptake was far less sensitive. It has recently been demonstrated that another ZIP transporter, hZip1, represents the endogenous zinc uptake system in K562 cells (Gaither and Eide 2001). Three important observations support this hypothesis. First, K562 cells express *hZIP1* mRNA, and the functional hZip1 protein is localized to the plasma membrane of these cells. Second, a twofold increase in zinc uptake activity was a consequence of a twofold overexpression of hZIP1 mRNA. It was not possible to distinguish, by biochemical means, the increased uptake of zinc in hZip1-overexpressing cells from the endogenous system. Last, but not least, antisense oligonucleotides targeted to inhibit *hZIP1* expression also inhibited the endogenous zinc uptake activity. The hypothesis that hZip1 is the endogenous transporter in K562 cells is thus strongly supported. The antisense *hZIP1* oligonucleotide treatment reduced zinc uptake to 10–20% of control levels, again corroborating the idea that hZip1 is the major pathway of zinc uptake in these cells (Gaither and Eide 2001).

A wide variety of different cell types demonstrates expression of hZIP1, in sharp contrast to the *hZIP2* gene. The results of Gaither and Eide (2001), therefore, suggest that hZip1 is an important candidate for being the primary factor for zinc uptake in many human tissues. A recent study by Franklin et al. (2003) in which a correlation was found between *hZIP1* expression levels and zinc uptake in human malignant cell lines derived from the prostate provided significant support for this conclusion. Prostate cell lines LNCap and PC-3 possess high levels of zinc uptake activity that is stimulated by prolactin and testosterone. It was found that *hZIP1* is expressed in LNCap and PC-3 cells, and this expression is increased by prolactin and testosterone treatment (Franklin et al. 2003). Expression of *hZIP1* was also repressed by adding zinc to the medium, which suggested that some regulation of zinc uptake occurs in response to cellular zinc status. Lioumi et al. (1999) recently reported a closely related ortholog of hZip1 obtained from the mouse; this protein was named Zirtl for zinc-iron regulated transporter-like protein. The *ZIRTL* gene is expressed in a wide variety of tissues as is *hZIP1*.

A conspicuous finding is that the transporters hZip1 and hZip2 have a surprisingly low affinity for their substrate. The K_m values of both proteins are about 3 μM for free Zn^{2+} ions. Additionally, zinc transporters in a wide variety of mammalian cells have K_m values of the same order. We are faced with an apparent paradox that arises when considering the free Zn^{2+} concentration in mammalian serum. The total zinc concentration of serum is about 15–20 μM, and very little of that amount is present in an unbound form (Zhang and Allen 1995). About 75% of Zn^{2+} is bound to albumin, and 20% is bound to α_2-macroglobulin. What does not exist in free form is complexed with amino acids such as histidine and cysteine. The serum has a high binding capacity for metals; thus, the free Zn^{2+} concentration in serum is estimated to be in the low nanomolar range. It is difficult to understand how such a low concentration of substrate would allow these transporters to contribute to zinc accumulation by mammalian cells under physiological conditions. The cellular requirements of zinc have to be considered and compared to transporter capacity. Recent studies showed that the capacity, expressed as V_{max}, for uptake is so high relative to the cellular demand for zinc that adequate levels can be obtained despite the apparent low affinity of the transporters (Gaither and Eide 2001).

Curiously, a ferrous iron transporter, the DCT1/DMT1/Nramp2 Fe^{2+}, may be involved in zinc uptake. This transporter is a member of the Nramp family of transporters and is not related to ZIP or CDF proteins. Experiments with *Xenopus* oocytes suggested that cation influx currents could indicate Zn^{2+} movement across the membrane; however, more recent results have indicated that the currents recorded in these oocytes result from Zn^{2+}-induced proton fluxes rather than transport of the metal ion (Sacher et al. 2001).

Export of zinc from the cell as well as transport into intracellular organelles is related to mammalian CDF family members. Seven CDF genes in humans and six in the mouse genome have been identified plus a small number of others from the rat and other mammals. Four of the mammalian

genes, *ZnT-1*, *ZnT-2*, *ZnT-3*, and *ZnT-4*, have been functionally characterized to such an extent that their roles in zinc metabolism are not in doubt. ZnT-1 is a zinc export transporter in the plasma membrane of mammalian cells; consequently, ZnT-1 may play a role in the cellular detoxification of zinc by exporting unnecessary metal ions out of the cell. An observation that cells overexpressing this transporter show higher zinc resistance than control cells further corroborates this role. ZnT-1 may also be involved in the dietary absorption of zinc in the intestine as well as in the reabsorption of zinc from urine in the renal tubules of the kidney. The intestinal enterocytes of the duodenum and the jejunum (*i.e.*, the primary sites of zinc absorption) express ZnT-1, and the protein is found localized to the basolateral membrane. This indicates a role in transporting zinc out of the enterocyte and into the bloodstream. The protein is also found on the basolateral surface of renal tubule cells, where it would be expected to appear to be involved in transporting zinc that has been reabsorbed from the urine back into the circulation. It is well established that the loss of zinc in urine is very low because of an efficient renal reabsorption (Gaither and Eide 2001).

Intracellular zinc sequestration and storage may be dependent on ZnT-2, a role similar to that proposed for Zrc1 and Cot1 in yeast. ZnT-2 is located in the membrane of the late endosome that accumulates zinc when cells are grown under high zinc conditions (Palmiter et al. 1996).

The third CDF transporter, ZnT-3, has a role similar to that of ZnT-2. It also transports zinc into an intracellular compartment where the metal may play a role in neuronal signaling. Messenger RNA of *ZnT-3* has been detected only in the brain and testis and is most abundant in the neurons of the hippocampus and the cerebral cortex. The protein is localized in membranes of synaptic vesicles in these neurons, which suggests that the protein transports zinc into this compartment. This hypothesis is further supported by the fact that a subset of glutamatergic neurons contains histochemically reactive zinc in their synaptic vesicles. The ZnT-3 protein was colocalized with these zinc-containing vesicles. Furthermore, a mouse line lacking *ZnT-3* did not accumulate zinc in these vesicles; therefore, the protein ZnT-3 must be required for the transport of zinc into synaptic vesicles in some types of neurons where it may play a neuromodulatory role (Cousins and McMahon 2000).

The protein ZnT-4 is expressed in the mammary gland, brain, and small intestine. In the mammary gland, it is responsible for zinc transport into milk. In fact, mutations in the *ZnT-4* gene produced a mutant mouse referred to as the lethal milk (lm) mouse. This mutant gene is the *lm* gene. Pups of any genotype suckled on *lm*/*lm* dams die before weaning, and the cause of death is zinc deficiency from an insufficient supply of zinc in the milk. In intestinal enterocyte ZnT-4 is localized in endosomal vesicles concentrated at the basolateral membrane. It seems that ZnT-4, in a manner similar to ZnT-1, may facilitate transport of zinc into the portal blood (Kambe et al. 2004).

It is becoming increasingly clear that regulation of zinc export in many cell types is managed by zinc. It has been demonstrated that *ZnT-1* mRNA is upregulated during ischemia, which is known to cause zinc influx into neurons. Cultured neurons transiently increased *ZnT-1* mRNA when exposed to zinc, a finding that is in accordance with *ZnT-1* regulation as a result of zinc influx (Gaither and Eide 2001). Transcriptional control of *ZnT-1* could therefore contribute to zinc detoxification by stimulating its export. As *ZnT-1* is expressed in many cell types, this could be a general mechanism of cellular zinc homeostasis. Zinc absorption may also be dependent on the transcriptional control of *ZnT-1*. Messenger RNA levels were found to be increased in enterocytes following an oral dose of zinc (Cousins and McMahon 2000). The location of the ZnT-1 protein on the basolateral membrane of these cells suggests that upregulation of ZnT-1 promotes zinc absorption by facilitating transport into the portal blood.

Regulation of zinc uptake transporters in mammals is less well known. Evidence suggests that the activity of these transporters is controlled by the levels of zinc. Zinc uptake in brush border membrane vesicles, for example, has been found to increase in zinc-deficient rats. Additionally, cultured endothelial cells grown under low-zinc conditions displayed a higher rate of zinc uptake than zinc-replete cells. Zinc deficiency may thus increase the expression or activity of zinc uptake transporters in some cell types. A hypothetic mechanism of this regulation could be similar to that described in yeast. This is supported by the finding that *hZIP1* mRNA levels in cultured malignant prostate cells were reduced when treated with zinc. This suggests a transcriptional control mechanism. Such a mechanism would play a critical role in mammalian zinc homeostasis (Gaither and Eide 2001; Kambe et al. 2004).

Although the intracellular trafficking of zinc from the apical cytoplasm to the basolateral cytoplasm of the enterocytes remains largely uncharacterized a few recent findings are worthy of mentioning. ZnT7 has been reported to be predominantly located in the Golgi apparatus. In addition, it seems that ZnT7 is required for the activity of alkaline phosphate. The expression pattern of ZnT7 suggests a role in dietary zinc uptake (Wang and Zhou 2010). Similar to ZnT7, another transporter ZnT5A transports zinc into the Golgi apparatus and contributes to the activity of zinc-dependent enzymes. It may be that ZnT5A and ZnT7 execute partially overlapping functions in intestinal zinc homeostasis. The smaller variant ZnT5B mediates bidirectional zinc transport (Valentine et al. 2007). It has been shown that ZnT6 forms a heteromeric complex with ZnT5A to activate zinc-dependent enzymes in the secretory

pathway. ZnT6 may have a role in zinc excretion from enterocytes into lumen via exocytosis. Recent studies have revealed two other Golgi-residing Zip transporters, Zip9 and Zip 13. However, no information was available regarding their expression in small intestine (Matsuura et al. 2009).

6.6 Uptake and Rejection of Copper

The transport and cellular metabolism of copper depends on a series of membrane proteins and smaller soluble peptides that comprise a functionally integrated system for maintaining cellular copper homeostasis. Inward transport across the plasma membrane appears to be a function of integral membrane proteins that form the channels that select copper ions for passage. Two membrane-bound, copper-transporting ATPase enzymes – ATP7A and ATP7B (the products of Menkes' and Wilson's disease genes, respectively) – catalyze an ATP-dependent transfer of copper to intracellular compartments or expel copper from the cell. ATP7A and ATP7B work in concert with a series of smaller peptides, the copper chaperones, which exchange copper at the ATPase sites or incorporate the copper directly into the structure of copper-dependent enzymes such as cytochrome *c* oxidase and Cu–Zn superoxide dismutase (CuZnSOD). This enzyme is found in the cytoplasm and scavenges the superoxide anion, a reactive oxygen species. These mechanisms come into play in response to a high influx of copper or during the course of normal copper metabolism.

In the gut, copper is absorbed by enterocytes before being distributed throughout the system. Daily net copper absorption in the intestine is approximately 0.6–1.6 mg day^{-1}. Roughly 4.5 mg day^{-1} copper is excreted into the gastrointestinal tract, predominantly in pancreatic juice (ca. 2 mg day^{-1}) and bile (ca. 2.5 mg day^{-1}), with small amounts of copper in saliva and gastric juice. Absorption of copper occurs mainly in the stomach and duodenum with the dominating part in the duodenum (van den Berghe and Klomp 2010).

6.6.1 Accessing the Intracellular Pool

In a defined culture medium (an artificial environment), cells can get copper ions from a great number of suitable donors. The uptake process is usually rapid and, curiously enough, does not depend on the ATP status of the cell, so it seems that the uptake process is not energy demanding. It is tempting to conclude that a passive copper transport system exists in the membrane (Tong and McArdle 1995). Plasma factors seem to have a similar capability to influence cellular access *in vivo*. The capacity of amino acids and chloride and bicarbonate ions to stimulate copper uptake gives some support for this idea (Harris 2000). Inhibitors of protein biosynthesis seem to influence increased uptake of copper. This indicates that the transport system has a two-sided character and accounts for both import and export (Harris 2000).

To work properly, a transport mechanism requires an easy exchange of copper. There has been some discussion about whether albumin- and ceruloplasmin-bound copper actually could be sources of copper for the tissues; however, a reducing environment, be it in plasma or in the membrane, can compromise binding strengths, making release of copper possible. These can be compromised by reducing systems in the membrane. Even amino acids could interact with the copper–protein complex (Harris 2000).

As dietary copper reaches the duodenum it comes in contact with copper transporter 1 (CTR1). This transporter is believed to be the primary protein responsible for intake of copper across the brush border microvilli. CTR1 is an integral plasma membrane protein with three membrane spanning domains that forms a homotrimeric pore for import of Cu^{+}. Dietary reducing agents like ascorbate may provide the suitable substrate. It might be that cupric reductases are necessary for this process. Perhaps does this include Dcytb, a putative Fe^{3+} reductase, or Steap2, a newly hypotheisized Fe^{3+} reducates that could also reduce Cu^{2+} (Prohaska 2008).

6.6.2 Albumin as a Copper Transport Factor

Albumin is the most abundant protein in plasma; therefore, it is a good candidate for copper transport because it can bind copper in several sites on the protein (Masuoka and Saltman 1994). The best candidates for transportation seem to be the sites at the extreme N terminal region or certain cysteine residues within the protein. At the N terminal, a histidine at position 3 may bind copper (Cu^{2+}); however, this site does not reach a rapid equilibrium with unbound copper. Histidine forms stable complexes with Cu^{2+}, and the histidine concentration in plasma is about 135 μM. These factors favor histidine as a transport ligand for copper. The interaction is reflected by a ternary complex with albumin–Cu(II) and histidine. Furthermore, the interaction induces the protein to release copper as a histidine–Cu(II) complex. It seems, however, that the histidine ligand is not transported across the membrane (Hilton et al. 1995). This would suggest that the role of histidine ends at the cell surface (Harris 2000).

6.6.3 Ceruloplasmin as Copper Transporter

Ceruloplasmin was first isolated from plasma and characterized as a copper-containing protein by Holmberg and Laurell (1948). This protein is a member of a class of proteins known as multicopper oxidases, which are characterized by three distinct copper sites. Ceruloplasmin is also a ferroxidase and plays an important role in oxidizing Fe^{2+} to Fe^{3+} for incorporation in apotransferrin. As a copper transporter, it contains about 95% of the copper in plasma. Multicopper oxidases utilize the electron chemistry of bound copper ions to couple substrate oxidation with the four-electron reduction of dioxygen. Electrons pass from the substrate to the type I copper, then to the trinuclear copper cluster, and subsequently to the oxygen molecule bound at this site (Garrick et al. 2003). Human ceruloplasmin is encoded in 20 exons encompassing about 65 kb of DNA localized to chromosome 3q23–q24. The human ceruloplasmin gene in hepatocytes is expressed as two transcripts of 3.7 and 4.2 kb. These come from use of alternative polyadenylation sites within the 3′ untranslated region. When these transcripts are expressed in the liver, the 1046-amino-acid protein ceruloplasmin is found in plasma (Bielli and Calabrese 2002).

Recent findings suggest that intracellular copper transport is regulated by the ATP7B, the Wilson disease gene product, ceruloplasmin and Niemann-Pick C1 protein. If ATP7B is properly localized in the late endosomes it makes possible the appropriate export of copper from hepatocytes. This is accomplished by means of ceruloplasmin secretion and biliary excretion of copper via lysosomes. Probably, some copper is incorporated into the late endosome lumen from cytoplasm and is transported to the *trans*-Golgi network by a Niemann-Pick-C1-dependent mannose-6-phosphate receptor transport system, where after it is incorporated into apo-ceruloplasmin in the *trans*-Golgi network. Holo-ceruloplasmin, the protein with copper incorporated, is then secreted to the sinusoidal space through the secretory pathway (Yanagimoto et al. 2011).

6.6.4 Membrane Transport of Copper

Some years ago, a major area of investigation was agents that transported copper in plasma. Lately, however, interest has been focused on interactions at the membrane surface and transport intracellularly. What brought about this change in focus was the identification of specific copper-transporting proteins in the membrane as well as within the cell. Yeast has both low- and high-affinity systems for copper uptake, and these mediate copper transport (Eide 1998). In yeast, the copper transport 1 (*CTR1*) gene was the first to be identified. Surprisingly, this gene was not directly for copper transport but was essential for iron transport in *Saccharomyces cerevisiae*. In fact, the copper was required for Fet3. This is a multicopper ferroxidase that catalyzes the oxidation of Fe^{2+} to Fe^{3+} to make absorption by a ferric transport protein possible (Dancis et al. 1994). The protein Ctrp1 is expressed by this gene and is a membrane-crossing protein. In addition, it is heavily glycosylated, with a serine- and methionine-rich composition in which the structural motif methionine–X–X–methionine is repeated 11 times. Ctrp1 transports Cu^{+} and not Cu^{2+} or any other metal ions. The identification of this protein establishes a mechanistic link between copper and iron uptake (Harris 2000). Strains of *S. cerevisiae* are equipped with a second high-affinity transporter gene, *CTR3* (Knight et al. 1996). The expressed protein, Ctrp3, restores copper-related functions in strains lacking the *CTR1* gene. A structurally similar transporter gene in *Arabidopsis thaliana*, *COPT1*, was later discovered as well as the human transporter *hCTR1*. In HeLa cells, the *hCTR1* gene is located on chromosome 9 (9q31/32, to be more precise) (Harris 2000). The human hCtr1 protein is much smaller than the yeast Ctrp1 protein, based on cDNA sequence data. All these copper transport proteins recognize only Cu^{+}; thus, there must be a reductase in the membrane to reduce Cu^{2+} to Cu^{+} at the moment of membrane penetration. In yeast, the expression of *FRE1* and *FRE2* reduces both Cu^{2+} and Fe^{3+} for transport. Expression of a third gene, *FRE7*, increases when extracellular copper becomes very limiting (Martins et al. 1998).

The first important factor to take into consideration is that the cytosolic environment is highly reducing. This means that reducing Cu^{2+} to Cu^{+} is a straightforward process. The nitrogen-containing molecule glutathione (γ-glutamylcysteinglycine, or GSH) is the most common intracellular thiol. It has a concentration in mammalian cells of 0.5–10 m*M* and is a reducing agent in many reactions. Glutathione was one of the earliest intracellular components identified with copper transport (Denke and Farburg 1989). Glutathione may also play the role of a general transporter of copper ions by delivering copper to Ctr1 in the plasma membranes. Glutathione reduces and binds Cu^{+} and delivers it to metallothioneins and to some copper-dependent apoenzymes such as superoxide dismutase and hemocyanin (Tapiero et al. 2003). Formation of Cu (I)–GSH is a spontaneous reaction apparently independent of enzyme involvement. When cellular GSH is low, cells are slower to take up copper from the medium and have a lower cellular concentration at steady state (Harris 2000). A decline in GSH levels in a cell impairs the subsequent binding of copper to apo-CuZnSOD or the delivery of copper to the cytosolic enzymes; thus, Cu–GSH complexes have the capacity to mediate Cu(I) transfer to a variety of binding sites on macromolecules. The function of GSH, then, extends beyond that of preventing copper toxicity to playing an important role in internal copper metabolism (Harris 2000).

A new family of soluble metal receptor proteins acting in the intracellular trafficking of metal ions is the metallochaperones. These metal receptors do not act as scavengers or detoxifiers. On the contrary, they act in a "chaperone-like" manner by guiding and protecting metal ions while facilitating appropriate partnerships (O'Halloran and Culotta 2000). Copper chaperones are a family of cytosolic peptides that form transient complexes with Cu^{+}. An invariant methionine–X–cysteine–X–X-cysteine metal-binding motif in the N-terminal region is a structural feature of most chaperones. Their function is to guide copper ions in transit to specific proteins that require copper (Harris 2000). Significant examples of metallochaperones are the gene products of the mercury resistance (*mer*) operon of the Gram-negative transposon Tn*21*, including MerP. MerP is a small, soluble periplasmic protein that transports Hg^{2+} to a membrane transporter and eventually to a reductase that reduces the Hg^{2+} to the volatile Hg^{0} as part of a detoxifying mechanism (Hamlett et al. 1992). Chaperones for copper act in a similar manner, moving copper from one location in the cell to another, often crossing membrane boundaries. In contrast to nuclear activating factors, chaperones demonstrate no capacity to enter the nucleus or interact with DNA. ATX1 and copZ are copper chaperones in *Saccharomyces cerevisiae* and *Enterococcus hirae*, respectively, and both have ferredoxin-like folds with the βαββαβ motif in the folded chains. Copper as Cu(I) is bound to two cysteine sulfur groups, forming a linear bidentate ligand. This structure probably allows for easy exchange of the bound copper to a structurally similar copper-binding site on the receiving protein (Harris 2000). Of utmost significance is that other structural features allow the peptide to specify the target proteins with virtually no possibility for mismatch. The targets for these copper chaperones are cytochrome *c* oxidase in the mitochondria, ATP7B in the trans-Golgi, and the apo form of copper- and zinc-dependent superoxide dismutase (CuZnSOD) in either peroxisomes or the cell cytosol. The obvious advantage of a chaperone is that it selects the receiving molecule with high precision. This is not the case with Cu(I)–GSH, which has no target-specifying property. A conceptual disadvantage is that chaperones force a partitioning of copper into multiple pools in order to replenish copper enzymes (Harris 2000). To date, three chaperones have been described in yeast, and all three are known to share structural features with mammalian and plant counterparts. The following is a brief description of known chaperones.

In yeast cells lacking superoxide dismutase activity (*sod1*Δ) and auxotrophic for lysine, *ATX1* (antioxidant 1) was identified as an antioxidant gene that suppressed oxygen toxicity (Lin et al. 1997). The human homolog of *ATX1*, named *ATOX1*, was shown to complement yeast lacking *ATX1*. Like ATX1, the human ortholog contains one methionine–threonine–cysteine–X–glycine–cysteine copper-binding domain (Harris 2000). The copper chaperone (CCH) in *Arabidopsis* has a 36% sequence identity with ATX1. Chaperones with the ATX1 structural domain target P-type ATPases. The ATPase in yeast is Ccc2p, a membrane-bound protein that mediates the transfer of copper to a late or post Golgi compartment. By the same reasoning, ATOX1 in mammals is thought to target ATP7B, which is a P-type ATPase that occurs in Wilson's disease. The ATPase ATP7B transfers copper to apoceruloplasmin or forces its extrusion into the bile, whereas Ccc2p transports copper to Fet3p, which is a multicopper oxidase that oxidizes Fe^{2+} to Fe^{3+} for incorporation into the ferric ion transporter and subsequent delivery (Harris 2000).

Another copper chaperone, COX17, carries out the transport of copper to cytochrome oxidase in the mitochondria of the yeast *Saccharomyces cerevisiae*. A human homolog of COX17 has also been reported. Cox17, acting as a mitochondrial copper shuttle, is the only known chaperone that violates the glycine–methionine–X–cysteine–X–X-cysteine consensus motif. The copper binding sites on Cox17 are cysteine residues occurring in tandem (Cys14 and Cys16) and positioned near the N terminus. A unique property is the binding of Cu(I) to Cox17, which is a binuclear cluster that is similar to the copper cluster in metallothionein, with the exception that the Cox17–Cu(I) complex is more labile. In *S. cerevisiae* the delivery of copper to cytochrome oxidase apparently involves two inner mitochondrial membrane proteins, SCO1 and SCO2, which are penultimate receivers of the copper (Harris 2000).

LYS7 is a 27-kDa copper chaperone that delivers copper to the apo-SOD1. There are mutants of yeast LYS7 mutants; they are defective in SOD1 (the gene encoding CuZnSOD) activity and are unable to incorporate copper into the protein. A single methionine–histidine–cysteine–X–X–cysteine consensus sequence is present in the N-terminal region of the protein. The copper chaperone for SOD (CCS) is a human counterpart of comparable size and 28% sequence identity (Harris 2000).

The most significant finding with regard to copper transport in cells is without a doubt the discovery, cloning, and sequencing of the genes responsible for Menkes' and Wilson's diseases. These diseases have constituted important models of abnormal copper metabolism in humans. The etiology of these diseases was unknown for a long time. Classical Menkes' disease is an X-linked (*i.e.*, strikes mainly males) copper deficiency. The incidence is about 1 in 200,000. Boys with the disease usually do not survive past 10 years of age. Manifestations of the disease are a series of enzyme defects. Hypopigmented hair is caused by a deficiency of the tyrosinase required for melanin synthesis. Connective tissue abnormalities, including aortic aneurysms, loose skin, and

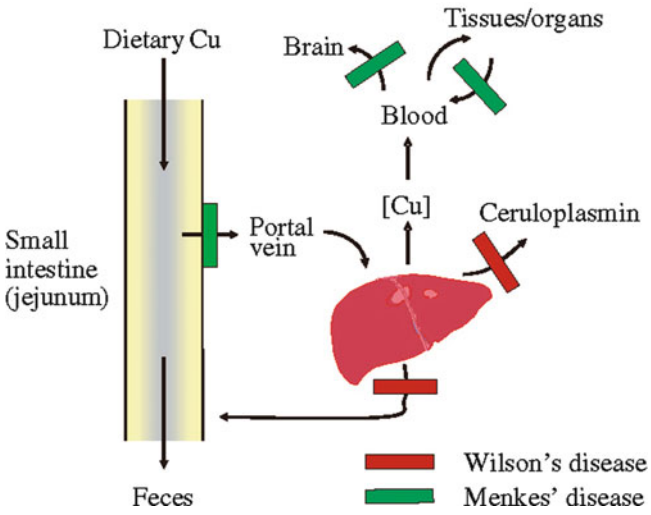


Fig. 6.11 A schematic view of copper pathways and the blocks in Menkes' disease and Wilson's disease. Copper is taken up by enterocytes in the small intestine and exported over the basolateral membrane into the portal circulation. The export mechanism is defective in Menkes' disease; consequently, an overload of copper occurs in the enterocytes and subsequent copper deficiency in the organism. The liver normally takes up most of the copper in the portal circulation. If there is a copper overload, excess copper is excreted in the bile. This process is blocked in Wilson's disease. Furthermore, the delivery of copper to ceruloplasmin is also blocked (Adapted from Mercer 2001)

fragile bones, result from reduced lysyl oxidase activity and consequent weak cross-links in collagen and elastin. Severe neurological defects are a predominant feature of the classical form of the disease and possibly result from the reduced activities of cytochrome oxidase; however, effects due to the reduced activity of superoxide dismutase, peptidylglycine-α-amidating monooxygenase, and dopamine-β-monooxygenase may also contribute to brain abnormalities (Suzuki and Gitlin 1999). All of these enzymes are copper dependent. A diagnostic feature is the unusual steely or kinky hair caused by reduced keratin cross-linking, a process that also is copper dependent. The copper deficiency in Menkes' disease is caused by reduced uptake of copper across the small intestine (Fig. 6.11) catalyzed by the Menkes' protein ATP7A, compounded by defective distribution of copper within the body wherever ATP7A is required for copper transport. ATP7A is involved in the transport of copper across the blood–brain barrier, which explains the marked brain copper deficiency and consequent severe neurological abnormalities in patients (Suzuki and Gitlin 1999).

Wilson's disease, or hepatolenticular degeneration, is an autosomal recessive copper toxicosis condition with an incidence of 1 in 50,000 to 1 in 100,000, depending on the population. Very high concentrations of copper accumulate in the liver because of impaired biliary excretion of copper or failure to incorporate copper into ceruloplasmin, the major copper-binding protein in the circulation. This will ultimately cause the death of hepatocytes. The disease has a variable age of onset but is rarely observed in children younger than five and can present as a hepatic or neurologic disease. Copper may be released from damaged hepatocytes and accumulates in extrahepatic tissues, including the central nervous system. The diagnostic copper deposits that can sometimes be seen in the cornea of the eyes are known as the Kayser-Fleischer rings disorder, which results from pathological accumulations of copper, predominantly in the liver and brain tissues. The dominant symptoms relate to a failure to release liver copper into bile. Both diseases have provided unprecedented molecular insights into genetic factors that regulate copper transport and bioavailability to organs and tissues (Suzuki and Gitlin 1999).

Menkes' and Wilson's diseases are caused by mutations in genes on the X chromosome (Xq13) and chromosome 13 (13q14.3). The isolation and sequencing of these disease genes have revealed that both code for P-type Cu-ATPases. The Wilson's (ATP7B) and Menkes' (ATP7A) proteins are specific copper transporters; furthermore, ATP7B and ATP7A have a 57% sequence homology to one another. In addition, they have remarkable parallels to copper-binding proteins in bacteria. The gene for Menkes' disease spans about 150 kb. Its mRNA is 8.3–8.5 kb, encompassing 23 exons that range in size from 77 to 4,120 bp, with a single open reading frame and an ATG start codon in the second exon. Exon 23 contains the TAA stop codon, 274 bp that are translated, and a 3.8-kb untranslated region that has the polyadenylation site. When analyzed as a cDNA, ATP7A mRNA encodes a protein of exactly 1,500 amino acids; however, it may have additional nucleotide sequences at the 5′ end. A 22-amino-acid presequence generated by an in-frame ATG site upstream occurs in some ATP7A transcripts (Harris 2000). Strong expression of Menkes' disease mRNA is observed in muscle, kidney, lung, and brain. In placenta and pancreas, the expressions are weaker, and liver shows only traces. The Wilson's disease transcript is 7.5 kb and encodes a protein of 1,411 amino acids. In contrast to the Menkes' disease gene, the Wilson's disease gene is strongly expressed in the liver and kidney (Harris 2000).

The biological functions of ATP7A and ATP7B are different, although their structures are similar. The Menkes' disease protein (ATP7A) seems to be responsible for the regulation of copper-ion release at the outer membrane. Experiments with Chinese hamster ovary cells have shown that overexpression of ATP7A makes the cells tolerate highly toxic amounts of copper in their immediate environment (Camakaris et al. 1995). Superior tolerance is manifested by forced expulsion that prevents copper accumulation. The similarity in overall appearance between ATP7A and ATP7B might lead to the conclusion that they are similarly distributed. ATP7B, in contrast to the membrane association

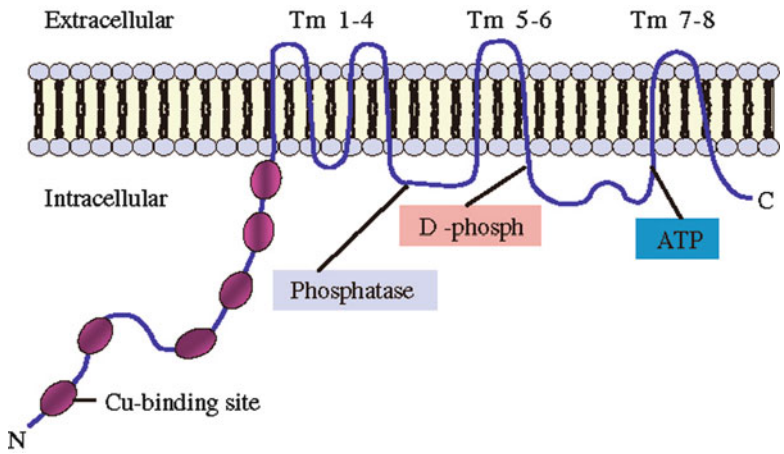


Fig. 6.12 An illustration of the Menkes' and Wilson's Cu-ATPases. Eight membrane-crossing (Tm) domains are predicted for both proteins. Most of the protein is localized in the cytoplasm. In addition to the ATP-binding site, phosphatase domain and phosphorylated aspartic acid (D-phosph) is common to all P-type ATPases. The N terminal contains six copper-binding motifs that interact with copper chaperones (Adapted from Mercer 2001)

of ATP7A, resides within an internal organelle of the cell, where it functions to incorporate copper into apo-ceruloplasmin. This process takes place in either the ER or a Golgi compartment. Additionally, the protein works to force the release of copper into the bile (Harris 2000).

The structure of ATP7A contains a comparatively large, heavy-metal binding domain (Hmb). This domain is comprised of six metal-binding cysteine clusters within the structural motif glycine–methionine–threonine/histidine–-cysteine–X–serine–cysteine which contain eight transmembrane (Tm) regions. The purpose of these Tm regions is to guarantee anchorage and orientation of the protein. The correct assembly defines the channel through which copper ions pass. ATP7A is a type II membrane protein and is thus defined because both the —NH_2 and the —COOH termini are on the cytosolic side of the membrane. Two flexible loops, one on a 135-residue chain and the second on a 235-residue chain, extend into the cytosol. The smaller one is between Tm 4 and 5 and the larger one is between Tm 6 and 7 (Harris 2000). The bacterial CopA protein contains a smaller Hmb with a single glycine–methionine–threonine/histidine–cysteine–X–X–X motif (Solioz et al. 1994). A yeast Cu-ATPase has the motif at most twice, which leads one to speculate whether the 650-residue Hmb region performs some transport-related function other than binding copper for export. A cysteine–proline–cysteine motif in Tm 6 is thought to build the channel that allows copper in the cytosol to be transported across the membrane (Fig. 6.12). Copper(I) is the preferred ion to bind to ATP7A and ATP7B. There is little or no affinity of the Hmb to bind to Fe(II), Fe (III), Ca(II), Mg(II), Mn(II), or Ni(II); however, Zn(II) shows some binding (Harris 2000).

The perinuclear area within the region of the cell thought to represent the Golgi seems to be the localization of ATP7A. ATP7A-loaded vesicles have been postulated to continually be moving between the Golgi and plasma membrane. Experiments show that high concentrations of copper in the exterior of the cell induce movement of the marked vesicles to the cell boundary. The Wilson's disease ATP7B is also localized in the Golgi. A truncated homolog of ATP7B that lacks four of the eight membrane-spanning domains seems to reside in the cytosol. A structural analog of the Wilson's disease gene or Menkes' disease gene, the *CCC2* gene in yeast, also encodes a P-type ATPase that exports cytosolic copper to the extracytosolic domain of Fet3p, a copper oxidase required for iron uptake (Harris 2000).

The mouse homolog of ATP7A, the Atp7a, is expressed in all tissues but is particularly strong in the choroid plexus of the brain. The localization in the brain, more specifically the blood–brain barrier, places the ATPase in a strategic position to control the flow of copper into the ventricles of the brain. An animal model of Menkes' disease is the macular mutant mouse. A gradual erosion of cytochrome *c* oxidase activity in the brain has been observed in this model. The effect can be partially prevented by a single injection of copper in an early perinatal period (Meguro et al. 1991). Copper shows a propensity to accumulate in brain blood vessels and in astrocytes which apparently hinders its movement to neurons (Kodama 1993); furthermore, embryonic mouse liver expresses Atp7a mRNA in contrast to adults. The rat homolog of ATP7B, Atp7b, is expressed early only in the central nervous system, heart, and liver. With development, Atp7b appears in intestine, thymus, and respiratory epithelia (Kuo et al. 1997). Transfection assays in yeast have demonstrated that the two ATPases share biochemical functions. Atp7b in hepatocytes from (LEC) rats, a model for Wilson's disease, mimics the sequestration of ceruloplasmin in cotransfected cells. The data support a metabolic connection between the plasma copper protein and Wilson's disease ATPase (Cox and Moore 2002). ATP7B may also be localized on the apical surface of hepatocytes, a location that allows the protein to expel copper ions into the bile (Fuentealba and Aburto 2003).

It seems that Atx1 is involved in copper supply to amine oxidase 1 in *Schizosaccharomyces pombe*, fission yeast, which is localized in the cytosol rather than the Golgi. Of

the copper chaperones, only disruption of Atx1 caused loss of enzymatic activity. *Arabidopsis thaliana* has a second Atx1 copper chaperone with an extra C-terminal domain that is inhibitory to its interactions with the P_1-type heavy metal ATPase 5 (Burkhead et al. 2009).

The expression of the *cop* operon, which regulates copper homeostasis, is low in mutants of *Enterococcus hirae* missing CopZ. The low expression includes copA and copB as well as hypersensitivity to copper resulting from insufficient copper-exporting P_1-type ATPase activity. If CopZ only donated Cu(I) ions to a copper exporter, intracellular copper levels should be elevated. In addition, copper-responsive expression of the *cop* operon should be higher. Consequently, an alternative target for CopZ has been suggested. Transcriptional regulation of the *cop* operon is mediated by CopY, which is a DNA-binding repressor. CopY, in an purified form, holds one Zn(II) ion and this can be displaced by two copper ions to form a Cu(I)-thiolate cluster. The replacement of zinc with copper impairs the binding of CopY to nucleotide sequences from the *cop* operon promoter region (Cobine et al. 2002).

Cyclin D1, which is involved in the transitions from G1 to S phase in the cell cycle, can be affected by Atx1 in a way that influences the transcriptional control. Even mouse fibroblast proliferation may be affected by Atx1. A two-hybrid interaction, in yeast, also takes place between Ccs1 and the β-secretase 1 that cleaves the amyloid precursor protein and generates the Aβ peptides that aggregate in Alzheimer's disease (Robinson and Winge 2010).

See Also the Following Chapter. Chapter 7 (Biological Functions of the Elements)

References

Andrews NC (1999) Disorders of iron metabolism. N Engl J Med 341:1986–1995

Bielli P, Calabrese L (2002) Structure to function relationships in ceruloplasmin: a "Moonlighting" protein. Cell Mol Life Sci 59:1413–1427

Burkhead JL, Reynolds KA, Abdel-Ghany SE, Cohu CM, Pilon M (2009) Copper homeostasis. New Phytol 182:799–816

Camakaris J, Petris MJ, Bailey L, Shen PY, Lockhart P, Glover TW, Barcroft C, Patton J, Mercer JF (1995) Gene amplification of the Menkes (MNK: ATP7A) P-type ATPase gene of CHO cells is associated with copper resistance and enhanced copper efflux. Hum Mol Genet 4:2117–2123

Cannone-Hergeaux F, Gruenheid S, Ponka P, Gros P (1999) Cellular and subcellular localization of the Nramp2 iron transporter in the intestinal brush border and regulation by dietary iron. Blood 93:4406–4417

Carpenter CE, Mahorey AW (1992) Contributions of heme and nonheme iron to human nutrition. CRC Crit Rev Food Sci Nutr 31:333–367

Cobine P, George GN, Jones CE, Wickramasinghe WA, Solinoz M, Dameron CT (2002) Copper transfer from the Cu(I)chaperone, CopZ, to the repressor, Zn(II)CopY: metal coordination environments and protein interactions. Biochemistry 41:5822–5829

Cohen CK, Fox TC, Garvin DF, Kochian LV (1998) The role of iron-deficiency stress responses in stimulating heavy-metal transport in plants. Plant Physiol 116:1063–1072

Collawn JF, Lai A, Domingo D, Fitch M, Hatton S, Trowbridge IS (1993) YTRF is the conserved internalisation signal of the transferrin receptor, and a second YTRF signal at position 31–34 enhances endocytosis. J Biol Chem 268:21686–21692

Conrad ME, Umbreit JN (2000) Iron absorption and transport – an update. Am J Hematol 64:287–298

Conrad ME, Umbreit JN, Moore EG (1999) Iron absorption and transport. Am J Med Sci 318:213–229

Costa E, Rocha S, Rocha-Pereira P, Reis F, Castro E, Teixeira F, Miranda V, do Sameiro Faria M, Loureiro A, Quintanilha A, Belo L, Santos-Silva A (2008) DMT1 (NRAMP2/DCT1) genetic variability and resistance to recombinant human erythropoietin therapy in chronic kidney disease patients under haemodialysis. Acta Haematol 120:11–13

Cousins RJ, McMahon RJ (2000) Integrative aspects of zinc transporters. J Nutr 130:1384S–1387S

Cox DW, Moore SD (2002) Copper transporting P-type ATPases and human disease. J Bioenerg Biomembr 34:333–338

Dancis A, Yuan DS, Haile D, Askwith C, Eide D, Moehle C, Kaplan J, Klausner RD (1994) Molecular characterization of a copper transport protein in *S. cerevisiae*: an unexpected role for copper in iron transport. Cell 76:393–402

Davidsson L, Lönnerdal B, Sandström B, Kunz C, Keen CL (1989) Identification of transferrin as the major plasma carrier protein for manganese introduced orally or intravenously or after *in vitro* addition in the Rat. J Nutr 119:1461–1464

Denke SM, Farburg BL (1989) Regulation of cellular glutathione. Am J Physiol 257:L163–L173

Donovan A, Brownile A, Zhou Y, Shepard J, Pratt SJ, Moynihan J, Paw BH, Drejer A, Barut B, Zapata A, Law TC, Brugnarall C, Lux SE, Pinkus GS, Pinkus JL, Kingsley PD, Pails J, Fleming MD, Andrews NC, Zon LI (2000) Positional cloning of Zebrafish ferroportin identifies a conserved vertebrate iron exporter. Nature 403:776–781

Dufner-Beattie J, Langmade SJ, Wang F, Eide D, Andrews GK (2003) Structure, function, and regulation of a subfamily of mouse zinc transporter genes. J Biol Chem 278:50142–50150

Eide DJ (1998) The molecular biology of metal Ion transport in Saccharomyces cerevisiae. Annu Rev Nutr 18:441–469

Eide D, Broderius M, Fett J, Guerinot ML (1996) A novel iron-regulated metal transporter from plants identified by functional expression in yeast. Proc Natl Acad Sci U S A 93:5624–5628

Eisenstein RS (2000) Iron regulatory proteins and the molecular control of mammalian iron metabolism. Annu Rev Nutr 20:627–662

Finch C (1994) Regulators of iron balance in humans. Blood 84:1697–1702

Fleming MD, Trenor CC, Su MA, Foernzler D, Beier DR, Dietrich WF, Andrews NC (1997) Microcytic anemia mice have a mutation in Nramp2, a candidate iron transporter gene. Nat Genet 16:383–386

Fleming RE, Miqas MC, Zhou XY, Jiang J, Britton RS, Brunt EM, Tomatsu S, Waheed A, Bacon BR, Sly WS (1999) Mechanism of increased iron absorption in murine model of hereditary hemochromatosis: increased duodenal expression of the iron transporter DMT1. Proc Natl Acad Sci U S A 96:3143–3148

Franklin RB, Ma J, Zou J, Guan Z, Kykoyi BI, Feng P, Costello LC (2003) Human ZIP1 is a major zinc uptake transporter for the accumulation of zinc in prostate cells. J Inorg Biochem 96:435–442

Fuentealba IC, Aburto EM (2003) Animal models of copper-associated liver disease. Comp Hepatol 2:5–16

Gaither LA, Eide DJ (2000) Functional characterization of the human hZIP2 zinc transporter. J Biol Chem 275:5560–5564

Gaither LA, Eide DJ (2001) Eukaryotic zinc transporters and their regulation. Biometals 14:251–270
Ganz T, Nemeth E (2008) The Hepcidin-Ferroportin system as a therapeutic target in anemias and iron overload disorders. Hematol Am Soc Hematol Educ Progr 2011:538–542
Ganz T, Nemeth E (2011) Hepcidin and disorders of iron metabolism. Annu Rev Med 62:347–360
Ganz T, Nemeth E (2012) Hepcidin and iron homeostasis. Biochim Biophys Acta. doi:10.1016/j.bbamcr.2012.01.014
Garrick MD, Nunez MT, Olivares M, Harris ED (2003) Parallels and contrasts between iron and copper metabolism. Biometals 16:1–8
Gruenheid S, Cannone-Hergeaux F, Gauthier S, Hackam DJ, Grinstein S, Gros P (1999) The iron transport protein NRAMP2 is an integral membrane glycoprotein that colocalizes with transferrin in recycling endosomes. J Exp Med 189:831–841
Gunshin H, Mackenzie B, Berger UV, Gunshin Y, Romero MF, Boron WF, Nussberger S, Gollan JL, Hediger MA (1997) Cloning and characterization of a mammalian proton-coupled metal-Ion transporter. Nature 388:482–488
Hamlett NV, Landale EC, Davis BH, Summers AO (1992) Roles of the *Tn21*, *merT*, *merP*, and *merC* gene products in mercury resistance and mercury binding. J Bacteriol 174:6377–6485
Hantke K (2001) Bacterial zinc transporters and regulators. Biometals 14:239–249
Harris ED (2000) Cellular copper transport and metabolism. Annu Rev Nutr 20:291–310
Harris ZL, Klomb LW, Gitlin JD (1998) Aceruloplasminemia: an inherited neurodegenerative disease with impairment of iron homeostasis. Am J Clin Nutr 96:10812–10817
Hilton M, Spenser DC, Ross P, Ramsey A, McArdle HJ (1995) Characterisation of the copper uptake mechanism and isolation of ceruloplasmin receptor/copper transporter in human placental vesicles. Biochim Biophys Acta 1245:153–160
Holmberg CG, Laurell CB (1948) Investigations in serum copper II isolation of the copper-containing protein and a description of some of its properties. Acta Chem Scand 2:550–556
Iacopetta BJ, Rothenberger S, Kuhn LC (1988) A role for the cytoplasmic domain in transferrin receptor sorting and coated pit formation during endocytosis. Cell 54:485–489
Kambe T, Yamaguchi-Iwai Y, Sasaki R, Nagao M (2004) Overview of mammalian zinc transporters. Cell Mol Life Sci 61:49–68
Kawabata H, Yang R, Hirama T, Vuong PT, Kawano S, Gombart AF, Koeffler HP (1999) Molecular cloning of transferrin receptor 2. A new member of the transferrin receptor-like family. J Biol Chem 274:20826–20832
Knight SAB, Labbé S, Kwon LF, Kosman DJ, Thiele DJ (1996) A widespread transposable element marks expression of a yeast copper transport gene. Genes Dev 10:1917–1929
Kodama H (1993) Recent developments in Menkes' disease. J Inherit Metab Dis 16:791–799
Kohgo Y, Torimoto Y, Kato J (2002) Transferrin receptor in tissue and serum: updated clinical significance of soluble receptor. Int J Hematol 76:213–218
Korshunova YO, Eide D, Clark WG, Guerinot ML, Pakrasi HB (1999) The IRT1 protein from *Arabidopsis thaliana* is a metal transporter with a broad substrate range. Plant Mol Biol 40:37–44
Kuo YM, Gitschier J, Packman S (1997) Developmental expression of the mouse mottled and toxic milk genes suggests distinct functions for the Menkes and Wilson disease copper transporters. Hum Mol Genet 6:1043–1049
Lasat MM, Pence NS, Garvin DF, Ebbs SD, Kochian LV (2000) Molecular physiology of zinc transport in the Zn hyperaccumulator *Thlapsi caerulescens*. J Exp Biol 51:71–79
Lee PL, Gelbart T, West C, Halloran C, Beutler E (1998) The human Nramp2 gene characterization of the gene structure, alternative splicing, promoter region and polymorphisms. Blood Cells Mol Dis 24:199–215
Lieu PT, Heiskala M, Peterson PA, Yang Y (2001) The roles of iron in health and disease. Mol Aspects Med 22:1–87
Lin SJ, Pufahl RA, Dancis A, O'Halloran TV, Culotta VC (1997) A role for the *Saccharomyces cerevisiae ATX1* gene in copper trafficking and iron transport. J Biol Chem 272:9215–9220
Lioumi M, Ferguson CA, Sharpte PT, Freeman T, Marenholz I, Mischke D, Heizmann C, Ragoussis J (1999) Isolation and characterization of human and mouse ZIRTL, a member of the IRT1 family of transporters, mapping within the epidermal differentiation complex. Genomics 62:272–280
Majuri R, Grasbeck R (1987) A rosette receptor assay with haem-microbeads. Demonstration of a haem receptor on K562 cells. Eur J Haematol 38:21–25
Mantzoros CS, Prasad AS, Beck FWJ, Grabowski S, Kaplan J, Adair C, Brewer GJ (1998) Zinc may regulate serum leptin concentrations in humans. J Am Coll Nutr 17:270–275
Marquès L, Oomen JFJ (2011) On the way to unravel zinc hyperaccumulation in plants: a mini review. Metallomics 3:1265–1270
Martins LJ, Jensen TL, Simons JR, Keller GL, Winge DR (1998) Metalloregulation of *FRE*1 and *FRE*2 homologs in Saccharomyces cerevisiae. J Biol Chem 273:23716–23721
Masuoka J, Saltman P (1994) Zinc(II) and copper(II) binding to serum albumin. J Biol Chem 269:25557–25561
Matsuura W, Yamazaki T, Yamaguchi-Iwai Y, Masuda S, Nagao M, Andrews GK, Kambe T (2009) SLC39A9 (ZIP9) regulates zinc homeostasis in the secretory pathway: characterization of the ZIP subfamily I protein in vertebrate cells. Biosci Biotechnol Biochem 73:1142–1148
McClelland A, Kuehn LC, Ruddle FH (1984) The human transferrin receptor gene: genomic organization, and the complete primary structure of the receptor deduced from a cDNA sequence. Cell 39:267–274
McKie AT, Marciani P, Rolfs A, Brennan K, Wehr K, Barrow D, Miret S, Bomford A, Peters TJ, Farzaneh F, Hediger MA, Hentze MW, Simpson RJ (2000) A novel duodenal iron-regulated transporter, IREG1, implicated in the basolateral transfer of iron to the circulation. Mol Cell 5:299–309
Meguro Y, Kodama H, Abe T, Kobayashi S, Kodama Y, Nishimura M (1991) Changes of copper level and cytochrome *c* oxidase activity in the macular mouse with age. Brain Dev 13:184–186
Mercer JFB (2001) The molecular basis of copper-transport diseases. Trends Mol Med 7:64–69
Moos T, Trinder D, Morgan EH (2000) Cellular distribution of ferric iron, ferritin, transferrin and divalent metal transporter 1 DMT1 in substantia nigra and basal ganglia of normal and beta 2-microglobulin deficient mouse brain. Cell Mol Biol 46:549–561
Nies DH, Silver S (1995) Ion efflux systems involved in bacterial metal resistance. J Ind Microbiol 14:186–199
O'Halloran TV, Culotta CV (2000) Metallochaperones: an intracellular shuttle service for metal ions. J Biol Chem 275:25057–25060
Palmiter RD, Cole TB, Findley SD (1996) ZnT-2, a mammalian protein that confers resistance to zinc by facilitating vesicular sequestration. EMBO J 15:1784–1791
Pietrangelo A, Rocchi E, Casalgrandi G, Rigo G, Ferrari A, Perini M, Ventura E, Cairo G (1992) Regulation of transferring, transferrin receptor, and ferritin genes in human duodenum. Gastroenterology 102:802–809
Prohaska JR (2008) Role of copper transporters in copper homeostasis. Am J Clin Nutr 88:826S–829S
Raskin I (1995) Plant genetic engineering may help with environmental cleanup. Proc Natl Acad Sci U S A 93:3164–3166
Robinson NJ, Winge DR (2010) Copper metallochaperones. Annu Rev Biochem 79:537–562
Rolfs A, Hediger MA (1999) Metal ion transporters in mammals: structure, function and pathological implications. J Physiol 518:1–12

Rothenberg BE, Voland JR (1996) Beta2 knockout mice develop parenchymal iron overload: a putative role for class I genes of the major histocompatibility complex in iron metabolism. Proc Natl Acad Sci U S A 93:1529–1534

Rutledge EA, Enns CA (1996) Cleavage of the transferrin receptor is influenced by the composition of the O-linked carbohydrate at position 104. J Cell Physiol 168:284–293

Sacher A, Cohen A, Nelson N (2001) Properties of the mammalian and yeast metal-ion transporters DCT1 and Smf1p expressed in *Xenopus laevis* oocytes. J Exp Biol 204:1053–1061

Solioz M, Odermatt A, Krapf R (1994) Copper pumping ATPases: common concepts in bacteria and man. FEBS Lett 346:44–47

Suhy D, O'Halloran TV (1995) Metal responsive gene regulation and the zinc metalloregulatory model. In: Siegel H (ed) Metal ions in biological systems, vol 32. Marcel Dekker, New York, pp 557–558

Suzuki M, Gitlin JD (1999) Intracellular localisation of the Menkes' and Wilson's disease proteins and their role in intracellular copper transport. Pediatr Int 41:436–442

Tacchini L, Bianchi L, Bernelli-Zazzera A, Cairo G (1999) Transferrin receptor induction by hypoxia: HIF-1-mediated transcriptional activation and cell-specific post-transcriptional regulation. J Biol Chem 274:24142–24146

Takeda A, Devenyi A, Connor JR (1998) Evidence for non-transferrin-mediated uptake and release of iron and manganese in glial cell cultures from hypotransferrinemic mice. J Neurosci Res 51:454–462

Tandy S, Williams M, Leggett A, Lopez-Jimenez M, Dedes M, Ramesh B, Srai SK, Sharp P (2000) Nramp2 expression is associated with pH-dependent iron uptake by the brain: effects of altered iron status. J Biol Chem 275:1023–1029

Tapiero H, Townsend DM, Tew KD (2003) Trace elements in human physiology and pathology: copper. Biomed Pharmacother 57:386–398

Testa U, Petrini M, Quaranta MT, Pelosi-Testa E, Mastroberardino G, Camagna A, Boccoli G, Sargiacomo M, Isacchi G, Cozzi A, Arosio P, Peschle C (1989) Iron up-modulates the expression of TfRs during monocyte–macrophage maturation. J Biol Chem 264:13181–13197

Theil E (2011) Iron homeostasis and nutritional iron deficiency. J Nutr 141:724S–728S

Tong KK, McArdle HJ (1995) Copper uptake by cultured trophoblast cells isolated from human term placenta. Biochim Biophys Acta 1269:233–236

Tsuji Y, Moran E, Torti SV, Torti FM (1999) Transcriptional regulation of the mouse ferritin H gene: involvement of p300/CBP adaptor proteins in FER-1 enhancer activity. J Biol Chem 274:7501–7507

Umbreit JN, Conrad ME, Moore EG, Latour LF (1998) Iron absorption and cellular transport: the mobilferrin/paraferritin paradigm. Semin Hematol 35:13–26

Valentine RA, Jackson KA, Christie GR, Mathers JC, Taylor PM, Ford D (2007) ZnT5 variant B is a bidirectional zinc transporter and mediates zinc uptake in human intestinal caco-2 cells. J Biol Chem 282:14389–14939

van den Berghe PVE, Klomp WJ (2010) New developments in the regulation of intestinal copper absorption. Nutr Rev 67:658–672

Vulpe CD, Kuo Y-M, Murphy TL, Cowley L, Askwith C, Libina N, Gischier J, Anderson GJ (1999) Hephaestin, a ceruloplasmin homolog implicated in intestinal iron transport, is defective in the sla mouse. Nat Genet 21:195–199

Wang X, Zhou B (2010) Dietary zinc absorption: a play of zips and ZnTs in the Gut. IUBMB Life 62:176–182

Wood RJ, Han O (1998) Recently identified molecular aspects of intestinal iron absorption. J Nutr 128:17648–17654

Yanagimoto C, Harada M, Kunemura H, Abe M, Koga H, Sakata M, Kawaguchi T, Terada K, Hanada S, Taniguchi E, Ninomiya H, Ueno T, Sugiyama T, Sata M (2011) Copper incorporation into ceruloplasmin is regulated by niemann-pick C1 protein. Hepatol Res 41:484–491

Zhang P, Allen JC (1995) A novel dialysis procedure measuring free Zn^{2+} in bovine milk and plasma. J Nutr 125:1904–1910

Zhao H, Eide D (1996) The yeast ZRT1 gene encodes the zinc transporter of a high affinity uptake system induced by zinc limitation. Proc Natl Acad Sci U S A 93:2454–2458

Biological Functions of the Elements

7

Ulf Lindh

Contents

7.1 Essentiality of Elements

Much discussion has taken place with regard to how to define the essentiality of elements, particularly trace elements. The earliest definition was actually borrowed from protein chemistry. In this definition, an element is essential if:

- It is present in living tissues at a relatively constant concentration.
- It provokes similar structural and physiological anomalies in several species when removed from the organism.
- These anomalies are prevented or cured by supplementation of the element.

The current definition, suggested by an expert consultation of the World Health Organization/Food and Agricultural Organization/International Atomic Energy Agency (Mertz 1998), states:

> An element is considered essential to an organism when reduction of its exposure below certain limit results consistently in a reduction in a physiologically important function, or when the element is an integral part of an organic structure performing a vital function in the organism.

The concept of essentiality has the practical consequence that it is necessary to supply an organism with adequate amounts of the concerned elements. An immediate question raised by this consequence is how much is adequate. For most elements, ranges of safe and adequate intakes have been defined. In some cases, however, there is considerable uncertainty regarding the limits of such ranges. Adequate intakes do vary substantially among elements, in both amount and width of the range. In very general terms, the range may be visualized as shown in Fig. 7.1. For a detailed discussion of deficiencies and toxicities, the reader is referred to Chap. 9.

7.2 Major, Minor, and Trace Elements in Biology

Eleven elements seem to be consistently abundant in biological systems: hydrogen, oxygen, carbon, nitrogen, sodium, potassium, calcium, magnesium, phosphorus, sulfur, and chlorine. In

U. Lindh (✉)
Biomedical Radiation Sciences, Rudbeck Laboratory, Research in Metal Biology, Uppsala University, Uppsala, Sweden
e-mail: Ulf.Lindh@ibg.uu.se

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_7,

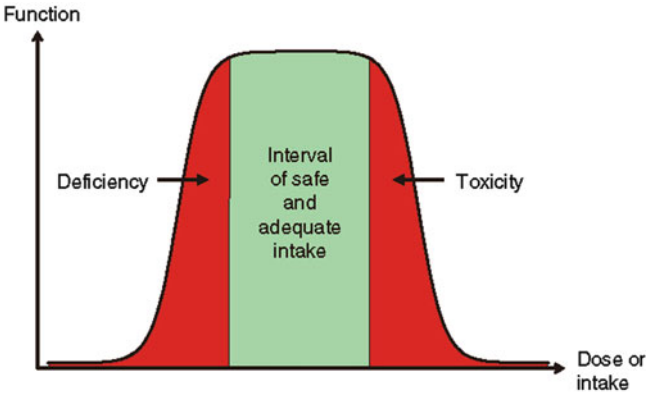


Fig. 7.1 Dose–response of essential trace elements

Table 7.1 Abundance by mass of major and minor elements in the human body

Element	Mass percent	Element	Mass percent
Oxygen	65.0	Magnesium	0.50
Carbon	18.0	Potassium	0.34
Hydrogen	10.0	Sulfur	0.26
Nitrogen	3.0	Sodium	0.14
Calcium	1.4	Chlorine	0.14
Phosphorus	1.0	–	–

humans, these elements comprise 99.9% of the atoms. Usually, these elements are divided into two groups of major and minor elements. The major elements – hydrogen, oxygen, carbon, and nitrogen – make up 99% of the atoms, or just over 96% of the body mass. Sodium, potassium, calcium, magnesium, sulfur, and chlorine comprise 3.78% of the body mass (Table 7.1).

The entire group of noble gases is excluded from consideration because their chemical properties make them unlikely to fulfill any biological function. The remaining elements are considered to be trace elements. There are 90 naturally occurring elements in the periodic table; thus, 73 are trace elements. Of these 73, 18 are considered to be essential or possibly essential trace elements: lithium, vanadium, chromium, manganese, iron, cobalt, nickel, copper, zinc, tungsten, molybdenum silicon, selenium, fluorine, iodine, arsenic, bromine, and tin. Their abundance in the human body is reflected in Table 7.2.

Several problems are associated with proving the essentiality of trace elements. Experiments are based upon the general acceptance that, if an essential trace element is completely withdrawn from the diet of experimental animals, signs and symptoms of a deficiency should occur, such as growth retardation and loss of hair. When a state of deficiency has been established, supplementation of the trace element should alleviate these symptoms and reverse the

Table 7.2 Abundance of certain trace elements in the human body by mass ($\mu g\ g^{-1}$)

Element	Mass fraction	Element	Mass fraction
Arsenic	0.26	Manganese	0.17
Bromine	2.9	Molybdenum	0.08[a]
Cobalt	0.021	Nickel	0.14
Chromium	0.094	Selenium	0.11
Copper	1	Silicon	260
Fluorine	37	Tin	0.24
Iron	60	Tungsten	0.008[a]
Iodine	0.19	Vanadium	0.11[a]
Lithium	0.009	Zinc	33

[a]Estimated from Iyengar et al. (1978) and Li (2000)

deficiency state. The first basic problem is that it is not possible to completely eliminate every bit of an element in food. Even if this were possible, the analytical techniques are inadequate due to their limits of detection. A second problem is that when essentiality is being evaluated, there is usually no well-grounded hypothesis for a possible biological function. Withdrawal of one essential trace element from the diet may result in altered uptake patterns for other trace elements, which makes results ambiguous. Most results have been obtained on plants and rodents. Veterinary medicine has contributed with information on the essentiality of elements in domestic animals. When it comes to humans, however, our knowledge of essential trace elements is less advanced for obvious reasons.

The essentiality of 12 of the trace elements in Table 7.2 is generally agreed upon, although perhaps not for all biological species. Without exhausting the available data on vanadium biology, a conspicuous property of ascidians must be pointed out. The blood cells of some ascidians accumulate vanadium at a degree of 10^7 as compared to the sea concentration. *Ascidia gemmata* has been shown to have the highest vanadium concentration – 350 m*M* – which corresponds to about 1.7% (Michibata et al. 2002).

7.3 Brief Description of the Function of Major Elements

Because the human body is roughly 71% water, it is not surprising that the quantities of hydrogen and oxygen are so substantial. Water makes up more than two-thirds of the weight of the human body, and without water a human would die in just a few days. All the cell and organ functions depend on water for functioning. It serves as a lubricant and forms the base for saliva and the fluids that surround the joints. Water regulates body temperature, as cooling and

heating occur through perspiration. Water helps to alleviate constipation by moving food through the intestinal tract, thereby eliminating waste. Water also contributes to the high contents of oxygen in the body.

We should also take into consideration the extraordinary chemical properties of water making it very suitable to being the solvent of life. Hydrogen bonds are the main feature that explains why water is the appropriate solvent. The current definition of the hydrogen bond by the International Union of Applied and Pure Chemistry (Arunan et al. 2011) is "The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation". Intramolecular hydrogen bonding is partly responsible for the secondary, tertiary, and quaternary structures of proteins and nucleic acids. It also plays an important role in the structure of polymers, both synthetic and natural.

7.3.1 A Few Important Points About Hydrogen

Hydrogen is a very special element in biology. It appears in three states, H^+ (cation), H– (a covalently bound state), and H^- (anion). It is a very strong acid as a proton (H^+). In the H– form, it takes part in stable non-metal bonds such as C–H and N–H. Even in the presence of dioxygen, these bonds are kinetically stable. Hydrogen can also be transferred from a non-metal, not just as H^+ or H– but also as H^- (Fraústo da Silva and Williams 2001). This makes it possible for hydrogen atoms to take part in one- or two-electron processes. Many biological redox reactions are based upon this property.

In the periodic table hydrogen is placed among the alkali metals. This may be surprising, however, at very low temperatures and very high pressures hydrogen acts like a metal, being more electrically and thermally conductive than silver. Other facts that propose hydrogen as a metal are fairly simple. All alkali metals react violently with oxygen, forming an oxide, or rust.

7.3.2 Carbon: The Backbone of Organic Chemistry and Biochemistry

In the chemical sense, use of the term "organic" means that carbon is involved; thus, an organic compound contains carbon atoms, and organic chemistry is the chemistry of carbon compounds. There are a few exceptions, however, as oxides, carbonates, and cyanides are considered inorganic compounds. Biochemistry, the chemistry of life, is a special branch of organic chemistry. Slightly contradictory is the fact that biochemistry involves both organic and inorganic compounds. To be more precise, it is appreciated in this field of study that the large molecules found in cells all contain carbon but many of the small molecules may be inorganic. A carbon atom is capable of combining with up to four other atoms, but in some cases a carbon atom combines with fewer than that. These bonds are covalent. Not only can a carbon atom form covalent bonds with four other atoms, but it can also combine with other carbon atoms; thus, carbon atoms can form chains and rings onto which other atoms can be attached. The atomic number of carbon is six; it has two electrons in the K shell and four in the L shell. Carbon must gain or lose four electrons to be ionized, but this process is difficult so instead it shares electrons to fill its L shell. In summary, carbon turns out to be a very versatile atom. The organic compounds in biology are represented by carbohydrates, lipids, proteins, and nucleic acids.

7.3.2.1 Carbohydrates

Carbon, hydrogen, and oxygen comprise most of the carbohydrates. For each carbon and oxygen there are two hydrogen atoms (i.e., CH_2O). Carbohydrates perform a series of important functions in biology, such as short-term energy storage (e.g., monosaccharides), long-term energy storage (e.g., starches and glycogen), and structural support (e.g., cellulose found in all plant cell walls), as well as important components of DNA and RNA.

7.3.2.2 Lipids

Lipid molecules are insoluble in polar solvents such as water. They dissolve in nonpolar solvents, and they are nonpolar. Lipids work as energy storage molecules, as insulation and protection for internal organs, as lubricants, and as hormones. Additionally, phospholipids are the major structural elements of membranes that are composed of a bilayer of phospholipids.

7.3.2.3 Proteins

Amino acids can be combined to form peptides, in which case the order of the amino acids is significant. The amino acids consist of an amino (−NH) group and a carboxylic acid ($-CO_2H$) group bonded to a central carbon atom. When the combination of amino acids exceeds more than about ten amino acids, the resulting peptide is referred to as a *polypeptide*. If the number of amino acids in a combination is more than 50, the molecule is referred to as a *protein*. The most important function of proteins is to maintain and drive the reactions of cells; however, additional properties include acting in supportive tissue like cartilage, and they are involved in muscle movement. Many proteins are enzymes, and their association with trace elements is described later in this chapter.

7.3.2.4 Nucleic Acids

Ribose, which is a monosaccharide, interacts with nitrogenous bases to form nucleosides such as adenosine. When nucleosides exist in cells mainly in the form of esters with phosphoric acid they are called *nucleotides*. If the sugar is ribose, the molecule is called *ribonucleotide*. The two nucleic acids, DNA and RNA, are polymers of nucleotides. In DNA, the bases are adenine, cytosine, guanine, and thymine; in RNA they are the same, except that uracil takes the place of thymine. The nucleotides in both DNA and RNA are joined by covalent bonds between the phosphate of one nucleotide and the sugar of the next one. A chain of many nucleotides is a polymer forming a nucleic acid.

7.3.3 Oxygen: The Savior and Reactionist

Without oxygen, humans and other mammals would not survive. Structures with a high degree of organization require specific transport mechanisms, and it is well known that in mammals oxygen is transported in the blood by hemoglobin; however, other oxygen-transporting molecules, such as hemerythrin, are used by some marine invertebrates, and hemocyanin is found, for example, in snails. In muscles, the diffusion of oxygen is facilitated by myoglobin. Oxygen is intimately involved in the production of energy-rich molecules (adenosine triphosphate, or ATP) which takes place in the mitochondrial membrane. The reduction of dioxygen to water requires the transfer of four electrons onto dioxygen at the same time:

$$O_2 + 4e^- + 4H^+ \rightarrow 2H_2O$$

Such a process, however, is not chemically possible; instead, on the way to becoming water, dioxygen passes through stages of aggressive power: the superoxide anion ($O_2^{-\bullet}$), hydrogen peroxide (H_2O_2), and the hydroxyl radical ($OH^\bullet$). All three intermediates are extremely reactive. Even hydrogen peroxide is very reactive, although it is not a radical. Electron- and proton-transfer reactions are extremely rapid, so the reactive intermediates are generally kept in the enzyme. In summary, the reduction of dioxygen to water involves four one-electron transfers:

$$O_2 \rightarrow O_2^{-\bullet} \rightarrow H_2O_2 \rightarrow OH^\bullet \rightarrow H_2O$$

Nature has evolved protective functions to take care of leaking intermediates. The superoxide anion is metabolized by superoxide dismutase to hydrogen peroxide, which is metabolized to water by catalase or glutathione peroxidase without releasing the extremely reactive hydroxyl radical. This balance, however, is delicate, and attacks from metals may result in the release of radicals. The plethora of reactions and functions that could be described for oxygen are too voluminous to present in this chapter, and the reader is referred to standard biochemistry textbooks.

7.3.4 Nitrogen Fixation

The supply of nitrogen to biological systems comes from gaseous N_2, which is very abundant in the atmosphere (80%). The bond between the nitrogen atoms is very strong; hence dinitrogen is chemically unreactive. Similar to dioxygen, dinitrogen has to be reduced in order to be biologically available. Reduction to ammonia requires a very specific and sophisticated enzyme in a process known as *nitrogen fixation*.

In the biosphere, most nitrogen fixation is carried out by a few species of bacteria that synthesize the enzyme nitrogenase. This enzyme is found in the species *Rhizobium*, living in symbiosis with root nodules of many leguminous plants, such as beans, peas, alfalfa, and clover. Nitrogen fixation also takes place in free-living soil bacteria such as *Azobacter*, *Klebsiella*, and *Clostridium* and by cyanobacteria found in aquatic environments.

Two proteins comprise nitrogenase. One of them contains a [4Fe–4S] cluster, and the other has two oxidation–-reduction centers. Iron is involved in one of these centers, molybdenum in the other. The net reaction is

$$N_2 + 8H^+ + 8e^- + 16ATP \rightarrow 2NH_3 + H_2 + 16ADP + 16P_i$$

To obtain the reducing power and ATP required for this process, symbiotic nitrogen-fixing microorganisms rely on nutrients obtained through photosynthesis carried out by the plant with which they are associated.

7.4 Brief Description of the Functions of Minor Elements

The minor elements in biology are comprised of sodium, magnesium, phosphorus, sulfur, chlorine, potassium, and calcium. The biological functions of the minor elements are discussed in three groups: (1) sodium, potassium, and chlorine; (2) magnesium and phosphorus; and (3) calcium.

Table 7.3 Concentrations of free cations and anions of calcium, magnesium, potassium, and sodium in seawater, human blood serum, and human red cells (mmol/L)

Ion	Seawater	Serum	Red cells
Calcium	10.25	2.20–2.55	10^{-4}
Magnesium	53.60	0.76–1.10	2.5
Potassium	9.96	3.5–5.1	92
Sodium	471	136–146	11
Chloride	549	98–106	50
Bicarbonate	–	22–29	10?
Phosphate	2×10^{-3}	0.74–3.07	10?

7.4.1 Sodium, Potassium, and Chlorine: Interactions and Ion Properties

Life evolved from water. Do concentrations of abundant elements in sea- or freshwater reflect this evolution? Table 7.3 makes a crude comparison between seawater and blood serum. It shows that intracellular concentrations are in most cases much lower than extracellular, the exception being potassium, which is an abundant intracellular ion. To function properly, organisms must actively pump sodium and chloride out of cells and actively take in potassium. It is not quite clear why life must reject the most abundant anion (Cl^-) and the most abundant cation (Na^+), although this is probably due to a need for the cellular stability that comes from maintaining an osmotic balance. Maintaining the intracellular and extracellular balance requires active transport processes that require considerable energy. For details regarding uptake and transport processes, the reader is referred to Chap. 6.

7.4.1.1 Biological Functions of Sodium, Potassium, and Chloride

Circuits maintained by K^+, Na^+, and Cl^- generally control the following properties in all cells of all organisms:

- Osmotic pressure
- Membrane potentials
- Condensation of polyelectrolytes
- Required ionic strength for activity

The first two properties are quite easy to understand; however, condensation of polyelectrolytes may have to be clarified. Biopolymers are polyelectrolytes; DNA, for example, possesses a linear series of charges. Other polyelectrolytes such as fats show two-dimensional arrays, whereas proteins have curved surfaces not seldom almost spherical. The surfaces of polyelectrolytes are stabilized by the surrounding ionic environment mainly due to sodium, potassium, and chloride ions.

Absorption of glucose, or any molecule for that matter, entails transport from the intestinal lumen, across the epithelium and into blood. The transporter that carries glucose and galactose into the enterocyte is the sodium-dependent hexose transporter, known more formally as SGLUT-1. As the name indicates, this molecule transports both glucose and sodium into the cell and, in fact, will not transport either alone.

The essence of transport by the sodium-dependent hexose transporter involves a series of conformational changes caused by the binding and release of sodium and glucose. It can be summarized as follows:

- The transporter is initially oriented facing into the lumen, and at this point it is capable of binding sodium, but not glucose.
- Sodium binds, which induces a conformational change that opens the glucose-binding pocket.
- Glucose binds, and the transporter reorients in the membrane such that the pockets holding sodium and glucose are moved inside the cell.
- Sodium dissociates into the cytoplasm, which causes glucose binding to destabilize.
- Glucose dissociates into the cytoplasm, and the unloaded transporter reorients back to its original, outward-facing position.

Sodium is intimately involved in vitamin transport. For example, the uptake of biotin, a member of the vitamin B complex, is dependent on the sodium-dependent multivitamin transporter, or SMVT (Stanely et al. 2002). There are also indications that the transport in brain parenchyma of *N*-acetylaspartate (the second most abundant amino acid in the adult brain) is accomplished by a novel type of sodium-dependent carrier that is present only in glial cells (Sager et al. 1999).

Sodium is also closely connected to the transport of vitamin C. It has been shown that the sodium-dependent vitamin C transporter (SVCT) is responsible for an age-dependent decline of ascorbic acid contents in tissues (Michels et al. 2003). The transport of vitamin C into the brain relies heavily on the sodium-dependent ascorbic-acid transporter Slc23a1 (Sotiriou et al. 2002). A conspicuous finding of Handy et al. (2002) was that copper uptake across epithelia seems to be sodium dependent.

Experiments have shown the presence of an ouabain-insensitive, potassium-dependent *p*-nitrophenylphosphatase in rat atrial myocytes. This enzyme is suggested to be an isoform of an H-transporting, potassium-dependent adenosine triphosphatase (Zinchuk et al. 1997). The enzyme pyruvate kinase requires potassium for maximal activity (Kayne 1971); however, it was recently shown that in the absence of potassium dimethylsulfoxide induces active conformation of the enzyme (Ramirez-Silva et al. 2001).

There is an interesting coupling of sodium and chloride in the transport of the neurotransmitter gamma-aminobutyric acid (GABA). The removal of GABA from the extracellular space is performed by sodium- and chloride-dependent high-affinity plasma membrane transporters (Fletcher et al. 2002). It has been suggested that chloride is involved in the

regulation of proteolysis in the lysosome through cathepsin C. This enzyme is a tetrameric lysosomal dipeptidyl–peptide hydrolase that is activated by chloride ion (Cigic and Pain 1999). Alpha-Amylases have also been shown to be chloride dependent (D'Amico et al. 2000).

7.4.2 Magnesium and Phosphate: Close Connections

In contrast to many metal and non-metal ions, magnesium is rather homogeneously distributed in organisms with about the same intra- and extracellular concentration of 10^{-3} *M*. It is also unique among the biological cations due to its size, charge density, and structure in aqueous solution, as well as its aqueous chemistry. These properties make magnesium generally different from all other cations, monovalent or divalent. Mg^{2+} has the largest hydrated radius of any common cation; in contrast, its ionic radius (*i.e.*, minus waters of hydration) is among the smallest of any divalent cation (Kehres and Maguire 2002). Magnesium is an essential element, a fact that was first demonstrated 77 years ago (Leroy 1926); however, its general role in cellular function is poorly understood.

Although there does not seem to be any significant chemical gradient, the net electrochemical gradient for Mg^{2+} is markedly directed inward because of the negative membrane potential inside; consequently, mechanisms must be in place for maintaining a low intracellular free Mg^{2+} concentration and to regulate Mg^{2+} homeostasis. At least two transport processes have been discovered; one is sodium dependent and the other works even in the absence of sodium. The sodium-dependent exchanger operates with a $3Na^{+}_{in}$:$1Mg^{2+}_{out}$ ratio (Romani and Maguire 2002). Various monovalent and divalent cations may replace the Na^{+} when sodium is absent. In that case, the stoichiometry is 1:1 (Romani and Maguire 2002). The physiological significance of this mechanism is, however, unclear as low extracellular concentrations of sodium are unlikely.

7.4.2.1 Hormonal Regulation of Magnesium Homeostasis

In physiological terms, it is reasonable to envision that, because hormones stimulate Mg^{2+} extrusion, other hormones or agents must also operate in eukaryotic organisms to promote Mg^{2+} accumulation and the maintenance of Mg^{2+} homeostasis (Romani and Maguire 2002). The mechanism behind hormonal regulation seems to be hormones or agents acting on the production of cyclic AMP. Another possibility is that activating a protein kinase C pathway decreases cyclic AMP in many tissues. An exiting finding was identifying the protein family Mrs2p in mitochondria; this is the first molecularly identified metal ion channel protein in the inner mitochondrial membrane.

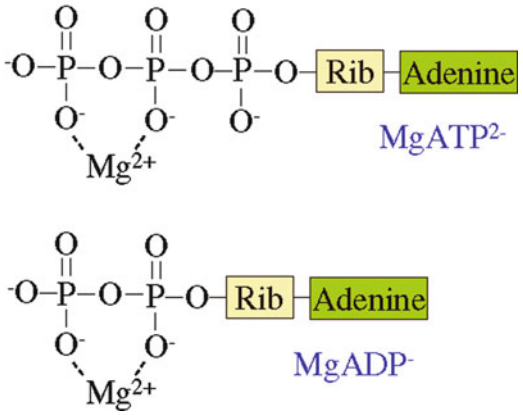


Fig. 7.2 Mg^{2+} complexes with ADP and ATP

7.4.2.2 Magnesium Binding and Magnesium Enzymes: Magnesium and Phosphates

The best known example of the very strong binding of magnesium ions is the Mg^{2+} in chlorophyll. A few other cases are ATP-synthetases in thylakoids and mitochondria and in the ATPases of muscles. Nucleoside diphosphates and triphosphate, both in aqueous solution and at the active site of enzymes, usually are present as complexes with magnesium (or sometimes manganese) ions. These cations coordinate with oxygen atoms of the phosphate groups and form six-membered rings with ADP or ATP (Fig. 7.2). A magnesium ion can form several different complexes with ATP. In solution, formation of the β,γ complex is favored.

Glycolysis is one of two important pathways in carbohydrate metabolism. The other is the pentose phosphate pathway. Both pathways provide energy to cells and both are also involved in the formation and degradation of other molecules such as amino acids and lipids. The glycolysis pathway has ten enzyme-catalyzed steps. In one of them, 2-phosphoglycerate is dehydrated to phosphoenolpyruvate in a reaction catalyzed by enolase. This enzyme requires Mg^{2+} for activity. Two magnesium ions participate in this reaction: A "conformational" ion binds to the hydroxyl group of the substrate and a "catalytic" ion participates in the dehydration reaction. Some other magnesium-dependent enzymes of the general metabolism are shown in Table 7.4 together with their biological functions.

7.4.2.3 Magnesium and Nucleic Acid Biochemistry

Monovalent metal ions such as Na^{+} and K^{+} act more or less as bulk electrolytes to stabilize surface charge. Divalent magnesium, on the other hand, interacts with nucleic acids with higher affinity. The role of Mg^{2+} is to neutralize negative charges from phosphates, either electrostatically or by forming hydrogen bonding networks from waters of solvation. Mg^{2+} may lower the pK_a of coordinated water, thereby facilitating phosphate ester hydrolysis. It is

Table 7.4 Some examples of magnesium-dependent enzymes and their functions

Enzymes	Function
Kinases	G-transfer reactions
Adenylate cyclase	cAMP formation from ATP
ATPases	Hydrolysis of ATP
Alkaline phosphatase	Splitting off phosphorus
Isocitrate lyase	Formation of succinate and glyoxylate in the citric acid cycle
Methyl aspartase	Glutamate receptor
Ribulose bisphosphate carboxylase	Carboxylation and oxygenation of ribulose bisphosphate
Myosin ATPase	Hydrolysis of ATP in muscles
Nucleases	Hydrolysis of phosphodiesters in nucleic acids
GTP-dependent enzymes	Restriction enzymes cleaving DNA, for example

becoming increasingly clear that many drug molecules interact with DNA in a specific and Mg^{2+}-dependent manner. An understanding of the latter may aid in the design of other novel DNA binding drugs (Sreedhara and Cowan 2002).

In addition to the enzymes noted in Table 7.4, it is worth mentioning a newly discovered class of enzymes requiring magnesium. Several categories of RNA have been found to catalyze reactions autonomously without protein or with only secondary protein assistance. Ribozymes are RNA molecules adopting three-dimensional structures that allow them to catalyze a variety of chemically important reactions, including but not restricted to phosphate ester hydrolysis, amide bond formation, and ligation (Sreedhara and Cowan 2002). Divalent magnesium seems to be involved in both the structure of RNA and the catalytic mechanism of ribozymes. Other metals such as manganese and calcium might take part in this structural and catalytic mechanism. Due to the intracellular abundance of magnesium, this would be the preferred metal ion. Cleavage of RNA is yet another exciting Mg^{2+}-dependent process of ribozymes.

7.4.2.4 Magnesium and Photosynthesis

As living organisms became abundant on the primitive Earth, their consumption of organic nutrients produced by geochemical processes outpaced production. Developing alternative sources of organic molecules that provided energy and the raw materials required for biosynthetic processes became critical for survival. The abundant CO_2 in the Earth's early atmosphere was a natural carbon source for organic synthesis; hence, photosynthesis became the pragmatic solution to the problem. Organisms capable of photosynthesis include certain bacteria, cyanobacteria (blue-green algae), algae, nonvascular plants, and vascular (higher) plants. Photosynthesis is also the source of the Earth's molecular oxygen. With the exception of anaerobic bacteria, all organisms capable of photosynthesis give off O_2 as an end product.

The net reaction of photosynthesis is:

$$CO_2 + H_2O \text{light} (CH_2O) + O_2$$

where (CH_2O) represents carbohydrate. The oxidation of water, a thermodynamically unfavorable reaction, is driven by solar energy. Electrons from this oxidation pass through electron-transport systems that resemble the mitochondrial electron-transport chain. Photosynthesis encompasses two major processes that can be described by two partial reactions:light

$$H_2O + ADP + P_i + NADP^+ \rightarrow O_2 + ATP + NADPH + H^+$$

$$CO_2 + ATP + NADPH + H^+ \rightarrow (CH_2O) + ADP + P_i + NADP^+$$

light

$$\text{Sum} : CO_2 + H_2O \rightarrow (CH_2O) + O_2$$

Photosynthesis takes place in chloroplasts. Chlorophyll is the most abundant pigment involved in the light harvesting, and the tetrapyrrole ring of chlorophylls, called *chlorine*, is similar to heme but contains Mg^{2+} chelated to the nitrogen atoms of the ring (Fig. 7.3). Chlorophylls a and b are the most abundant types. Magnesium ions are inserted into the chlorin ring through the action of a chelatase enzyme. Without the enzyme, magnesium could not combine with chlorins. Other metal ions, such as Zn^{2+}, Cu^{2+}, and Fe^{2+}, are more likely to combine with chlorins.

Mg^{2+} has an additional function in the chloroplast. When light shines on the thylakoids, pH drops to about 4.0. The protons force Mg^{2+} out of the thylakoid, and some chloride enters. The consequence is increased levels of Mg^{2+} in the stroma, where magnesium activates the carboxylase rubisco for incorporation of CO_2 into ribulose bisphosphate. This reaction is a step in the Calvin cycle of photosynthesis.

7.4.3 Calcium: Messenger and Support

Calcium has a plethora of functions in biology, and new roles are still being discovered. Calcium triggers new life at fertilization. It controls several developmental processes, and when cells have differentiated it functions to control such diverse cellular processes as metabolism, proliferation, secretion, contraction, learning, and memory (Jaiswal 2001). The best known function of calcium is its being an integral component of bone and teeth phosphates. Of the approximately 2.14 kg of calcium in the entire body, the skeleton contains 1.1 kg and the soft tissues 13.8 g. The bulk of the calcium, therefore, is contained in the skeleton as more or less crystalline calcium phosphate (called *bone apatite*), and it promotes the stability and rigidity of bones. Although only a minor portion is contained in soft tissues, the role of calcium here should not be overlooked. Control over metabolic processes probably involves a calcium-dependent step (Table 7.5).

7.4.3.1 Calcium Homeostasis

Only a minor proportion of skeletal calcium (about 1%) is rapidly exchangeable with plasma, although remodeling of bone results in the turnover of nearly 20% of skeletal calcium each year. Approximately 1% of body calcium is present in the extracellular fluid, where functions include the regulation of neuromuscular excitability and acting as a cofactor for clotting enzymes. The gradient of extracellular to intracellular free calcium is around 10,000:1, although cellular calcium is also found as insoluble complexes. Within cells, free calcium ions (Ca^{2+}) regulate the activity of various enzymes directly and also exert second-messenger hormonal functions by interaction with calcium-binding proteins such as troponin C and calmodulin.

The normal daily calcium exchange is represented in Fig. 7.4. Calcium in the gastrointestinal tract originates from the diet and also from secretions. Approximately half is absorbed, mainly in the upper small intestine (jejunum and

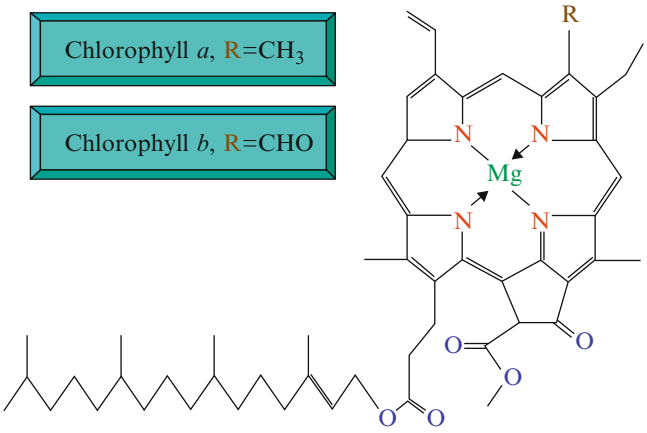


Fig. 7.3 Magnesium in chlorophyll

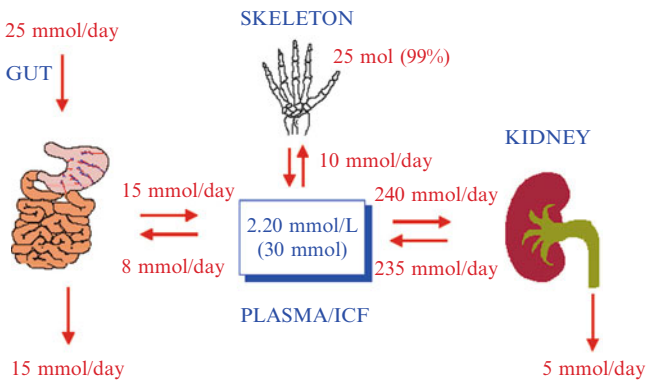


Fig. 7.4 Normal daily calcium exchange

Table 7.5 Calcium-controlled events in cells

Activity	Controlled events or systems
Photosynthesis	Dioxygen release
Oxidative phosphorylation	Dehydrogenases
Receptor responses	Nerve synapse
	IP3-linked reactions
Contractile devices	Muscle triggering (actomysin)
	Cell filament controls
Phosphorylation	Activation of kinases (e.g., in fertilization)
Metabolism	Numerous enzymes inside cells
Membrane/filament organization	Annexin-like proteins modulate tension
Cell division	S-100 proteins, immune system
Cell death (apoptosis)	Internal proteases
Hormone/transmitter release	Homeostasis
Binding to membranes	C-2 domains of enzymes
Cross-linking	Outside cells
Enzyme-activation	Outside cells; in membranes

Source: Adapted from Fraœsto da Silva and Williams (2001)

duodenum), by active transport. Up to 250 mmol calcium is filtered daily by the kidney. With the majority reabsorbed in the proximal tubule and loop of Henle, urinary excretion is normally 2.5–7.5 mmol L^{-1}, depending on uptake. The small amounts lost in sweat are usually insignificant, unless profuse sweating occurs for a prolonged period.

Parathyroid hormone (PTH) is a key regulatory hormone of calcium metabolism whose secretion is stimulated by low plasma Ca^{2+} concentrations and by low plasma magnesium concentrations. The physiological importance of regulation by magnesium is unclear, although magnesium depletion can cause hypocalcemia. The secretion of PTH is inhibited by increased Ca^{2+} levels. Within the parathyroid glands, PTH is transcribed as an 115-amino-acid polypeptide and processed to form an 84-amino-acid polypeptide, in which form it is stored prior to secretion. The biological activity of PTH is contained within a 32- to 34-amino-acid fragment of the molecule located at the N-terminal end. Following secretion, PTH is cleaved, mainly in the liver, to produce two fragments, one of which is an inactive (C-terminal) fragment. PTH is usually measured in blood by assays that depend on immunoreactivity rather than biological activity (immunoassays). In renal failure, the rate of metabolism of the C-terminal fragment is reduced as the kidney normally removes this. If assays for measuring PTH are based on the immunoreactivity of the C-terminal end of the molecule, levels may appear higher than would be apparent if biological activity were determined. This is important, because renal failure causes secondary hyperparathyroidism.

The main effect of PTH is to raise plasma Ca^{2+} concentrations through actions on bones, the kidney, and, indirectly, the gastrointestinal tract (Fig. 7.5). In the bones, PTH stimulates osteoclast activity, while in the kidney it increases the reabsorption of calcium and reduces the rate of transport of phosphate and bicarbonate. PTH also stimulates the hydroxylation of 25-hydroxycholecalciferol to form calcitriol, which then acts on the gut to increase calcium and phosphate absorption.

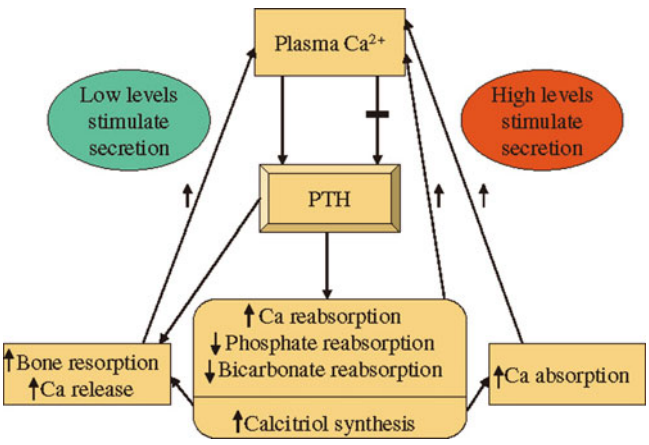


Fig. 7.5 Hormonal control of calcium metabolism

Vitamin D is converted to its biologically active form, 1,25-dihydroxycholecalciferol (calcitriol) by successive hydroxylations in the liver and kidney. Calcitriol stimulates intestinal absorption of calcium and phosphate by regulating the synthesis of a protein that transports calcium across the enterocyte. In addition, calcitriol is required for normal mineralization of bone, which is defective in deficiency states. Weakness of skeletal muscles also occurs in vitamin D deficiency, which responds to supplements. This suggests that vitamin D is important for normal skeletal function, although the basis of this is not understood.

7.4.3.2 Calcium Signaling

The development and improvement of analytical methods to qualitatively and quantitatively determine the presence of specific elements are of profound importance for understanding their biological roles. Calcium is no exception in this respect. The ability of cells to precisely regulate the cellular concentrations of free and bound calcium both in time and space is a feature that adds to the versatility of the calcium ion. Calcium plays an important role in cellular signaling and therefore has been classified as one of the major second-messenger molecules. Calcium signals control a vast array of cellular functions, ranging from short-term responses such as contraction and secretion to longer term control of transcription, cell division, and cell death (Berridge et al. 2000; Venkatchalam et al. 2002). Eukaryotic cells have developed specialized internal Ca^{2+} stores that are localized in the sarco- or endoplasmic reticulum (Carafoli and Klee 1999; Sorrentino and Rizzuto 2001), at least for higher vertebrate cells. These stores represent a significant contribution to intracellular signaling, as the stores can release Ca^{2+} either in conjunction with or independently of the Ca^{2+} entry pathways localized on the plasma membrane after stimulation by agonists (Sorrentino and Rizzuto 2001). Molecular studies have both identified the genes and characterized the proteins participating in intracellular Ca^{2+} homeostasis (Carafoli and Klee 1999; Sorrentino and Rizzuto 2001).

Intracellular Ca^{2+} stores have a specific molecular organization with regard to the way in which Ca^{2+} is released through specific intracellular Ca^{2+} release channels (ICRCs). Accumulation and storage, however, are mediated by dedicated Ca^{2+} pumps and Ca^{2+} binding proteins, respectively.

Ca^{2+} release from intracellular stores is mediated by channels encoded by a superfamily of genes, which includes three genes encoding channels capable of binding the vegetal ryanodine (RY) receptors and three genes encoding channels that bind inositol (1,4,5)-trisphosphate (IP3 receptors). These should be common for vertebrates

(Sorrentino and Rizzuto 2001). In most nonexcitable cells, the generation of receptor-induced cytosolic calcium signals is complex and involves two interdependent and closely coupled components: the rapid, transient release of calcium from stores in the endoplasmic reticulum and then the slow and sustained entry of extracellular calcium (Putney and McKay 1999; Berridge et al. 2000; Venkatchalam et al. 2002). Through the activation of phospholipase C subtypes, G-protein-coupled receptors and tyrosine-kinase-coupled receptors generate the second messenger inositol (1,4,5)-trisphosphate and diacylglycerol. Doing so functions as a chemical message that diffuses rapidly within the cytosol and interacts with inositol-trisphosphate receptors located on the endoplasmic reticulum lumen and generates the initial calcium signal phase (Sorrentino and Rizzuto 2001).

7.4.3.3 Structural Role of Calcium

Bone is not a static tissue, although it would be easy to think that when growth is finished bone does not change anymore; however, bone is a very dynamic tissue undergoing continuous destruction and remodeling. This process is important for the maintenance of bone volume and calcium homeostasis. The cells responsible for these dynamics are osteoblasts and osteoclasts, respectively. Osteoblasts produce bone matrix proteins, of which type I collagen is the most abundant extracellular bone protein. In addition, osteoblasts are critical to mineralization of the tissue (Aubin and Triffitt 2002; Katagiri and Takahasi 2002). Undifferentiated mesenchymal cells become osteoblasts, chondrocytes, myocytes, and adipocytes. Progenitor cells acquire specific phenotypes that are under the control of regulatory factors during the differentiation. Bone morphogenetic proteins (BMPs) play critical roles in the differentiation of the mesenchymal cells to osteoblasts (Katagiri and Takahasi 2002).

Osteoclasts are multinucleated cells responsible for bone resorption. They are differentiated from hematopoietic cells of monocyte/macrophage lineage under the control of bone microenvironments. The presence of ruffled borders and a clear zone is characteristic for these cells (Väänänen and Zhao 2001). Osteoblasts or bone marrow stromal cells have been shown to regulate osteoclast differentiation, thus providing a microenvironment similar to bone. The resorbing area under the ruffled border is acidified by vacuolar H^+–ATPase. A clear zone surrounds the ruffled border. This zone allows for the attachment of osteoclasts to the bone surface and maintains a microenvironment favorable for bone resorption. The recent discovery of the tumor necrosis factor (TNF) receptor–ligand family has clarified the molecular mechanism of osteoclast differentiation (Katagiri and Takahasi 2002).

7.4.4 Sulfur Bioinorganic Chemistry

About ten nonmetallic elements are essential for life. Of these, sulfur stands out due to its astounding chemical versatility. This versatility is reflected by several papers dealing with the biology of sulfur – for example, the chemistry of sulfane sulfur, disulfide as a transitory covalent bond and its significance in protein folding, the role of protein sulfenic acids in enzyme catalysis and redox regulation, and sulfur protonation in $[3Fe–4S]^{0,2-}$ clusters. In addition, it has been suggested that sulfur plays a role with regard to the μ_3 sulfide, where μ_3 indicates the number of metal ions connected to a multivalent ligand of Fe–S clusters in the catalysis of homolytic reactions, such as in pyruvate–formate lyase, anaerobic ribonucleotide reductase, 2,3-lysine amino mutase, biothin synthase, and the thioredoxin reductase (Beinert 2000). A close neighbor in the periodic table, phosphorus, works in biological systems by ionic reactions in the form of its oxo-anions. In contrast, sulfur in its most oxidized form, sulfate, is of limited use to higher organisms except for sulfation and detoxification reactions; rather, it is the chemical versatility of the lower oxidation states of sulfur that is vital for anabolic reactions (Beinert 2000).

7.4.4.1 Inorganic Systems

A conspicuous example of Fe–S clusters is cytoplasmic c-aconitase/iron regulatory protein (IRP). The aconitases lose one Fe readily on exposure to O_2, or even faster to O_2^-, with formation of the enzymatically inactive 3Fe form $[3Fe–4S]^+$, which can be reconstituted to form active $[4Fe–4S]^{2+}$ enzyme (Fig. 7.6a). In addition, c-aconitase alternates as aconitase and, after complete loss of its Fe–S cluster, as an IRP (Fig. 7.6b) that can bind to iron-responsive elements (IREs) in the untranslated regions of mRNAs and thereby regulate translation (Haile et al. 1992; Beinert et al. 1996; Beinert 2000). There is some spontaneous breakdown of aconitase; however, this process is probably too slow to be used in regulation but for subtle adjustments. Agents such as

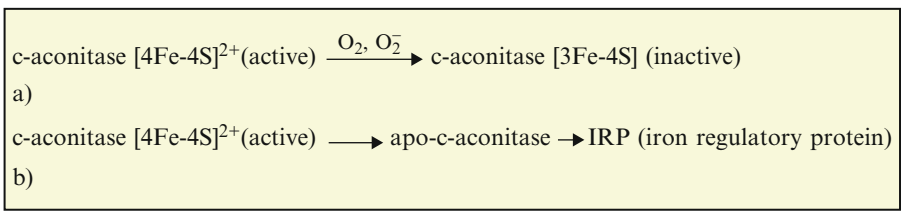


Fig. 7.6 Possibilities for aconitases

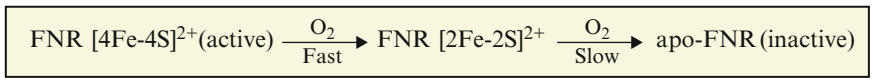


Fig. 7.7 Transition from anaerobic to aerobic metabolism in *E. coli*

Table 7.6 Concentrations of vanadium in the tissues of several ascidians compared with human serum (mmol/L)

Species	Tunic	Mantle	Branchial basket	Serum	Blood cells
Ascidia gemmata	ND	ND	ND	ND	347.2
A. ahodori	2.4	11.2	12.9	1.0	59.9
A. sydneiensis	0.06	0.7	1.4	0.05	12.8
Phallusia mammillata	0.03	0.9	2.9	ND	19.3
Ciona intestinalis	0.003	0.7	0.7	0.008	0.6
Homo sapiens	–	–	–	0.000003–0.000018	–

NO or H_2O_2 can speed up the process significantly. Transition of the organism from anaerobic to aerobic metabolism (*Escherichia coli*) is governed by the global transcription regulator fumarate nitrate reduction (FNR). FNR alternates between the active, O_2-sensing, holoform $[4Fe–4S]^{2+}$ and the inactive form $[2Fe–2S]^{2+}$ and, finally, apoform during O_2 sensing (Bates et al. 2000) by means of its very O_2-sensitive Fe–S cluster (Fig. 7.7). Obviously, in all these cases, there must be a continuous cycle of disassembly and reassembly or repair of clusters *in vivo*.

7.4.4.2 Organic Systems

If the processes in which Fe–S clusters are involved are considered primarily inorganic, it is interesting to look for sulfur processes without the involvement of these clusters. It appears that at least one process is devoid of Fe–S clusters—the biosynthesis of thiamine. This pathway is complicated and not yet fully understood (Begley et al. 1999). Thiamine synthesis involves at least six peptides (ThiF, S, G, H, I, J). Cystein is the original source of the sulfur, and the immediate donor, which completes the heterocycle synthesis, is thiocarboxylate formed at the carboxyl terminus of ThiS in an ATP-demanding reaction.

7.5 The Functional Value of Trace Elements

Trace elements are found in minute amounts in tissues and body fluids. Nevertheless, they play an extremely important role in biology. The paramount function is to be necessary for the structure and function of significant biomolecules, mainly enzymes. An intriguing property of many of the trace elements is that they are found among the transition elements, and most of them are also in the fourth row of the periodic table. Only boron, selenium, molybdenum, iodine, and tungsten are found at other sites.

7.5.1 Vanadium, Cobalt, and Nickel Are Not Used Extensively

These trace elements are treated more or less collectively without any specific chemical argument. They can be found in the transition metal area in the periodic table and are known to have or are suspected to have biological functions.

7.5.1.1 Vanadium

The rich aqueous chemistry of vanadium suggests that its metabolism is complex. Both the anion vanadate (VO^-_3) and the cation vanadyl (VO^{2+}) can complex with molecules of physiological significance. Vanadium species such as the oxyanion and oxycations VO_4^{2-}, VO^{2+}, and VO^+_2 are oxidizing agents capable of reacting by one-electron transfer or by two-electron steps when V acts as an O-atom donor. In addition, vanadium can form sulfur-containing anionic centers (e.g., VS_4^{3-}) and sulfur-containing cationic centers such as VS^{2+} and $VSSH^+$ (Fraústo da Silva and Williams 2001). Apart from the strikingly high concentrations of vanadium in some ascidians, for which there still is no clear biological function, vanadium concentrations in tissues and body fluids generally are low, as can be seen from Table 7.6.

Most of the vanadium in the vanadocytes of ascidians, which are known to be the signet ring cells, has been shown to be in the V(III) state, with a minor part occurring in the V (IV) state; consequently, reducing agents have to participate in the accumulation of vanadium in vanadocytes. Among the proposed agents are tunichromes. Earlier it was thought that the vanadium incorporated by ascidians was dissolved as ionic species or associated with low-molecular-weight substances rather than proteins. This is in contrast to other metals that generally bind to macromolecules such as proteins and are incorporated into the tissues of living organisms. Michibata et al. (2002), however, demonstrated the presence of a vanadium-associated protein in vanadocytes. This protein was estimated to associate with vanadium

in an approximate ratio of 1:16. The protein was comprised of three peptides of estimated molecular weights of 12.5, 15, and 16 kDa; however, the physiological roles of vanadium still remain to be elucidated.

In sea water, vanadium is dissolved in the 5+ oxidation state (HVO_4^{2-}, $H_2VO_4^-$). During the accumulation process V(V) is reduced to the 3+ oxidation state (V(III)) via the 4+ oxidation state (VO^{2+}). NADPH is a strong candidate to take part in the reduction from V(V) to V(IV) since the enzymes involved in the pentose phosphate pathway are expressed exclusively in the cytoplasm of vanadocytes (Ueki et al. 2000) and detailed studies have pointed to that this reaction occurs with the assistance of several chelating substances, vanabins. These are vanadium-binding protein capable of binding as many as 10–20V(IV) ions per molecule of protein and are believed to be important in V(V)-reduction, V(IV)-transport and V(IV)-storage processes in the cytoplasm of vanadocytes (Kawakami et al. 2009).

The milieu of the vanadocyte vacuole is maintained at an extremely low pH approach 1.9 or less and there is a tight relation between the vanadium concentration and protons. Recent data suggest that there is a vanadium transporter, Nramp, unique for the ascidians vanadocytes. V(V) seems to be readily reduced to V(IV) and V(IV) is stabilized by vanabins that act as vanadium reductase and vanadium chaperone in the cytoplasm. The Nramp should be a VO^{2+}/H^+ antiporter expressen on the vacuolar membrane and that proton electrochemical gradient generated by V-ATPase is the driving force for VO^{2+} transport from the cytoplasm into the vacuole (Ueki et al. 2011).

Other functions of vanadium are associated with various defense systems, such as peroxidase and catalase activity or haloperoxidases, as well as dinitrogen fixation. The peroxidase and catalase activity is exerted by the vanadium-containing substance amavadine (V(IV) *bis*-complex of *N*-hydroxyimino-di-a-propionate), the mechanism of which is illustrated in Fig. 7.8 (Matoso et al. 1998). Amavanadine is found in some *Amamita* toadstools. It seems that amavadine could be a kind ofprimitive protective substance able to use H_2O_2 for self-regeneration of damaged tissues or to defend against foreign pathogens and predators, decomposing it when not necessary (Fraústo da Silva and Williams 2001).

An intriguing and exciting property of vanadium is that it mimics insulin both *in vitro* and *in vivo*. Even more fascinating is that vanadium as sodium vanadate ($NaVO_3$) was used to treat patients with diabetes mellitus as early as 1899 (Lyonnet et al. 1899), before insulin was even discovered. The proposed mechanism by which vanadyl acts is on at least three sites (Sakurai 2002), as illustrated in Fig. 7.9. Vanadate behaves in a manner similar to that of phosphate; therefore, the effects of vanadium are thought to inhibit the production of protein phosphotyrosine phosphatase that follows stimulation of protein tyrosine phosphorylation. Vanadate also activates autophosphorylation of solubilized insulin receptors in a way similar to the insulin activation. Vanadate also stimulates the tyrosine kinase activity of the insulin receptor β-subunit. Both vanadate and vanadyl are effective in stimulating glucose metabolism in adipocytes. Vanadate restores the expression of the insulin-sensitive glucose transporter and induces the recruitment of the GLUT4 glucose transporter to the plasma membrane. The 3′,5′-cyclic adenosine monophosphate (cAMP)-mediated protein phosphorylation cascade in adipocytes is activated during diabetes or in the presence of adrenalin. Both glucose and vanadyl, which are incorporated in the adipocytes in response to vanadyl treatment, lead to the restored regulation of this cascade. Free fatty acid (FFA) release is thought to be inhibited by vanadyl. Vanadyl, thus, acts on at least three cell sites—phosphatidyl inositol 3-kinase, glucose transporter, and phosphodiesterase—to normalize both glucose and FFA levels in diabetes.

It has been demonstrated that vanadium compounds can be used in the treatment of diabetes mellitus, and vanadium also seems to be able to prevent the onset of diabetes. A novel hypothesis has been proposed in which nitric oxide (NO) production can be attributed to macrophages (Mϕ). The possible mechanism is shown schematically in Fig. 7.10 (Sakurai 2002). In the macrophages of normal animals treated with $VOSO_4$, incorporation of vanadium and responses to enhanced NO production are low. In the prediabetic phase of mice treated with streptozotocin (STZ), activated Mϕ exudes through pancreatic islets, and the NO produced concomitantly by activated Mϕ destroys normal islet β cells. The onset of diabetes by STZ administration is proposed to be based on the enhancement of the generation of superoxide anions ($\cdot O^-_2$) in β cells. NO reacts with $\cdot O^-_2$ to produce peroxynitrite ($ONOO^-$). One of the degradation products of peroxynitrite is a hydroxyl radical (·OH). The radicals $\cdot O^-_2$ and ·OH destroy the β cells. Treatment with $VOSO_4$ suppresses NO production and consequently formation of ·OH that damages the β cells.

7.5.1.2 Cobalt

A spontaneous association with cobalt biological functions is most likely to be with vitamin B_{12}, as cobalt is an integral cofactor in this vitamin. Although the biological functions of vitamins are slightly out of the scope of this book, a few comments are necessary here. Vitamin B_{12} is involved in four types of important reactions:

1. Reduction of ribose to deoxyribose
2. Rearrangement of diols and similar molecules
3. Rearrangement of malonyl to succinyl
4. Transfer of methyl groups

Worth mentioning is the enzyme ribonucleotide reductase that requires vitamin B_{12} for the reduction of

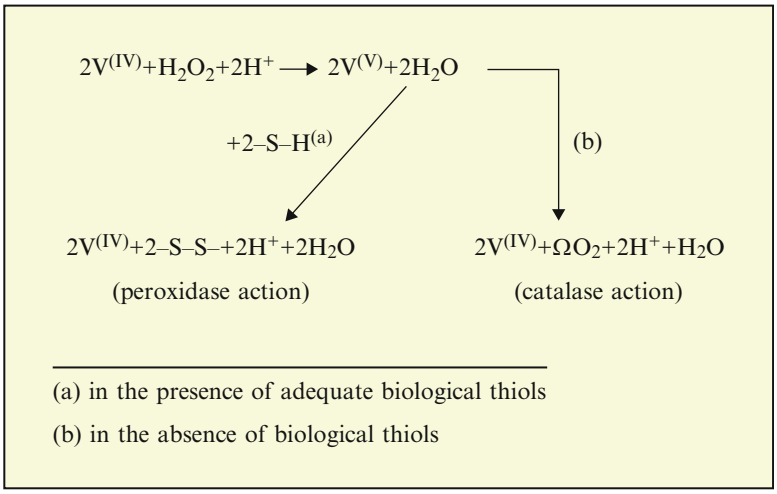


Fig. 7.8 Peroxidase and catalase activity of amavadine (Adapted from Matoso et al. 1998)

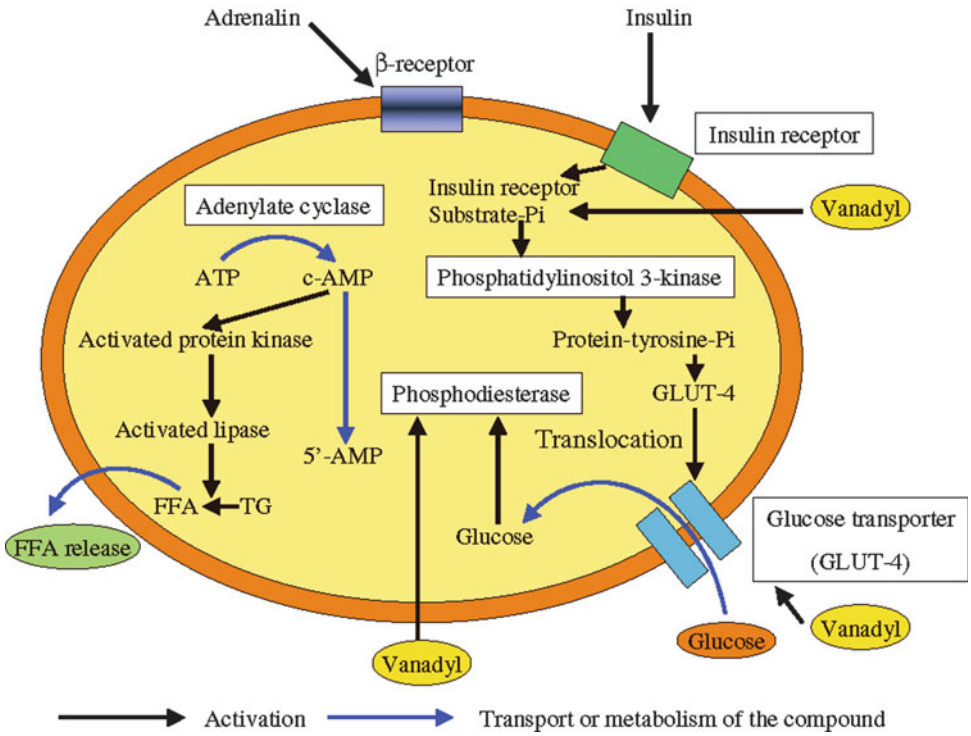


Fig. 7.9 Proposed mechanisms of vanadyl action (Adapted from Sakurai 2002)

ribonucleotides to the corresponding deoxyribonucleotides. Not only is this reaction dependent on the enzyme ribonucleotide reductase and vitamin B_{12}, but the reaction is also considered to proceed via the radical pathway. This constitutes an interesting example of how radicals can have important functions in biological processes that are not destructive in nature. In vitamin B_{12}, cobalt is associated with a corrin ring, a relative of porphyrin. Non-corrin cobalt is receiving increased interest, and ten non-corrin-cobalt-containing enzymes have been isolated (Kobayashi and Shimizu 1999) and characterized (Table 7.7).

Metionine aminopeptidase cleaves the N-terminal methionine from many newly translated polypeptide chains in both prokaryotes and eukaryotes. It is an important catalyst for N-terminal modification involved in functional regulation, intracellular targeting, and protein turnover. The *E. coli* methionine aminopeptidase is a monomering protein of 29 kDa and consisting of 263 residues that bind two Co^{2+} ions in its active

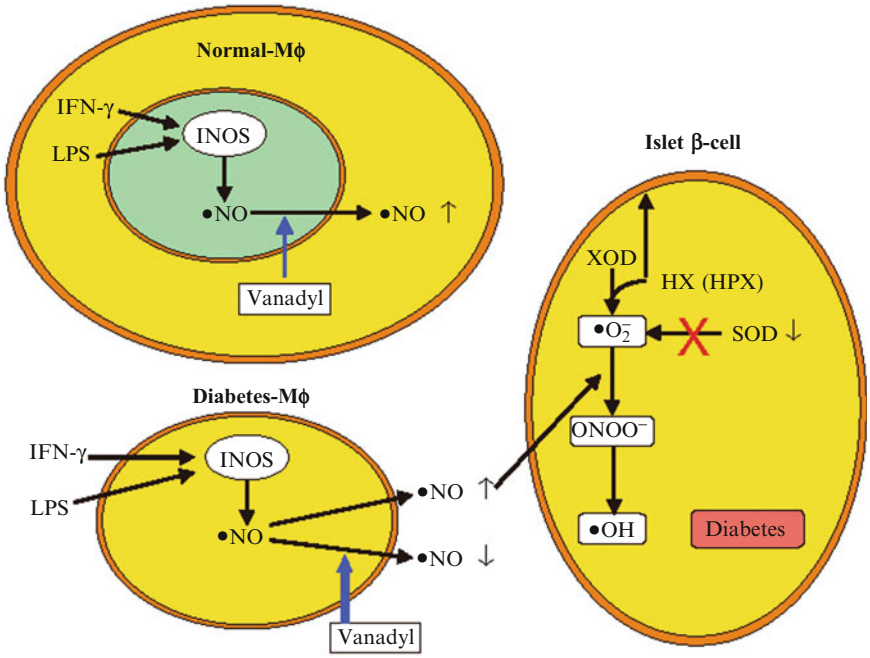


Fig. 7.10 A possible mechanism of vanadium prevention of diabetes (Adapted from Sakurai 2002)

Table 7.7 Cobalt-containing proteins

Enzyme or protein	Source	Cofactor content	Postulated role of cobalt
Methionine aminopeptidase	Animals, yeast, bacteria	2 Co per subunit	Hydrolysis
Prolidase	Archaea	1–2 Co per subunit	Hydrolysis
Nitrile hydratase	Actinomycetes and bacteria	1 Co in each α-subunit	H_2O activation, CN-triple-bond hydration and protein folding
Glucose isomerase	Actinomycetes	1 Co per 4 subunits	Isomerization
Cobalt transporter	Actinomycetes and yeast	–	Cobalt uptake
Methylmalonyl-CoA carboxytransferase	Bacteria	1 Co, 1 Zn per subunit	Carboxytranserfation
Aldehyde decarbonylase	Algae	1 Co-porphyrin per αβ-subunit	Decarbonylation for aldehyde
Lysine-2,3-aminomutase[a]	Bacteria	0.5–1 Co per subunit	Mutation
Bromoperoxidase	Bacteria	–0.35 Co per 2 subunits	Bromination
Cobalt-porphyrin-containing protein	Bacteria	1 Co-porphyrin per protein	Electron carrier

[a]Lysine-2,3-aminomutase also contains an iron–sulfur cluster, zinc, and PLP as cofactors
Source: Adapted from Kobayashi and Shimizu (1999)

site. The two subfamilies of cobalt-containing methionine aminopeptidase are a prokaryotic class (type I) and a human class (type II).

Prolidase (or praline dipeptidase) specifically cleaves Xaa–Pro dipeptides. In concert with other endopeptidases and exopeptidases, prolidase is thought to be involved in the terminal degradation of intracellular proteins and may also function in the recycling of proline (Ghosh et al. 1998).

Nitrile hydratase catalyzes hydration of nitriles to amides (see below) and is a key enzyme involved in the metabolism of toxic compounds. The presence of cobalt as an essential cofactor may be explained by the effective catalysis of CN-triple bond hydration as well as a requirement for the protein folding.

Glucose isomerase catalyzes the reversible isomerization of D-glucose to D-fructose. Although this enzyme requires divalent cations for the activity, its specific requirement depends on the source of the enzyme.

Most metals that play essential roles as cofactors in biological processes have to be actively incorporated into cells against concentration gradients. COT1 acts as a cobalt toxicity suppressor in *Saccharomyces cerevisiae* by sequestration or compartmentalization within the mitochondria of cobalt ions that cross the plasma membrane (Conklin et al. 1992). The cobalt transporter NhlF mediates cobalt incorporation into the cell in an energy-dependent manner. The sequence of this transporter shows eight putative hydrophobic membrane-spanning domains. NhLF contains nine histidine residues and two cysteine residues. One histidine (His306) and one cysteine (Cys301) residue are located on the fourth outside loop and may comprise a cobalt-binding

site for the initial fixation of the metal. The transmembrane segments could form a cobalt channel in which the four histidines might function as transient cobalt-binding ligands.

The biotin-containing enzyme methylmalonyl–CoA carboxytransferase (transcarboxylase) is a complex multisubunit enzyme that catalyzes the transfer of a carboxyl group from methylmalonyl–CoA to pyruvate to form propionyl–CoA and oxaloacetate.

Aldehyde decarboxylase converts a fatty aldehyde to hydrocarbon and carbon monoxide. This enzyme is responsible for a key step in the biosynthesis of hydrocarbon compounds. Lysine-2,3-aminomutase catalyzes the reversible isomerization of l-lysine to l-β-lysine, a reaction in which the hydrogen on the 3-pro-R position of lysine is transferred to the 2-pro-R position of β-lysine and the 2-amino group of lysine migrates to carbon-3 of β-lysine. The enzyme contains three cofactors: pyridoxal phosphate, Fe–S centers, and cobalt or zinc.

Bromoperoxidase catalyzes the formation of a carbon–bromine bond in the presence of peroxides. More studies, however, are needed to elucidate the function of bromoperoxidase. Other proteins contain cobalt–porphyrins in plants and in sulfate-reducing bacteria. The cobalt–porphyrins are not covalently bound to the proteins. The prosthetic groups of these cobalt-containing proteins are found to be cobalt isobacteriochlorins.

To maintain homeostasis, a set of systems has to work in close cooperation: uptake systems, efflux systems, metallochaperones and metal-storage proteins. In the case of cobalt, many of the details of homeostasis still remain unclear. In general, the genes encoding cobalt transporters are clustered with those encoding Co-dependent enzymes or enzymes involved in the coenzyme B12 biosynthesis. Several cobalt-specific metallochaperons have recently been discovered in the cobalamin biosynthesis pathway and in the maturation of Co-NHases, respectively. Bacteria have developed mechanisms to get rid of cobalt ion. Efflux systems that belong to the major facilitator superfamily permeases and the resistance nodulation pumps were found to act in cobalt export (Okamoto and Eltis 2011).

7.5.1.3 Nickel

Nickel-containing enzymes are involved in at least five metabolic processes, including the production and consumption of molecular hydrogen, hydrolysis of urea, reversible oxidation of carbon monoxide under anoxic conditions, methanogenesis, and detoxification of superoxide anion radicals. The active sites of the relevant enzymes harbor unique metallocenters, which are assembled by auxiliary proteins. Different ligand environments are involved in the coordination of nickel in the various metalloenzymes (Eitinger and Mandrand-Berthelot 2000). The reactions

a. Superoxide dismutase

$$2H^+ + 2O_2^{-\bullet} \rightarrow H_2O_2 + O_2$$

b. Urease

$$H_2N-CO-NH_2 + 2H_2O \rightarrow 2NH_3 + H_2CO_3$$

c. Hydrogenase

$$2H^+ + 2e^- \leftrightarrow H^+ + H^- \leftrightarrow H_2$$

d. Methyl-CoM reductase

$$CH_3-CoM + CoB-SH \rightarrow CH_4 + CoM-S-S-CoB$$

e. CO dehydrogenase

$$CO + H_2O \rightarrow 2H^+ + CO_2 + 2e^-$$

f. Acetyl-CoA synthase

$$CH_3-CFeSP + CoA-SH + CO \rightarrow CH_3-CO-SCoA + CFeSP$$

Fig. 7.11 Reactions catalyzed by nickel-dependent enzymes

catalyzed by nickel-dependent enzymes are summarized in Fig. 7.11.

Some transporters are potentially able to transport nickel; however, CorA, MgtA, and MgtB have affinities that are too low to be of physiological importance (Ragsdale 1998). Two different types of high-affinity nickel transporters have been identified. One is a multicomponent ATP-binding cassette (ABC) transporter system such as NikABCDE that uses ATP. The other is a one-component transporter such as NixA, UreH, HupN, and HoxN. NixA and HoxN are integral membrane proteins that have eight transmembrane-spanning helices and a sequence motif that is essential for function. HypB can sequester nickel and release it for incorporation into apoproteins when nickel becomes limiting (Ragsdale 1998).

A nickel–superoxide dismutase (Ni–SOD) has been isolated from *Streptomyces*. The Ni-SOD gene, *sodN*, has been found in cyanobacteria (blue-green algae), marine gammaproteobacteria and a marine eukaryote (Dupont et al. 2008). The protein is a homotetramer of four 13-kDa subunits with little sequence similarity to earlier known SODs. Nickel induces its expression, represses the FeZn–SOD, and is involved in maturation of a precursor polypeptide (Ragsdale 1998).

Urease plays a key role in the nitrogen metabolism of plants and microbes and acts as a virulence factor for some human and animal pathogens (Mobley et al. 1995). There are two nickel ions in the enzyme, and CO_2 is required to generate a carbamylated lysine bridge between the two nickel ions.

Hydrogenases are of different types. Some hydrogenases involve Ni, Fe, and Se; some Ni and Fe. In addition, some depend only on Fe. In fact, one hydrogenase has no association with metals at all (Hartmann et al. 1996).

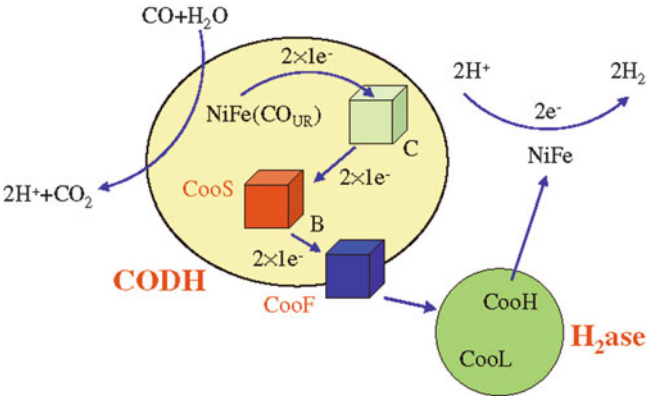


Fig. 7.12 CO-dehydrogenase oxidizes CO to CO_2 using a NiFe cluster and a nickel-dependent hydrogenase (H2ase) to produce hydrogen gas (Adapted from Watt and Ludden 1999)

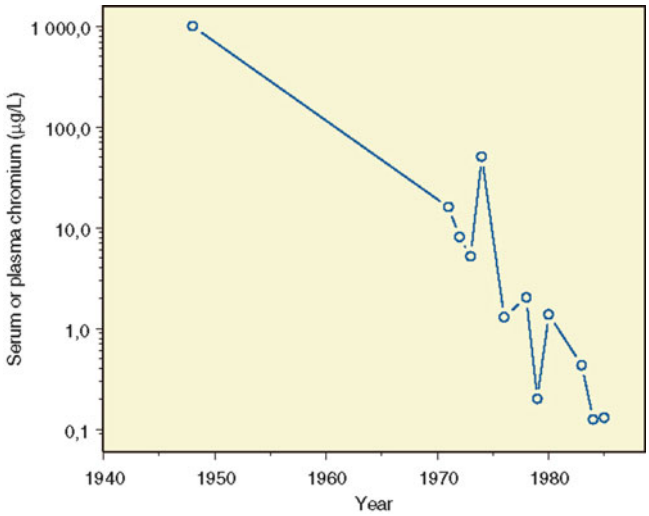


Fig. 7.13 ÒNormalÓ serum levels of chromium over 30–40 years

Methyl-coenzyme M reductase is a remarkable enzyme composed of six subunits forming a heterohexamer. Three unusual coenzymes are embedded in a long channel between the subunits. Of particular interest is the binding of coenzyme F430, a Ni–porphinoid that occurs exclusively in this enzyme. As substrates, the enzyme binds methyl-coenzyme M (methyl-thioethane sulfonate) and coenzyme B (7-thioheptanoyl threoninephosphate). Structural studies of this enzyme have been hampered by the low activity of the purified enzyme; however, substantial improvements in this respect have been seen in recent years (Finazzo et al. 2003).

Carbon monoxide (CO) dehydrogenase oxidizes CO to CO_2 in the half reaction shown in Fig. 7.12 using a Ni–Fe cluster. Another enzyme, a hydrogenase, also contains a Ni–Fe cluster and uses the electrons in the first half reaction to reduce protons to hydrogen gas: $2\ H^+ + 2e^- \times H_2(g)$ (Watt and Ludden 1999). Carbon monoxide dehydrogenases are bifunctional enzymes that perform reversible CO oxidation and also function to synthesize or degrade acetyl–CoA. The bifunctional characteristics of this enzyme led to changing its name to CO–dehydrogenase/acetyl–CoA synthase (Hausinger 1993). Both catalytic sites for the individual reactions require nickel for catalysis. Methanogenic bacteria convert acetate to methane and CO_2 and couple acetate degradation to ATP synthesis. This metabolic process uses the CO–dehydrogenase/acetyl–CoA synthase (Watt and Ludden 1999).

Nickel compounds are recognized as human carcinogens. It has been suggested that the molecular mechanism of the genotoxicity underlying the induction of carcinogenesis is the delivery of nickel into a cell from particulates outside the cell. Water-soluble nickel salts penetrate cells poorly. Following phagocytosis, the particles are contained in vacuoles that become highly acidified, and this greatly enhances the dissolution of soluble nickel from the particles. This increases the intracellular load of nickel, which can subsequently attack chromatin and particularly histones and produce effects associated with carcinogenic activity (Zoroddu et al. 2002).

The essentiality of nickel for plants now seems to be fairly well established (Yusuf et al. 2011). The toxicity of nickel in microorganisms is even further consolidated, although it was believed to be a problem to cells exposed to industrial pollution or as a result of naturally-occurring nickel-contaminated soils. This assumption, however, was overruled when a nickel defense system, RcnA, was characterized in the mesophile *Escherichia coli*. In addition, putative homologues of RcnA were detected in archaea and throughout bacteria (Macomber and Hausinger 2011).

7.5.2 Chromium, Molybdenum, and Tungsten

This triad of trace elements comprises group 6 in the periodic table; consequently, they should have some chemical properties in common. An intriguing property of these trace elements is their relationship with the non-metal elements of group 16; hence, the trace elements of group 6 often appear covalently bonded and not as simple metal ions.

7.5.2.1 Chromium

Chromium(III) was already established as an essential dietary component in the 1950s. Many studies have been conducted to elucidate the biological role of chromium but have had a low degree of success. A major problem is the low concentration in tissues and body fluids and the high risk of contamination of samples. It was not until the 1980s that reliable analytical methods were developed. Over the course of 30–40 years of analytical efforts (Fig. 7.13), the so-called normal serum level of chromium in humans has been

decreasing by orders of magnitude. It has been suggested that chromium is integral to the glucose tolerance factor (GTF). Efforts to purify this factor have led to the detection of nicotinic acid, glycine, glutamic acid, and cysteine, as well as chromium (Mertz et al. 1974). Later studies, though, have failed in establishing an integral chromium factor.

Recent years have witnessed a plethora of activity related to the elucidation of a potential role for trivalent chromium in mammalian carbohydrate and lipid metabolism at a molecular level. In the 1980s, the isolation and characterization of a unique chromium-binding oligopeptide known as low-molecular-weight chromium-binding substance (LMWCr) or chromodulin were reported (Yamamoto et al. 1987). The oligopeptide has a molecular weight of about 1,500 Da and is comprised of only four types of amino acid residues (*i.e.*, glycine, cysteine, glutamate, and aspartate). Despite its small molecular weight, it binds four equivalents of chromic ions, apparently in a tetranuclear assembly, as necessitated by charge balance arguments (Vincent 2000). Chromodulin has an intriguing ability to potentiate the effects of insulin on the conversion of glucose into carbon dioxide or lipid. No other naturally occurring chromium-containing species potentiates insulin action in this manner (Vincent 2000).

How chromium is absorbed and transported is still uncertain. It appears that transport is mediated by transferrin, the main iron-transporting protein (molecular weight, 80 kDa). This may be due to the fact that transferrin usually carries only about a 30% load of iron so it has unused transportation capacity. Transferrin is also thought to be a transporter of various trace elements. Recent reports on the effects of insulin on iron transport and the relationship between hemochromatosis and hepatic iron overload and diabetes suggest that transferrin may actually be the major physiologic chromium transport agent (Vincent 2000).

Hexavalent chromium compounds have been established as being carcinogenic. Chromate easily enters cells through the sulfate channel and is quickly reduced by, for example, glutathione. The ultimate step of the metabolic pathway yields Cr(III) inserted within the cell nucleus, where it cross-links DNA to proteins. Recent results indicate that glutathione is not only a primary target for oxidation by chromate but also acts as an efficient ligand-stabilizing Cr (V) in a dimeric bridged cluster (Gaggelli et al. 2002).

Although the underlying mechanism by which Cr(III) exerts its beneficial effects remain unclear, numerous in vitro and in vivo studies have indicated that Cr(III) plays and important role in the maintenance of plasma cholesterol and triglycerides and the inhibition of oxidative stress and inflammatory cytokine secretion (Lau et al. 2008).

7.5.2.2 Molybdenum

Molybdenum is found in the second row of transition metals in the periodic table. It is the only metal in this row that is required by most living organisms. Although only a minor constituent of the Earth's crust, molybdenum is readily available to biological systems because of the solubility of molybdate salts in water. In fact, molybdenum is the most abundant transition metal in seawater. It is not, therefore, surprising that molybdenum has been incorporated widely in living organisms. Molybdenum is redox active under physiological conditions (ranging between oxidation states VI and IV). The V valence state is also available, and molybdenum can act as a transducer between obligatory two- and one-electron oxidation–reduction systems such as the hydroxylation of carbon centers under more moderate conditions than are required by other systems (Hille 2002).

An important feature of molybdenum-containing enzymes is the molybdenum cofactor (Moco) that is able to associate with different apoenzymes to form the Mo–holoenzymes where, depending on the type of apoenzymes, molybdenum catalyzes redox reactions on C, N, and S atoms. The only exception is bacterial nitrogenases that contain an FeMo cofactor not related to Moco (Mendel 1997). It has been shown that in Moco molybdenum is complexed by a pterin with a four-carbon alkyl side chain containing a Mo-coordinating dithiolene group and a terminal phosphate ester (Mendel 1997). This pterin has been named molybdopterin (Fig. 7.14).

There are two different kinds of molybdenum enzymes. One is exemplified by the nitrogenase enzyme family, which is characterized by a multinuclear center with iron, sulfur, and molybdenum (Fig. 7.15). In fact, molybdenum can be replaced by vanadium or iron. Nitrogenases catalyze the reaction from dinitrogen to ammonia and are the basic components in nitrogen fixation. Indeed, most cycling of nitrogen in the biosphere depends on the trace element molybdenum. The second type of molybdenum enzyme is characterized by a dependence on a mononuclear center, which is associated with molybdopterin. Three enzyme families are included in this group: xanthine oxidases, sulfite oxidases, and dimethylsulfoxide DMSO reductases. Examples of molybdenum enzymes of these families are discussed below.

Xanthine oxidases comprise the largest family of molybdenum enzymes, with up to 20 members. The xanthine oxidase catalyzes purine or pyrimidine catabolism by inserting oxygen and/or removing hydrogen from the substrates. It is a well-studied enzyme that exists as a dimer with a molecular weight of about 300 kDa. Each subunit contains one molybdenum, one flavin adenine dinucleotide, and two Fe_2S_2 centers. The reaction catalyzed is xanthine $+H_2O$ ∅uric acid + 2 H^+ + $2e^-$. Members of this family have been found broadly distributed within eukaryotes, prokaryotes and archaea (Romão 2009).

Sulfite is a highly reactive and potentially toxic compound. Like other reduced inorganic sulfur compounds such as

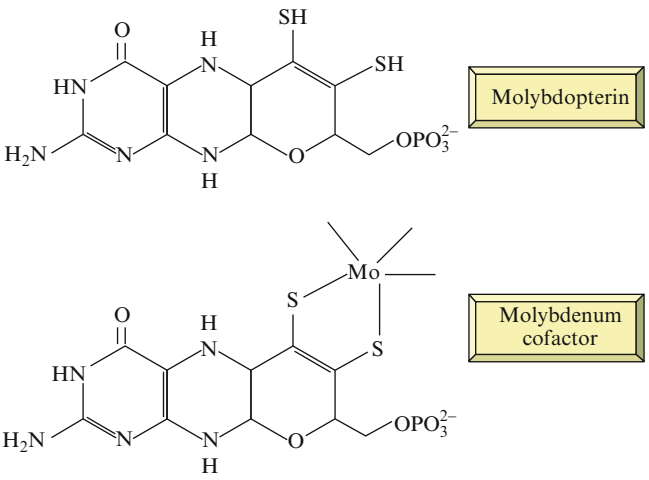


Fig. 7.14 Molybdopterin and molybdenum cofactor

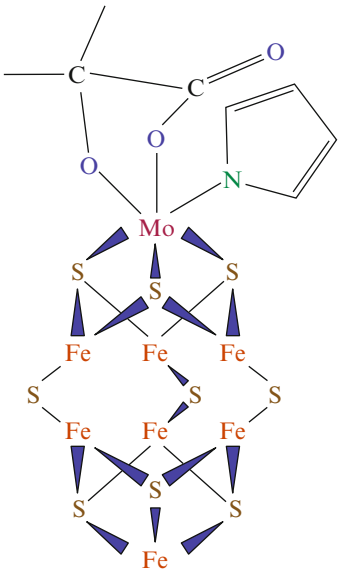


Fig. 7.15 Multinuclear center with iron, sulfur, and molybdenum in nitrogenase

hydrogen sulfide or thiosulfate, it occurs in nature as a consequence of geological and industrial processes and the anaerobic mineralization of organic matter by dissimilatory sulfate reduction (Kappler and Dahl 2001). There is no doubt that sulfite oxidases have developed during evolution as a response to sulfite reactivity. Sulfite oxidase contains only a single prosthetic group, a heme group, in addition to the Moco-derived center. The holoenzyme is a dimer with a molecular weight of 120 kDa. The enzyme can oxidize sulfite by using oxidized cytochrome *c*, ferricyanide, or dioxygen as an electron acceptor. Sulfite oxidation occurs at the molybdenum center, with the heme center serving to couple this two-electron oxidation to the reduction of two molecules of cytochrome *c*. The members of this family include not only sulfite oxidizing enzymes in plants, animals and bacteria but, additionally, the assimilatory eukaryotic nitrate reductases. These nitrate reductases catalyze the first and rate-limiting steps of nitrate assimilation in plants, algae and fungi and are very different from the bacterial nitrate reductases with respect to primary and tertiary structures (Romão 2009).

The dimethyl sulfoxide (DMSO) reductase family of molybdenum (and tungsten) enzymes is a considerably larger and more diverse family compared to xanthine oxidase and sulfite oxidase families. They are found in bacteria and archaea and include diverse enzymes like dissimilatory nitrate reductases and formate reductases (subfamily I), respiratory nitrate reductases I and ethylbenszene dehydrogenase (subfamily II); DMSO reductase and trimethylamine N-oxide reductase (subfamily III). With the exception of fomate dehydrogenases, members of the DMSO reductase family catalyze the reduction of dimethylsulfoxide to dimethylsulfide and liberate the oxygen atom of DMSO as water (Romão 2009). The reducing equivalents come from a specific pentaheme cytochrome. In addition to the enzymes described here are a few other molybdenum enzymes, such as pyridoxal oxidase, xanthine dehydrogenases, and pyropallol transhydrolases.

7.5.2.3 Tungsten

Tungsten is by far the heaviest metal with a biological function (*i.e.*, an essential trace element). The three classes of tungsten enzymes are aldehyde ferredoxin–oxidoreductase, formate dehydrogenase, and acetylene hydratase. Formate is a common metabolite in most life forms. In most cases, its production and consumption involve formate dehydrogenase (FDH), which catalyzes the reversible two-electron conversion of CO_2 to formate according to:

$$CO_2 + 2H^+ + 2e^- \Longleftrightarrow HCOO^-$$

Not all FDHs are tungsten dependent. On the other hand, most FDHs of aerobic organisms do not contain metals or other cofactors. The tungsten-dependent FDHs have mostly been identified in Chlostridiae (e.g., *C. thermoaceticum*, *C. formioaceticum*, and *C. acidiurici*) (Kletzin and Adams 1996).

Formyl methanofuran dehydrogenase (FMDH) catalyzes the reversible formation of *N*-formylmethanofuran from CO_2 and methanofuran (MFR). The first step in CO_2 utilization by methanogens is thus:

$$CO_2 + MFR^+ + H^+ + 2e^- \Longleftrightarrow CHO - MFR + H_2O$$

However, FMDH was found to be a molybdoenzyme in the first methanogens examined (Kletzin and Adams 1996).

Tungsten-dependent FMDHs have been isolated from moderate thermofils such as *Methanobacterium wolfei* and *M. thermoautotrophicum*.

Aldehyde-oxidizing enzymes catalyze oxidation of aldehydes of one type or another and again have been found in microorganisms. The reaction catalyzed is

$$CH_3CHO + H_2O \Longleftrightarrow CH_3COO^- + 3H^+ + e^-$$

Among the aldehyde-oxidizing enzymes are carboxylic acid reductase (CAR) from acetogens which was the second tungsten-dependent enzyme isolated. Another tungsten-dependent enzyme is aldehyde ferredoxin oxidoreductase (AOR), from *Pyrococcus furiosus*, which catalyzes the oxidation of a range of aliphatic and aromatic aldehydes and reduces ferredoxin. Formaldehyde ferredoxin oxidoreductase (FOR) has also been demonstrated in *P. furiosus*. These enzymes also use ferredoxin as the physiological electron carrier and are maximally active at temperatures above 95°C. In contrast to AOR, however, they oxidize only C1–C3 aldehydes and are of much lower specific activity (Kletzin and Adams 1996).

The enzyme glyceraldehyde-3-phosphate ferredoxin oxidoreductase (GAPDH) has been found in the peculiar hyperthermophilic archaeon *P. furiosus* and is thought to play a role in gluconeogenesis. A fourth tungsten-containing enzyme from *P. furiosus* has recently been characterized (Roy and Adams 2002). This enzyme, preliminarily named WOR 4, is thought to play a role in SO reduction. It has an interesting structure in that it contains approximately one W atom, three Fe atoms, three or four acid-labile sulfides, and one Ca atom per subunit (Roy and Adams 2002). Aldehyde dehydrogenases (ADH) have been shown in *Desulfovibrio gigas*.

Acetylene hydratase (AH) converts acetylene to acetaldehyde according to:

$$H - C \equiv C - H + H_2O \rightarrow H_3C - HC = O$$

which is a hydration reaction.

In addition to availability a key factor in tungsten utilization appears to be its redox properties relative to molybdenum. When tungsten substitutes for molybdenum in molybdoenzymes, some will be inactive, probably because the tungsten site has a lower reduction potential as compared to the molybdenum site. Conversely, to catalyze a reaction of extremely low potential, tungsten should be preferred over molybdenum (Kletzin and Adams 1996).

The first protein to be able to discriminate between the similar oxyanions of tungstate and molybdatae was TupA (tungsten uptake) of *Eubacterium acidaminophilum*. The genes for this ABC transporter are situated within the cluster of genes involved in molybdopterin cofactor biosynthesis. This cluster is close the genes endocing the tungsten-containing FDH 1 and adjacent to two genes of *moeA*, which is responsible for the final incorporation of Mo or W into the pyranopterin cofactor (Andreesen and Makdessi 2008).

Table 7.8 Manganese-containing vesicles

Vesicle	Known Mn protein
Chloroplast	Superoxide dismutase
Golgi apparatus	Glycosyl-transferases
Lysosome (acid)	Acid phosphatase (Mn(III))
Mitochondria	Superoxide dismutase
Thylakoid (acid)	Generation of molecular oxygen
Vacuole (acid)	Mn(II) is free

7.5.3 Manganese: Photosynthesis and Defense Against Oxygen

Although fairly abundant in the biosphere, manganese is found only in trace amounts in living organisms. Manganese exhibits the widest range of oxidation states of any of the first row *d*-block metals. The lowest states are stabilized by π-acceptor ligands, usually in organometallic complexes. Most of the biochemistry of manganese can be explained by two properties: It is redox active, and it is a close but not exact analog of Mg^{2+}. Manganese plays many roles in biological systems ranging from acting as a simple Lewis acid catalyst to being an element that can transverse several oxidation states to carry out water oxidation. The presence of Mn^{2+} in the cytoplasm in significant concentrations would pose a serious problem for aerobic eukaryotes. One reason is that manganese is considered to be mutagenic, and it binds relatively weakly to proteins. Prokaryotes, on the other hand, may make use of this property to promote mutations to generate variation. Manganese in aerobic eukaryotic cells, then, has to be pumped into various organelles. Examples of such organelles are presented in Table 7.8.

An intriguing property of manganese in its divalent form is that it can be used for magnetic resonance imaging (MRI). Being paramagnetic it induces T1-weighted image enhancement, which increases the quality of anatomic MRI. Because manganese enters synaptically activated neurons through voltage-gated calcium channels, it is also an interesting marker for functional MRI (Schmitt et al. 2011).

7.5.3.1 Generation of Dioxygen from Water

Higher plants and algae complete oxygenic photosynthesis whereby water is oxidized by a cluster of manganese ions. This process takes place in Photosystem II (PS II), which is a multisubunit complex embedded in the thylakoid membranes. This photosynthetic enzyme catalyzes the most thermodynamically demanding reaction in biology,

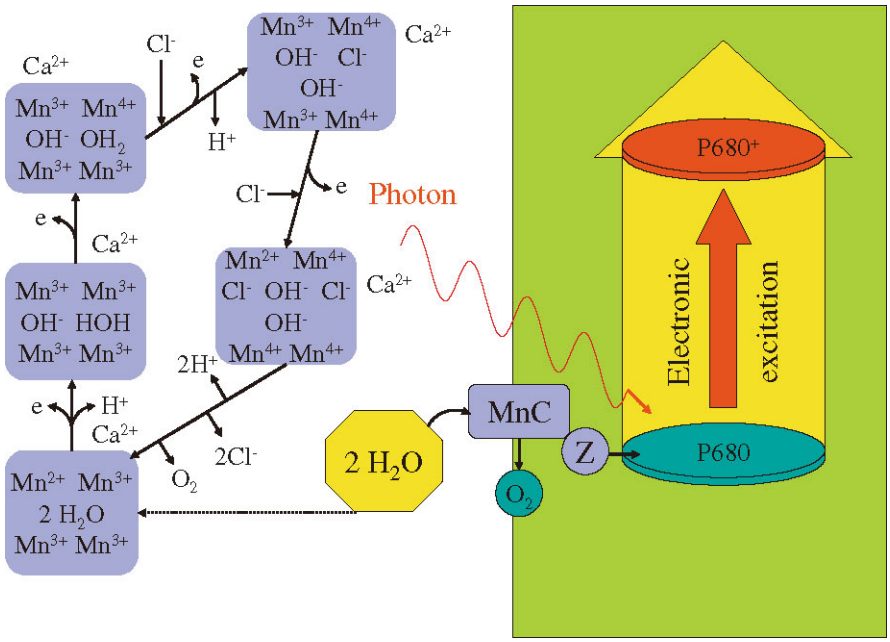


Fig. 7.16 A partly hypothetical model of the splicing of water

the splitting of water into dioxygen and reducing equivalents (Barber 2003). Light is captured by antennae chlorophyll and funneled to the primary reaction center, which contains chlorophyll (P680). The oxidized P680 first oxidizes a redox-active tyrosine, hydrogen bonded to a histidine (Yz). Subsequent oxidation of the manganese cluster occurs via this generated tyrosine radical. The enzyme can exist in five oxidation levels named S_n states, where S_0 is the most reduced and S_4 the most oxidized. Oxygen is liberated upon the $S_3 \varnothing S_4 \varnothing S_0$ transition (Yocum and Pecoraro 1999). A partly hypothetical model of this process is depicted in Fig. 7.16. It is interesting to note that the oxidation of water to dioxygen takes place in the thylakoid membranes of plants and algae. These membranes correspond to the mitochondrial membrane of eukaryote cells, in which the reverse process of reducing dioxygen to water takes place. Even more intriguing is that manganese plays important roles in both processes.

An interesting question is why manganese was chosen in the evolution of the oxidation of water to molecular oxygen. The metal selectivity of the oxygen evolving complex is highly specific and only assembles with manganese. Additional ions such as bicarbonate and chloride also influence the activity of the complex. However, they are outside the first coordination sphere on manganese where the water splitting reaction occurs. It could be that early oxygenic autotrophs initially used manganese because of its abundance as $Mn^{2+}(H_2O)_6$ in the Archean oceans. The Mn cation is very stable and exchanges ligands rapidly. It also forms soluble salts. These properties together with the high natural abundance and the fact that the demands of a cell for manganese are very low means that supply often exceeds demand (Williamson et al. 2011).

7.5.3.2 Defense for a Life with Oxygen

The enzyme most commonly associated with manganese is superoxide dismutase, which is a scavenger of the very reactive superoxide anion produced during reduction of oxygen in cellular respiration. As the name suggests, the enzyme acts on the superoxide anion produced in the first step in the reduction of dioxygen to water. It catalyzes the following reaction:

$$\mathrm{Mn(III) + O_2^- \rightleftharpoons [Mn(III) - O_2^-] \rightarrow Mn^{2+} + O_2}$$

$$\mathrm{Mn^{2+} + O_2^- \rightleftharpoons [Mn^{2+} - O_2^-] + 2H^+ \rightarrow Mn(III) + H_2O_2}$$

The end product, hydrogen peroxide, although not a radical, is a reactive oxygen species. MnSODs are widespread in bacteria, plants, and animals. In most animal tissues and yeast, MnSOD is largely confined to mitochondria (Fridovich 1995). It is conspicuous that, although manganese is deeply involved in the generation of oxygen from water, it is also necessary in scavenging reactive oxygen species when dioxygen is metabolized in mitochondria. MnSOD is the most phylogenetically widespread manganese-dependent enzyme (Kehres and Maguire 2003).

There are some aspects of redox biochemistry that have to be taken into account to understand the unexpected tolerance of relatively high concentrations of Mn^{2+}. Similar to iron, manganese can cycle readily *in vivo* between the 2+ and 3+

oxidation states; however, the reduction potential of any molecule depends on its ligand environment. Another important feature is that manganese is less reducing than iron under most biological conditions (Kehres and Maguire 2003). "Free" (solvated) Mn^{2+} has a reduction potential too high to reduce H_2O_2 in aqueous solution; however, replacing one or two inner hydration shell waters with hydrogen carbonate results in rapid catalase activity by a Mn^{2+}-dependent disproportionation reaction. That the reducing potential of manganese is smaller than that of iron can be explained by the different $3d$ electron occupancies of the two cations. The electrochemical consequence is that $Mn^{3+} + e^- \varnothing Mn^{2+}$ has a standard reduction potential of +1.51 V, while $Fe^{3+} + e^- \varnothing Fe^{2+}$ has a standard reduction potential of +0.77 V.

Two consequences of this redox chemistry are obvious. The similarity between Mn^{2+} and Fe^{2+} is that their intrinsic reduction potentials are close enough to those of many common biological molecules that each metal can take part in biologically important redox catalysis. In addition, the critical difference between the two metals is that the higher reduction potential of Mn^{2+} renders the free (*i.e.*, solvated) Mn^{2+} innocuous under conditions (notably in an aerobic environment) where free Fe^{2+} would actively generate toxic radicals. Cells can thus tolerate very high cytoplasmic concentrations of Mn^{2+} with essentially no negative redox consequences. This is not the case with iron or with any other biologically relevant redox-active metal (Kehres and Maguire 2003).

Currently, very little is known about how Mn-SOD acquires its essential cofactor inside the cell. Studies on MnSOD from *E. coli* show that metal binding in vivo shows only a modest selectivity. This means that Mn, Fe and mixed (Mn, Fe) forms of the protein are all present. In spite of this peculiar binding behavior, only the manganese form is catalytically active. This conspicuous contrast between low selectivity in metal binding and strict specificity for catalysis constitutes a paradox of MnSOD function (Whittaker 2010).

7.5.3.3 Similarity to Magnesium

The greatest similarity between these two metals is in the context of structure. Mg^{2+} is an ideal structural cation for biological molecules, especially phosphorylated ones such as nucleic acids and many intermediary metabolites. This is due to a complete lack of d electrons in Mg^{2+}, and its $2s^2 2p^6$ electron configuration confines it to strict octahedral liganding geometry with liganding bond angles very close to 90°. This geometry is useful in organizing the conformations of complex compounds or macromolecules (Kehres and Maguire 2003). The lack of d electrons is important because there is very little covalent interaction between Mg^{2+} and its ligands; therefore, Mg^{2+} is a labile and rapidly exchangeable cation that does not interpose itself in the way of other close intermolecular interactions. Mn^{2+}, with its relatively similar ionic radius and relatively minor involvement of its stable symmetric $3d^5$ shell electrons in bonding, readily exchanges with Mg^{2+} in most structural environments and exhibits much of the same octahedral, ionic, labile chemistry. However, less similarity is seen in catalysis. The $3d^5$ electrons of Mn^{2+} do interact to some extent with electrophilic ligands. Thus, Mn^{2+}-ligand bonds are generally much more flexible than Mg^{2+}-ligand bonds, in both length and angle (Kehres and Maguire 2003).

7.5.3.4 Manganese and Reactive Oxygen Species

The manganese-dependent superoxide dismutase has already been mentioned. In addition to this enzyme is a set of enzymes involved in the detoxification of reactive oxygen species that are dependent on manganese. There is a family of manganese-dependent catalases that often are referred to as "non-heme catalases." They are structurally and mechanistically unrelated to conventional catalases that are cofactored by iron in a prosthetic heme group. Both classes operate on hydrogen peroxide. In addition to catalases, a family of catabolic heme enzymes known as manganese peroxidases couples the redox activity of H_2O_2 to the degradation of nutrients such as lignin via oxidation of a manganese bound to a propionate side-chain of the heme group. An intriguing phenomenon is that salens, synthetic chelators that are derivatives of *N,N′-bis*(salicylidene)ethylendiamine chloride, form complexes with Mn^{2+} that show efficacy as combined superoxide dismutase/catalase mimics. They have been shown, in their oxo forms, to oxidize nitric oxide and nitrite to the more benign nitrate *in vitro* (Kehres and Maguire 2003). A summary of manganese-related enzymes and proteins can be found in Table 7.9.

7.5.3.5 Relation Between MnSOD and the Apoptotic Transcription Factor

The apoptotic transcription factor p53 regulates a myriad genes engaged in many cellular functions, including cell cycle regulation, apoptosis, proliferation, DNA replication and repair and stress response. In addition, p53 is an important tumor suppressor and is often mutated in cancer (Holley et al. 2010).

The transcription factor also affects the production of reactive oxygen species. Knockout of p53 in primary human fibroblasts results in a decrease of mitochondrial mass, which correlates with a significant increase in cellular hydrogen peroxide and a reduction in both mitochondrial and cellular superoxide levels (Lebedeva et al. 2009).

The antioxidant enzyme MnSOD is regulated by p53 and this regulation is important for controlling mitochondrial generation of reactive oxygen species. This relationship between MnSOD expression and p53 levels correlates with different cancer cell lines that express reduced levels or mutated forms of p53, indicating that p53 may regulate MnSOD expression (Holley et al. 2010).

Table 7.9 Manganese-dependent enzymes and proteins

Enzyme/Protein	Function	Occurrence
MnSOD	Detoxify superoxide radical anion	Bacteria, archaea, and eukaryotes
Non-heme Mn-catalase	Detoxify hydrogen peroxide	Bacteria
Transcription factor (*mntR*)	Repress Mn^{2+} uptake transporter expression	Homologs in diverse bacteria; extent unknown
ppGpp hydrolase (*spoT*)	Hydrolyze RNA synthesis regulator ppGpp	Practically ubiquitous in bacteria
Protein phosphatases	Dephosphorylate many cellular proteins	All cells; highly conserved between prokaryotes and eukaryotes
Agmatinase	Synthesize osmoprotectant putrescine from angmatine (decarboxylated arginine)	Many Gram-negative bacteria
Aminopeptidase P	Hydrolyze typical X-Pro sequence	Enterobacteria, most other bacteria
Phosphoglyceromutase	Catalyzes interconversion of 3-phosphoglycerate and 2-phosphoglycerate	Enterobacteria, other bacteria, and plants
Fructose-1,6-BP phosphatase	Convert fructose-1,6-BP to fructose-6-phosphate	Enterobacteria; extent otherwise unknown
Adenyl cyclase	Synthesize cyclic AMP	*Mycobacterium tuberculosis*
Aromatic hydrocarbon metabolism	Oxidation of catechols and other aromatics	*Arthrobacter globiformis* and similar enzymes in many soil bacteria
Lipid phosphotransferases	Modify or remove polar headgroups on lipids	Enterobacteria, Gram-positive bacteria
Polysaccharide polymerases	Synthesize capsular or secreted polysaccharide	Some Gram-positive and -negative bacteria
Protein kinases	Phosphorylation of unknown proteins	Extent unknown
Pyruvate carboxylase	Catalyze carboxylation of pyruvate to oxaloacetate	Eukaryotes, *Bacillus licheniformis*, and *Mycobacterium smegmatis*
Ribonucleotide reductase	Convert ribonucleotides to deoxyribonucleotides	Most bacteria, eurkaryotes
Arginase	Convert arginine to urea + ornithine	Higher eukaryotes, liver, and macrophages/monocytes; *Bacillus* sp.
Concavalin A	Plant lectin binding	Plants
Mn-lipoxygenase	Synthesize lipoxins from fatty acids	Fungi
Mn-peroxidase	Degrade lignin	White- and brown-rot fungi
Photosynthetic reaction center	Convert H_2O to O_2	Photosynthetic bacteria and plants

Source: Adapted from Kehres and Maguire (2003)

7.5.4 Iron: Savior and Threat

Iron is the most important of all metals and is the fourth most abundant element in the Earth's crust. It is a *d*-block element that can exist in oxidation states ranging from −2 to +6. In biological systems, however, these oxidation states are limited primarily to the ferrous (+2), ferric (+3), and ferryl (+4) states. Fe^{3+} is quite water insoluble and significant concentrations of water-soluble Fe^{3+} species can be attained only by strong complex formation. The interconversion of iron oxidation states is not only a mechanism whereby iron participates in electron transfer but also a mechanism whereby iron can reversibly bind ligands. Iron can bind to many ligands by virtue of its unoccupied *d* orbitals. The preferred biological ligands for iron are oxygen, nitrogen, and sulfur atoms. Iron(III) is a hard acid that prefers hard oxygen ligands, while iron(II) is on the borderline between hard and soft and favors nitrogen and sulfur ligands. The electronic spin state and biological redox potential (from +1,000 mV for some heme proteins to −550 mV for some bacterial ferredoxins) of iron can change according to the ligand to which it is bound. By exploiting the oxidation state, redox potential, and electron spin state of iron, nature can precisely adjust iron's chemical reactivity (Beard 2001). Biologically important iron-containing proteins carry out oxygen transport and storage, electron transfer, and substrate oxidation–reduction. Four major classes of protein carry out these reactions (Beard 2001) in mammalian systems: (1) iron-containing, nonenzymatic proteins (hemoglobin and myoglobin); (2) iron–sulfur enzymes; (3) heme-containing enzymes; and (4) iron-containing enzymes that are non-iron-sulfur, non-heme enzymes (Fig. 7.17).

7.5.4.1 Hemeproteins

Proteins with heme as the prosthetic group carry out important functions in biological systems; thus the biosynthesis of heme is very important. It begins with the synthesis of tetrapyrroles, where the term *tetrapyrrole* indicates compounds containing four linked pyrrole rings. Four classes of such compounds are very common in biology: (1) the widely distributed iron porphyrin, heme; (2) the chlorophylls of plants and photosynthetic bacteria; (3) the phycobilins,

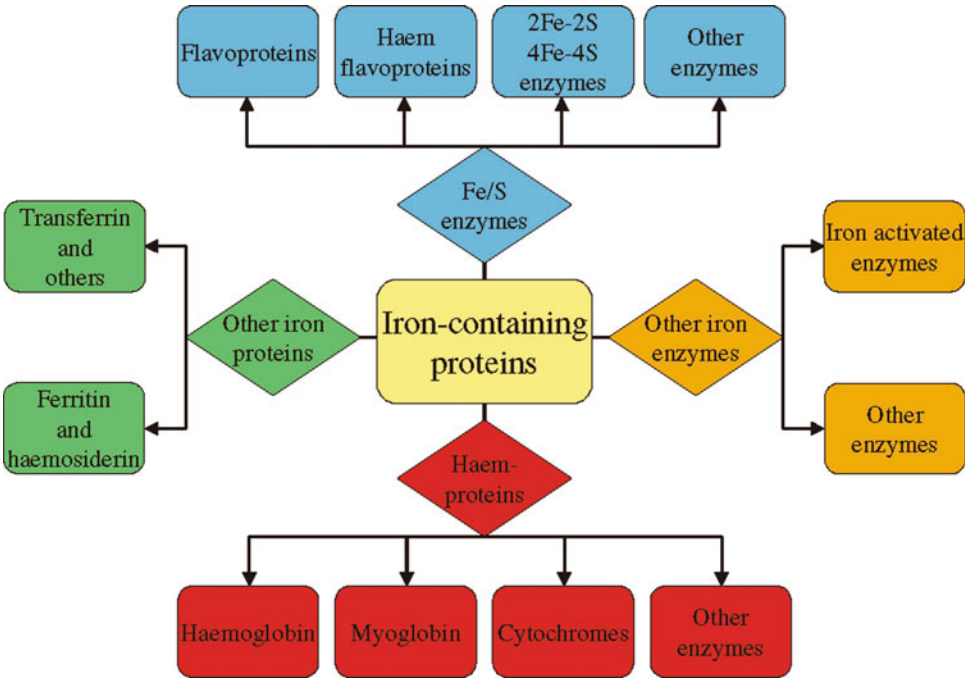


Fig. 7.17 Iron-containing proteins

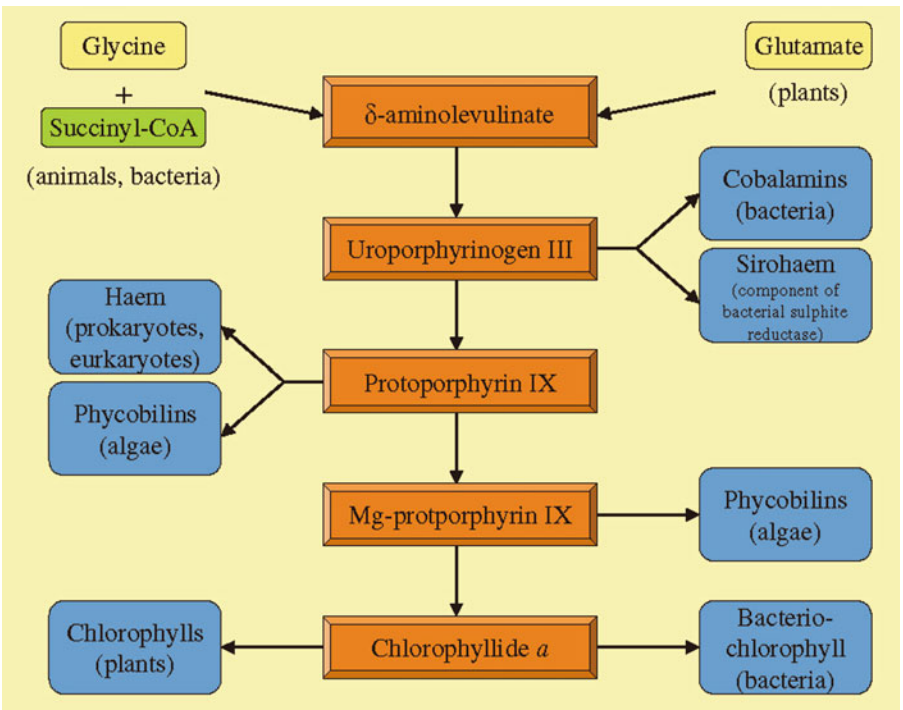


Fig. 7.18 The principal steps of heme biosynthesis

photosynthetic pigments of algae; and (4) the cobalamins, especially vitamin B_{12} and its derivatives. All tetrapyrroles are synthesized from a common precursor, δ-aminolevulinic acid (ALA). Figure 7.18 summarizes the relationships between the synthetic pathways.

7.5.4.2 Hemoglobin and Myoglobin

Evolution has provided animals with hemoglobin and myoglobin for oxygen transport and storage. Aerobic metabolism requires some kind of oxygen transporter because relying on diffusion would be adequate for only very small animals. Insects, however, have solved the problem with oxygen transport through their networks of tubes (tracheae) leading from the body surface to the inside tissues.

Oxygen transport and storage are tricky problems to solve. The molecules used must bind dioxygen without allowing oxidation of other substances, thereby reducing dioxygen.

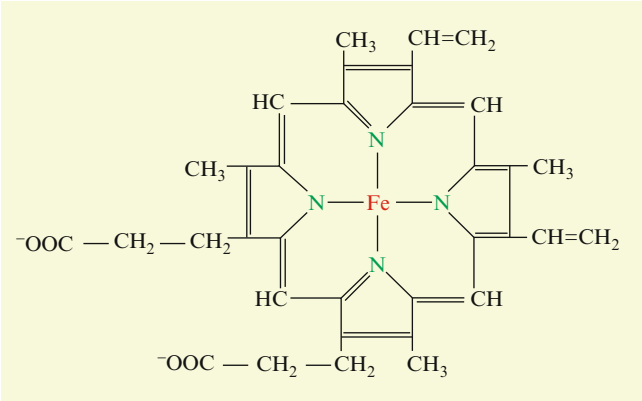


Fig. 7.19 Iron binding in protoporphyrin IX. The complex is called *heme*

Are proteins suited for direct oxygen binding? The answer is no; however, *d*-block metals such as iron and copper in their lower oxidation states readily bind oxygen. Proteins can bind Fe(II) in various ways. In hemoglobin and myoglobin, iron is bound by the tetrapyrrole ring protoporphyrin IX (see Fig. 7.19). When iron is bound to protoporphyrin IX, the system is referred to as ferroprotoporphyrin or heme. In addition to its use in hemoglobin and myoglobin, the heme group is a prosthetic group in a variety of proteins (see below). This makes the biosynthesis of heme very important. Myoglobin is a smaller molecule relative to hemoglobin and is the principal oxygen storage protein.

An interesting property of hemoglobin, in addition to its oxygen-binding ability, is its enzymatic activity. It has been thought that the oxygen carrier function was so specialized that globins were not recruited to new tasks; however, it has recently been found that the globin of some marine worms (*Amphitrite ornata*) has evolved into a powerful peroxidase, more precisely dehaloperoxidase. This enzyme catalyzes the oxidative dehalogenation of polyhalogenated phenols in the presence of hydrogen peroxide at a rate at least ten times faster than all known halohydrolases of bacterial origin. *A. ornata* can thus successfully survive in an environment where other species secrete brominated aromatics and other halocompounds as repellents (Lebioda 2000).

7.5.4.3 Iron Usage in Mitochondria

The use of iron in cells is mostly confined to mitochondria where it is used for the synthesis of heme and iron-sulfur clusters (ISCs). Although the trafficking of iron is rather well understood as a whole, the trafficking to mitochondria is less well known. Kinetic data and microscopy studies in erythroid cells, however, provide support for the "kiss and run" hypothesis, which suggests the direct delivery of transferrin-derived iron to mitochondria through a transient contact with the endosome (Richardson et al. 2010). It is generally accepted, in other cell types, that iron acquired during the transferrin cycle is first released into the cytosol and then targeted to the mitochondria. An essential role in intracellular iron sensing and trafficking seems to be played by the conserved cytosolic glutaredoxins Grx3 and Grx4 (Mühlenhoff et al. 2010). The entry of iron into mitochondria requires the SLC transporter mitoferrin. All organisms synthesize the tetrapyrrol porphyrin ring of heme from the universal precursor 5-aminolevlinic acid (ALA). Most eukaryotes, except plants, generate ALA in mitochondria. The transport of heme and its metabolic intermediates across the mitochondrial membranes probably involve members of the ABC family of transporters and the transporter SLC25A39.

The pathways of assembly and repair of ISCs are complex; the functions of several proteins have been defined but new players continue to emerge. The mitochondrial proteins Isu1/2 and Isa1/2 provide a scaffold for the early steps of ISC biogenesis. The cysteine desulfarase nitrogen fixation homolog 1 in complex with iron-sulfur protein biogenesis, desulfarase interacting protein 11 generates elemental sulfur, whereas the iron-binding protein frataxin seems to serve as an iron donor (Wang and Pantopoulos 2011).

7.5.4.4 Cytochromes

Cytochromes constitute a group of hemeproteins with distinctive visible-light spectra that function as electron carriers from biological fuels to oxygen. The major respiratory cytochromes are classified as *b*, *c*, or *a*, depending on the wavelengths of the spectral absorption peaks. They are vital members of the mitochondrial electron-transfer chain. Figure 7.20 shows a schematic representation of part of this electron-transfer chain. In addition, the cytochromes are also essential components in plant chloroplasts for photosynthesis. It is the ability of the iron center to undergo reversible Fe(III) × Fe(II) changes to allow them to act as electron-transfer centers. In the mitochondrial electron-transfer chain, cytochrome *c* accepts an electron from cytochrome c_1 and then transfers it to cytochrome *c* oxidase. The electron is ultimately used in the four-electron reduction of dioxygen.

The iron-reducing bacterium *Shewanella oneidensis* MR-1 has the capacity to iron cycling over the long term by respiring on crystalline iron oxides like hematite when poorly crystalline phases are depleted. The ability of outer membrane cytochromes OmcA and MtrC of MR-1 to bind to and transfer electrons to hematite has led to the suggestion that they function as terminal reducteases when this mineral is used as a respiratory substrate (Mitchell et al. 2012).

7.5.4.5 Cytochrome P-450 Enzymes

Cytochrome P-450 enzymes are hemeproteins that function as monooxygenases to catalyze the insertion of oxygen into a C—H bond of an aromatic of aliphatic hydrocarbon

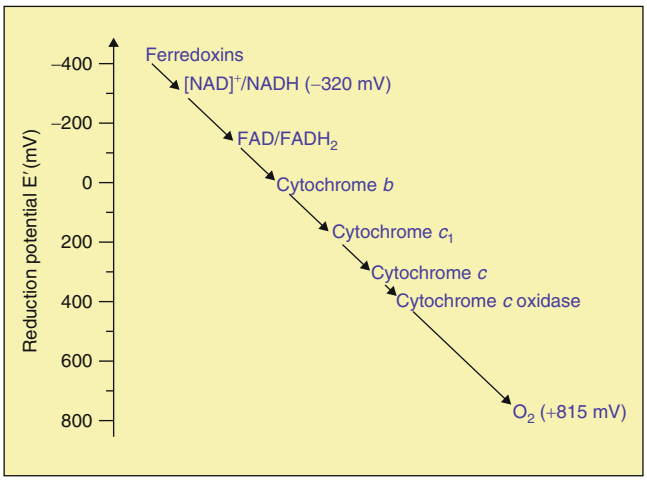


Fig. 7.20 The mitochondrial electron-transfer chain

(*i.e.*, the conversion of RH to ROH). Examples of the biological functions of cytochrome P-450 are drug metabolism and steroid synthesis. Carbon monoxide adducts of cytochrome P-450 absorb at 450 nm; thus its name. One of the many interesting aspects of cytochrome P-450 enzymes is that some are inducible, which means that following exposure of the cell to an inducing chemical enzyme activity increases, in some cases several orders of magnitude. It has been proposed that the catalytic cycle begins with the enzyme in a resting state with iron present as Fe(III). The hydrocarbon substrate then binds, followed by one-electron transfer to the iron porphyrin. The resulting Fe(II) complex with bound substrate proceeds to bind O_2. A key reaction is the reduction of the porphyrin ring of the oxygen complex by a second electron, which produces the ring radical anion. Uptake of two H^+ ions then leads to the formation of the Fe(IV) oxo-complex, which attacks the hydrocarbon substrate to insert oxygen. Loss of ROH and uptake of an H_2O molecule at the vacated coordination position bring the cycle back to the resting state (Ortiz de Montellano and De Voss 2002).

Three cytochrome P-450 (CYP) hydroxylases are responsible for vitamin D synthesis and degradation. The presence of CYP2R1, CYP27B1 and CYP24A1 in the prostate indicates that analogs of vitamin D might be developed and used for the prevention and treatment of prostate cancer (Chen et al. 2012).

7.5.4.6 Iron-Activated Enzymes

The most prominent example of this class of enzymes is heme oxygenase, which has evolved to carry out the oxidative cleavage of heme, a reaction essential in several physiological processes as diverse as iron reutilization and cellular signaling in mammals and synthesis of essential light-harvesting pigments in cyanobacteria and higher plants, as well as the acquisition of iron by bacterial pathogens (Wilks 2002). Heme oxygenase (HO) is concentrated in both blood vessel endothelium and adventitial neurons which suggests that HO subserves functions that are handled by NO-generating enzymes. The gene of this cellular stress protein, mediating the catabolism of heme to biliverdin in brain and other tissues, is strongly induced by dopamine, oxidative stress, and metal ions. In the brain, it is primarily expressed in the astroglia, and, when upregulated, HO promotes mitochondrial iron deposition in these cells (Schipper 1999). HO protects cells by degrading prooxidant metalloporphyrins and appears to facilitate iron efflux from the cell (Berg et al. 2001). HO is responsible not only for iron export but also for the intracellular sequestration of iron by glial mitochondria (Schipper 1999).

In addition to heme oxygenase, new alcohol dehydrogenases (ADHs) have been discovered. At first an iron-activated was found in *Zymomonas mobilis*. Shortly thereafter similar enzymes were demonstrated in *Saccharomyces cerevisiae* and *Escherichia coli*. It was thought that these new iron-activated ADHs were restricted to microbes. Later it was shown that an iron-activated ADH, named ADHFE1, was found in human fetal brain as well as in adipose tissue (Kim et al. 2007a, b). Another finding partly related to iron is the increase of reactive oxygen species (ROS). Exogenous compounds inducing production of ROS can be used in the treatment of cancer. Examples are arsenic trioxide and ferrocenium ion containing slats. These ROS regulating drugs not only kill cancer cells but also increase the ROS amount in normal cells. Prodrugs that are converted to toxic species at cancer-specific conditions, for example high contents of ROS, potentially lack this dangerous side effect. Aminoferrocene-based prodrugs have been tested in human promyelocytic leukemia, human glioblastoma-astrocytoma and non-malignant fibroblasts. They were found to simultaneously inhibit the antioxidative cellular system and to induce catalytic generation of ROS (Hagen et al. 2012).

7.5.4.7 Fe/S Enzymes

The porphyrin ligand environment of iron that occurs in hemoglobin and myoglobin is also important in redox enzymes. Thus, in the large class of biochemically important heme proteins iron is coordinated to a porphyrin ligand. All other iron proteins are defined as non-heme. Those that contain iron in a tetrahedral environment of four sulfur atoms are particularly important. Iron–sulfur clusters, however, were not familiar to inorganic chemists prior to recognition of their biochemical importance. Iron–sulfur clusters are simple inorganic groups that are contained in a variety of proteins having functions related to electron transfer, gene regulation, environmental sensing, and substrate activation.

Table 7.10 Some examples of proteins and enzymes with [Fe–S] clusters and their function

Enzyme or Protein	Function	Cofactors
Rubredoxin	Electron transfer	Fe
Ferredoxin, Rieske ferredoxin	Electron transfer	[2Fe–2S]
Phthalate dioxygenase reductase	Electron transfer	[2Fe–2S], FAD
Naphthalene dioxygenase	O_2-activation	[2Fe–2S], Fe(II)
Adenine glycosylase, glutamine PRPP amidotransferase, endonuclease III	Structural stabilization	[4Fe–4S]
Aconitase	Electrophilic catalysis	[4Fe–4S]
IRP-1[a], FNR[b], soxRS[c]	Regulation of gene expression	[4Fe–4S], [2Fe–2S]
Dinitrogenase reductase, dinitrogenase	Nitrogen fixation	[4Fe–4S], FeMoCo, P-clusters

[a] Iron regulatory protein-1
[b] Ferredoxin:NADP + oxidoreductase
[c] Superoxide regulatory system

Table 7.11 Protein components of the mitochondrial electron-transfer chain

Enzyme Complex	Mass (kDa)	Number of subunits	Prosthetic group(s)
I NADH dehydrogenase	850	42	FMN[a], [Fe–S]
II Succinate dehydrogenase	140	5	FAD[b], [Fe–S]
III Ubiquinone: cytochrome *c* oxidoreductase	250	11	Hemes, [Fe–S]
Cytochrome *c*	13	1	Heme
IV Cytochrome oxidase	160	13	Hemes, Cu_A, Cu_B

[a] Flavin mononucleotide
[b] Flavin adenine dinucleotide

Biological Fe–S clusters, however, are not formed spontaneously, but a consortium of highly conserved proteins is required for both the formation of Fe–S clusters and their insertion into various protein partners. The formation or transfer of Fe–S clusters appears to require an electron-transfer step (Frazzon et al. 2002). Table 7.10 provides examples of proteins with Fe–S clusters and their function.

Although not all of the enzymes involved in the mitochondrial respiratory chain are iron–sulfur proteins, it is enlightening to look through the various parts of the chain. Electrons move from reduced nicotinamide adenine dinucleotide (NADH), succinate, or some other primary electron donor, through flavoproteins, ubiquinone, iron–sulfur proteins and cytochromes, and finally to O_2. Table 7.11 summarizes the protein components of the mitochondrial electron-transfer chain.

In the context of iron–sulfur proteins, we can observe an intriguing link to another essential metal (namely, molybdenum) through xanthine dehydrogenase and xanthine oxidase. These are the most studied of the small but important class of molybdenum-containing iron–sulfur flavoproteins. Xanthine oxidoreductase catalyzes the hydroxylation of a wide variety of purine, pyrimidine, pterin, and aldehyde substrates. The active form of the enzyme is a homodimer of molecular mass 290 kDa, and each of the monomers acts independently in catalysis. Each subunit contains one molybdopterin group, two identical 2Fe–S centers, and one flavin adenine dinucleotide cofactor (Nishino and Okamoto 2000).

7.5.4.8 Transferrin, Lactoferrin, and Hemopexin

Transferrin originally was the name of the serum protein that binds and transports iron for delivery to cells. Today, it is the name applied to a wider family of homologous proteins that includes serum transferrin, lactoferrin, ovotransferrin, and melanotransferrin (Baker et al. 2003). Only serum transferrin has a proven transport function; however, transferrins seem to be involved in the homeostatic control of free iron in all the places where it might be found. The serum transferrin has two structurally similar lobes. Both lobes contain one iron-binding center that is very specific for iron(III) and has a binding constant of about 10^{20}. It is only when transferrin is loaded with two ferric ions that it binds strongly to the receptor for internalization. In addition to iron(III), transferrin can bind strongly to a range of other metal ions. Many such complexes are still recognized by the transferrin receptor, and some bacteria have transferrin receptors. Transferrin, therefore, holds promise for use in the development of antimicrobial therapies (Andrews et al. 2003).

Lactoferrin is an iron-binding protein that binds iron even more tightly than transferrin. It is present in milk, many other exocrine secretions, and white blood cells (Baker et al. 2003). One of the first functions attributed to lactoferrin was the ability to inhibit bacterial growth and viral infection. It is thought that lactoferrin is able to sequester iron from certain pathogens, which inhibits their growth. Another important function is its ability to stimulate the release of the neutrophil-activating polypeptide interleukin-8 (IL-8). This suggests that lactoferrin may

function as an immunomodulator for activating the host defense system (Kruzel and Zimecki 2002).

Hemopexin is a recycler and transporter of heme. Turnover of heme proteins, notably hemoglobin, leads to the release of heme into extracellular fluids with potentially severe consequences. Like free iron, heme is a source of essential iron for invading bacterial pathogens and is highly toxic because of its ability to catalyze free-radical formation. Protection is given by hemopexin, a 60-kDa serum glycoprotein that sequesters heme with very high affinity from the bloodstream; transports it to specific receptors on liver cells, where it undergoes receptor-mediated endocytosis; and releases the bound heme into cells. It thus serves both to protect against heme toxicity and to conserve and recycle iron (Baker et al. 2003).

The role of hemopexin in the control of heme is well established. Recent findings add to this role in several aspects. It acts as a multifunctional agent in important processes like iron homeostasis, antioxidant protection, signaling pathways to promote cell survival and gene expression (Tolosano et al. 2010). There are indications of activities such as hemopexin as a hyaluronidase, serine protease, pro-inflammatory or anti-inflammatory mediator, necrosis-suppressing factor, inhibitor of cellular adhesion and metal ion binding protein (Mauk et al. 2011).

7.5.4.9 Ferritin and Hemosiderin

Ferritins constitute a class of iron storage proteins found in bacterial, plant, and animal cells. They form hollow, spherical particles in which up to 4,500 iron atoms can be stored as iron(III). Although ferritins are quite small molecules about 8–12 nm in diameter, they are very effective iron stores. Additionally, they can provide iron on demand. The biosynthesis of ferritin is controlled by the level of iron in the cell via the iron regulatory protein (Andrews et al. 2003). Hemosiderin is another iron-storage complex; however, knowledge of its structure is minimal. It is found solely in cells, in contrast to ferritin, which can also be found in the circulation. It has been suggested that hemosiderin could be a complex of ferritin, denatured ferritin, and some other material. Iron present in hemosiderin deposits is poorly available to provide iron on demand. The storage complex is found in macrophages and appears to be especially abundant following hemorrhage; thus, its formation might be related to phagocytosis of red blood cells and hemoglobin (Trinder et al. 2000).

An interesting feature of ferritin and its relation to iron is that it may be used to differentiate between bacterial and viral infections. Kossiva et al. (2012) showed that ferritin and ferritin to iron ratio (FIR) were significantly higher than iron values in febrile patients than its controls. This could provide an inexpensive, rapid and easily performed complementary index for diagnosis of bacterial infections.

7.5.5 Copper: The Master of Oxidases

Historical records show that copper and copper compounds had been used medicinally at least as early as 400 BC (Mason 1979). Many copper compounds were used to treat a variety of diseases during the nineteenth century, and the presence of copper in plants and animals was recognized more than 150 years ago. For quite some time it has been widely accepted that copper is an essential trace element required for survival by all organisms, from bacterial cells to humans (Linder and Hazegh-Azam 1996). Copper ions undergo a unique chemistry due to their ability to adopt distinct redox states, either oxidized [Cu(II)] or in the reduced state [Cu(I)]. Consequently, copper ions serve as important catalytic cofactors in redox chemistry for proteins that carry out fundamental biological functions; however, copper provides a challenge to biological systems. The very properties that make copper indispensable to biology become toxic when copper is present in excess. Copper's outstanding redox properties, however, such as the transitions between Cu(II) and Cu(I), can in certain circumstances result in the generation of reactive oxygen species such as superoxide radicals and hydroxyl radicals. Susceptible cellular components can be damaged by these reactive species if an effective scavenging mechanism is not in operation. Copper can also bind with high affinity to histidine, cysteine, and methionine residues of proteins. This may result in the inactivation of the proteins (Camarakis et al. 1999). Consequently, there is a great need for effective homeostatic mechanisms controlling the cellular concentration of copper.

7.5.5.1 Copper Proteins

Copper is present in three different forms in proteins: (1) blue proteins without oxidase activity (e.g., plastocyanin), which function in one-electron transfer; (2) non-blue proteins that produce peroxidases and oxidize monophenols to diphenols; and (3) multicopper proteins containing at least four copper atoms per molecule and acting as oxidases (e.g., ascorbate oxidase and laccase). Table 7.12 provides an arbitrary selection of copper-dependent proteins to emphasize the versatility of copper proteins (Peña et al. 1999; Fraústo da Silva and Williams 2001).

The structure and function of type I copper in multicopper oxidases has recently been reviewed. Blue copper proteins show an "alkaline transition" of their absorption wavelength, absorption intensity and redox potential. This phenomenon has not been studied in detail. However, bilirubin oxidase shows a dramatic alkaline transition. The function of type I copper in multicopper oxidases is to withdraw an electron from the substrate and transfer it to the trinuclear copper center (Sakurai and Kataoka 2007).

Table 7.12 Examples of copper-dependent proteins

Protein	Function
Cytochrome oxidase	Reduction of O_2 to H_2O
Laccase	Oxidation of phenols
Ceruloplasmin	Oxidation of Fe(II) to Fe(III), Cu transport
Hemocyanin	Transport of O_2
Lysine oxidase	Cross-linking of collagen
Ascorbate oxidase	Oxidation of ascorbate
Galactose oxidase	Oxidation of primary alcohols to aldehydes in sugars
Amine oxidase	Removal of amines and diamines
Blue proteins	Electron-transfer (many kinds)
Superoxide dismutase	Superoxide dismutation (defense)
Nitrate reductase	Reduction of NO^-_2 to NO
Nitrous oxide reductase	Reduction of N_2O to N_2
Metallothionein	Cu(I) storage
Dopamine monooxygenase	Hydroxylation of Dopa
Co-proporphyrin decarboxylase	Production of protoporphyrin IX
Ethylene receptor	Hormone signaling
Methane oxidase	Oxidation to methanol
Terminal glycine oxidases	Production of signal peptides
Tyrosinase	Melanin production
Clotting factors V and VII	Blood clotting
Angiogenin	Induction of blood vessel formation
Hephaestin	Iron egress from intestines
CP-x type ATPase	Copper pump
Atx-1 (Lys 7)	Copper transfer

7.5.5.2 Intracellular Distribution of Copper

To be distributed in the cell, copper has to be taken up in the cell. This is accomplished by copper transport 1 (Ctr1p), which is a membrane-spanning protein. Ctr1p specifically transports Cu(I), not Cu(II) or any other metals. A conspicuous finding was that the gene *CTR1* (Cu transport 1) was first discovered not as a gene for copper but as a gene essential for iron transport in *Saccharomyces cerevisiae* (Harris 2000). The copper was required for Fet3, a multicopper ferroxidase that catalyzed the oxidation of Fe(II) to Fe(III). Recently, the 3D-structure of the human copper transporter was elucidated and that the transporter has a design similar to a "traditional" ion channel, in which a membrane-spanning pore is created by a symmetrical homotrimer (De Feo et al. 2009).

The intracellular environment is generally reducing. This means that Cu(II) is rapidly reduced to Cu(I). Glutathione was identified early as being involved in copper transport. Glutathione (GSH) is a cysteine-containing tripeptide present more or less exclusively within cells at concentrations in the millimolar range. Copper(I) reacts directly with the internal cysteine sulfhydryl group of glutathione; however, it is less likely that Cu(II) binds to glutathione because of the propensity of Cu(II) to catalyze the oxidation of sulfhydryl groups. The formation of Cu(I)–GSH is a spontaneous reaction probably independent of enzyme involvement. Besides transferring copper to metallothionein, glutathione is required for biliary excretion of copper. In addition, there is evidence that a Cu–GSH complex can mediate stable Cu (I) binding to apocuproproteins (Harris 2000).

Copper chaperones are a family of cytosolic peptides. They are usually referred to as metallochaperones and serve as an intracellular shuttle service for metal ions. In the case of copper, they form transient complexes with Cu (I). These chaperones escort copper ions on their way into proteins that require copper. The prototype of metallochaperones is MerP, a small soluble mercury-binding protein that transports Hg^{2+} to a membrane transporter and eventually to a reductase that reduces Hg^{2+} to the volatile Hg^0 as part of a detoxifying process (Lund and Brown 1987). The best known metallochaperone protein is Atx1 (antioxidant 1), which was originally isolated as an antioxidant protein in *Saccharomyces cerevisiae* and functions in a high-affinity iron uptake pathway in yeast. It works together with Ccc2, a copper-transporting P-type ATPase. The cooperation of these two proteins is necessary for copper loading of the multicopper oxidase Fet3, which is required for the high-affinity iron uptake (Huffman and O'Halloran 2001).

The human homologs of Ccc2 are the Wilson's (Atp7ab) and Menkes' (Atp7a) disease proteins, and the human homolog of Fet3 is ceruloplasmin. Cytochrome oxidase is essential for cellular respiration and requires three copper atoms for

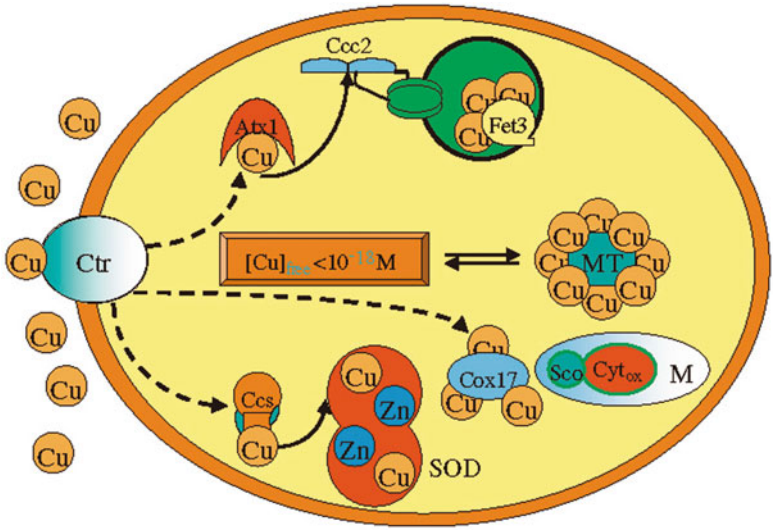


Fig. 7.21 A schematic view of copper localization in the cell

proper function. Two are situated in a dinuclear site in one subunit and one in another subunit next to heme a_3. It has been suggested that Cox17, a small 8-kDa protein, functions in a copper trafficking pathway of cytochrome oxidase (Huffman and O'Halloran 2001). The CCS metallochaperone seems to be necessary for the incorporation of copper into the radical scavenging enzyme Cu,Zn–superoxide dismutase.

Several copper-dependent enzymes like tyrosinase and ceruloplasmin are processed in the Golgi apparatus before being inserted in membranes or secreted from cells. This means that copper is required in the Golgi apparatus. The cytoplasmatic copper chaperone Hah1seems to be involved in the process as well as the two highly homologous transporters ATP7A and ATP7B. One Cu(I) ion is bound by Hah1and delivered to ATP7A and ATP7B, which are able to pump the copper ions into the Golgi organelle for incorporation into copper enzymes (Banci et al. 2010). A schematic view of copper uptake and cellular localization is shown in Fig. 7.21. A few intriguing examples of copper-binding proteins are presented in the next section.

7.5.5.3 Ceruloplasmin

This protein was first isolated from plasma and characterized as a copper-containing protein as early as 1948 (Holmberg and Laurell). It was later discovered that the concentration of this protein was low in patients with Wilson's disease. The first proposed physiological role of the protein was in iron homeostasis and as a ferroxidase (Hellman and Gitlin 2002). Ceruloplasmin belongs to the group of multicopper oxidases. This group of proteins has three distinct copper sites, which are type I copper sites. Charge transfer between the cysteine ligand sulfur and the copper at these sites results in strong absorption at 600 nm. This is why they are called blue proteins. A single copper of type II is coordinated by four imidazole nitrogens and is in close proximity to two antiferromagnetically coupled type III copper ions absorbing at 330 nm. The type II and type III coppers form a trinuclear copper cluster, which is the site of oxygen binding during the catalytic cycle (Hellman and Gitlin 2002).

Multicopper oxidases use the smooth electron chemistry of bound copper ions to couple substrate oxidation with the four-electron reduction of molecular oxygen. Electrons pass from the substrate to the type I copper and then to the trinuclear copper cluster and subsequently to the oxygen molecule bound at this site (Messerschmidt et al. 1989). In addition to ceruloplasmin, several multicopper oxidases have been identified that play a critical role in iron homeostasis. Fet3 is a ferroxidase essential for iron uptake in yeast, and hephaestin is a ceruloplasmin homolog required for efficient iron efflux from the placenta and enterocytes in mammals (Vulpe et al. 1999). Ceruloplasmin is an acute-phase reactant, and its serum concentration increases during inflammation, infection, and trauma largely as a result of increased gene transcription in hepatocytes mediated by the inflammatory cytokines (Hellman and Gitlin 2002).

The liver largely exercises copper homeostasis, and the hepatocytes are the primary site of copper metabolism. Hepatocytes are highly polarized epithelial cells that regulate copper excretion in the bile dependent on the intracellular copper concentration. The copper chaperone Atx1 is required for the delivery of copper to ceruloplasmin; however, the mechanism of copper incorporation into ceruloplasmin is not well understood. Studies of *Saccharomyces cerevisiae* show that both the H^+-transporting V-type ATPase and the CLC chloride channel Gef1 are necessary for copper incorporation in the homologous multicopper oxidase Fet3 (Gaxiola et al. 1998). Ceruloplasmin is capable, *in vitro*, of catalyzing oxidation of a number of different substrates. This has caused some confusion as to the physiologic role of this protein. Ceruloplasmin from human serum has considerable ferroxidase activity necessary for the oxidation of the ferrous iron and

incorporation of ferric iron in apotransferrin (Osaki et al. 1971).

Ceruloplasmin is not only synthesized in the liver but in several other tissues including the central nervous system. The brain is particularly vulnerable to oxidative damage due to its high metabolic rate, high content of oxidizable polyunsaturated fatty acids and the presence of redox-active metals such as iron and copper. In a recent study, it was found that in spite of no significant differences in the concentration of ceruloplasmin, the oxidative activity of ceruloplasmin was significantly higher in patients with Alzheimer's disease and Parkinson's disease as compared to healthy controls (Torsdottir et al. 2010). In addition to this finding, it seems that ceruloplasmin interacts with other proteins participating in inflammation (Vasilyev 2010).

7.5.5.4 Superoxide Dismutase

Superoxide dismutase (SOD) is a member of the oxidoreductase family of enzymes. There are several forms of superoxide dismutase, including Mn–SOD, Fe–SOD, and Cu,Zn–SOD. The Mn–SOD is found exclusively in the mitochondria, whereas the Fe–SOD is generally found in prokaryotes. The Cu,Zn–SOD is active in the cytoplasm of eukaryotes and is the most abundant form of SOD. Cu, Zn–SOD is a dimer of identical subunits with a molecular weight of 16 kDa. One Cu(II) and one Zn(II) are included in each dimer. The role of zinc ions in SOD is structural, and the copper ions take part in the catalytic process. It is based on a redox process, and zinc does not take part in those reactions. SOD catalyzes the disproportion of the superoxide anion ($O_2^{-\bullet}$) to a less dangerous reactive oxygen species. SOD catalyzes the conversion of the superoxide anion according to the following process:

$$Cu^{2+} + O_2^{\bullet -} \rightarrow Cu^{+} + O_2$$
$$Cu^{+} + O_2^{\bullet -} + 2H^{+} \rightarrow Cu^{2+} + H_2O_2$$
$$\text{Net reaction} : 2O_2^{\bullet -} + 2H^{-} \rightarrow H_2O_2 + O_2$$

A Ni-containing SOD was found earlier in *Streptomyces griseus* and *S. coelicolor* (Youn et al. 1996). The structure of this enzyme was revealed later. However, the details of the roles of the protein remain to be elucidated (Abreu and Cabelli 2010). The subunits of Cu,Zn–SOD are stabilized by an intrachain disulfide bond but associated by noncovalent forces. This enzyme requires copper and zinc for its biological activity, and the loss of copper results in its complete inactivation, which can lead to the development of human diseases.

An intriguing connection between mutations in SOD1 and familial amyotrophic lateral sclerosis (ALS) was found about two decades ago. Familial ALS represents about 10% of ALS cases and mutations in SOD1 are the most common cause of familial ALS; they are found in about 20% of ALS patients (Siddique and Ajroud-Driss 2011).

A lot of attempts have been made to establish a link between oxidative stress and schizophrenia. The results are contrasting with either an increased or a decreased activity of SOD in patients (Ciobica et al. 2011).

7.5.6 Zinc: The Ubiquitous Trace Element

Zinc has been known to be essential to life since 1869, when it was discovered that it was required by *Aspergillus niger*. It differs chemically from its neighbors in the transition metal area of the periodic table. Zinc does not take part in redox reactions, but it is a good Lewis acid. In fact, it could be called Nature's Lewis acid. It has a hard metal center and is ideally suited for coordination of N- and O-donors. It is also highly polarizing, and the activity of Zn(II)-dependent enzymes is due to the Lewis acidity of the metal center. In addition, it is characterized by fast ligand exchange. The first enzyme to be recognized as a zinc metalloenzyme was carbonic anhydrase, an enzyme essential for respiration in mammals. At present, zinc metalloenzymes have been recognized in all classes of enzymes. Today, several hundred enzymes are known to be dependent on zinc for catalytic, structural, regulatory, and noncatalytic functions in the animal kingdom. In the realm of plants there are more than 1,000 zinc-dependent enzymes. Examples of enzymes in which zinc plays a catalytic role include carbonic anhydrase, carboxypeptidase, thermolysin, and aldolases. Zinc stabilizes the quaternary structure of oligomeric holoenzymes. It dimerizes *Bacillus subtilus* α-amylase without affecting its enzymatic activity and stabilizes the pentametric quaternary structure of asparatate–-transcarbamylase. Zinc acts as an activator of bovine lens leucine aminopeptidase and inhibits the activity of porcine kidney leucine aminopeptidase and fructose-1,6-bisphosphatase.

The biochemistry of zinc has attracted immense interest during several decades and the section that follows tries to give only a few examples. Even the reasons for zinc selection in biology and its evolutionary history attract increasing interest (Williams 2012).

7.5.6.1 Zinc and Enzymes

Zinc has three types of roles with regard to enzymes: catalytic, cocatalytic, and structural (see Table 7.13). Carbonic anhydrases are a widely expressed family of enzymes that catalyze the reversible reaction $CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+$. These enzymes therefore both produce HCO_3^- for transport across membranes and consume HCO_3^-, which has been transported across membranes (Sterling et al. 2001).

Table 7.13 Examples of zinc metalloproteins

Enzyme	Role of Zinc
Class I: Oxidoreductases	
Alcohol dehydrogenase	Catalytic, non-catalytic
D-Lactate cytochrome reductase	?
D-Lactate dehydrogenase	Catalytic
Superoxide dismutase	Structural (copper catalytic)
Class II: Transferases	
Aspartate transcarbamylase	Structural
Cobalamin-dependent methionine synthase	Catalytic
Cobalamin-independent methionine synthase	Catalytic
DNA polymerase	Catalytic
Mercaptopyruvate sulphur transferase	?
Nuclear poly(A) polymerase	Catalytic
Phosphoglucomutase	?
Protein farnesyltransferase	Catalytic
Reverse transcriptase	Catalytic
RNA polymerase	Catalytic
Terminal dNT transferase	Catalytic
Transcarboxylase	?
Class III: Hydrolases	
α-Amylase	Structural
α-D-Mannosidase	?
β-Lactamase II	Catalytic
Adenosine deaminase	Catalytic
Alkaline phosphatase	Catalytic, noncatalytic
Aminocyclase	?
Aminopeptidase	Catalytic, regulatory
Aminotripeptidase	Catalytic
AMP deaminase	?
Angiotensin-converting enzyme	Catalytic
Astacin	Catalytic
Carboxypeptidase (other)	Catalytic
Carboxypeptidase A	Catalytic
Carboxypeptidase B	Catalytic
Collagenase	Catalytic
Creatinase	?
Cytidine deaminase	Catalytic
D-Carboxypeptidase	Catalytic
DD carboxypeptidase	Catalytic
Dihydropyrimidine aminohydrolase	?
Dipeptidase	Catalytic
Elastase	?
Fructose-1,6-bisphosphatase	Regulatory
Neutral protease	Catalytic
Nuclease P_1	?
Nucleotide pyrophosphatase	Catalytic
Phosphodiesterase (exonuclease)	Catalytic
Phospholipase C	Catalytic
Procarboxypeptidase A	Catalytic
Procarboxypeptidase B	Catalytic
Thermolysine	Catalytic
Class IV: Lyases	
δ-Aminolevulinic acid dehydratase	Catalytic

(continued)

Table 7.13 (continued)

Enzyme	Role of Zinc
Carbonic anhydrase	Catalytic
Fructose-1,6-bisphosphatase adolase	Catalytic
Glycoxalase	Catalytic
L-Rammulose-1-phosphate adolase	Catalytic
Class V: Isomerases	
Phosphomannose isomerase	?
Class VI: Ligases	
Pyruvate carboxylase	?
TRNA synthetase	Catalytic

Source: Data from Prasad (1995) and McCall et al. (2000)

Erythrocytes of mammals have two isoenzymes of carbonic anhydrase: CAI, which has a low activity, and CAII, which has high activity. The molecular weight of human CAI is 30,000 Da, and it contains one atom of zinc per molecule. Human CAII has 259 amino acids, while human CAI has 260 amino acids, and the two isoenzymes share 60% sequence homology. In carbonic anhydrase, zinc is catalytic.

Recent years have produced an increasing number of results concerning CAs. The CAs being present in prokaryotes and eukaryotes are encoded by four distinct, evolutionary unrelated gene families (Hilvo et al. 2005). New intriguing aspects of CAs are therapeutic agents based on the inhibition of CAs present in pathogens like *Candida albicans*. In higher plants algae and cyanobacteria, all member of the CA families are present. For example, in *Arabidopsis* 14 different CAs are present (Supuran and Scozzofava 2007).

In addition to the examples of zinc metalloproteins in Table 7.13, zinc enzymes have been found the classes lyases, isomerases and ligases (Andreini and Bertini 2012).

An example of the structural role of zinc in enzymes is aparatate transcarbamylase (ATCase). This enzyme catalyzes the first step in pyrimidine biosynthesis, condensation of aspartate, and carbamyl phosphate. It is an allosterically regulated enzyme, and its activity is inhibited by cytidine triphosphate (CTP) and activated by ATP. Both responses seem to make sense from a physiological perspective. If CTP levels are already high, additional pyrimidines are not needed, and high ATP signals offers both a purine-rich state and an energy-rich cell condition under which DNA and RNA synthesis can be active. There are similarities with hemoglobin in that the allosteric regulation of ATCase involves changes in the quaternary structure of the molecule. A change in the molecule from the tense state to the relaxed state involves a major rearrangement of subunit positions (Purcarea et al. 1997).

Leucine aminopeptidase (LAP) is a prototypic dizinc peptidase that has been studied intensely. The enzyme is present in animals, plants, and bacteria and has various tissue-specific physiological roles in the processing or degradation of peptides. Human LAP has been shown to catalyze postproteosomal trimming of the N terminus of antigenic peptides for presentation on major histocompatibility complex class I molecules (Beninga et al. 1998; Sträter et al. 1999). The two zinc atoms bound to LAPs work in two different ways: One has a catalytic function, and the other regulates the activity induced by the zinc atom at the first site.

Alcohol dehydrogenase is a zinc-dependent enzyme. At least in some species, zinc may be essential for protection against oxidative damage (Tamarit et al. 1997). In humans, there are at least nine different forms of the enzyme, most of which are found in the liver. It should be noted that an unusual iron- and zinc-containing alcohol dehydrogenase has been identified in the hyperthermophilic archaeon *Pyrococcus furiosus* (Ma and Adams 1999). Alcohol dehydrogenase is our primary defense against alcohol intoxication. It catalyzes the following reaction:

$$\text{Ethanol} \xrightleftharpoons{\text{Alcohol dehydrogenase}} \text{Acetaldehyde} \xrightleftharpoons{\text{Aldehydede hydrogenase}} \text{Acetic acid}$$

In fact, alcohol dehydrogenase catalyzes a transformation to a yet more toxic product, acetaldehyde. So, this toxic molecule is transformed in the next step by aldehyde dehydrogenase to acetic acid and other molecules (Duester 1998) that can be used by the cells. Alcohol dehydrogenase also catalyzes the transformation of retinol in the eye to retinaldehyde and by aldehyde dehydrogenase to retinoic acid. This first line of defense against alcohol, however, is beset with some dangers because alcohol dehydrogenase also modifies other alcohols, often producing dangerous products. For instance, methanol is converted into formaldehyde, which causes damage to proteins and possibly cancer. Small amounts of methanol cause blindness when the sensitive proteins in the retina are attacked, and larger amounts lead to widespread damage and death (Barceloux et al. 2002).

The alcohol dehydrogenases are NAD(H) dependent and have two subunits. Each subunit contains two zinc atoms and

binds one molecule of NAD(H). One zinc atom is essential for the catalytic effect, and the role of the second is largely still unknown; however, it does not seem to be necessary for structure in many cases. There is evidence that zinc may be of importance for the conformational stability of yeast alcohol dehydrogenase (Yang and Zhou 2001).

7.5.6.2 Zinc and Gene Expression

It is by now well established that zinc plays a very important role in gene expression. The importance can be appreciated from the fact that about 25% of the zinc content of rat liver is found in the nucleus, and a significant amount of zinc is incorporated into nuclei *in vitro* (Cousins 1998). One subject of major importance is genetic stability. The correct sequence of nucleotides in DNA is essential for proper replication, gene expression, and protein synthesis. Zinc is involved in the processes of genetic stability and gene expression in a variety of ways, including the structure of chromatin, replication of DNA, and transcription of RNA through the activity of transcription factors and RNA and DNA polymerases, as well as playing a role in DNA repair and programmed cell death (Falchuk 1998).

Until the late 1980s, DNA-binding proteins were not well represented among the nearly 300 zinc-containing proteins known at that time. The multisubunit bacterial RNA polymerases were found to be zinc dependent in the early 1970s. Following these findings, the eukaryotic RNA polymerases, containing many more subunits than the bacterial enzymes, were also found to be zinc enzymes. Replicative DNA polymerases are essential for the replication of the genome of all living organisms. They catalyze the chemical reaction of DNA synthesis:

$$\text{template} - \text{primer} - (\text{dNMP})_n + \text{dNTP} \rightarrow \text{template} - \text{primer} - (\text{dNMP})_{n+1} + \text{PP}_i$$

where dNMP and dNTP are deoxynucleoside 5′-monophosphate and 5′-triphosphate, respectively. During the reaction, inorganic pyrophosphate is released. The reaction requires a 3′-hydroxyl group of the primer for the nucleophilic attack on the α-phosphate of the incoming dNTP. The released pyrophosphate is hydrolyzed, thus providing energy for the reaction.

The sequence similarities may be used to classify them into three types. Type A polymerases are homologous to bacterial polymerases, type B includes archaeal DNA polymerases and eukaryotic DNA polymerase α, and type C is made up of the bacterial polymerase III class. The catalytic mechanism of all three types involves two metal-binding acidic residues in the active site.

One very important aspect of zinc is the so-called zinc fingers, which were first discovered as the transcription factor IIIA (TFIIIA). TFIIIA is a site-specific DNA-binding protein

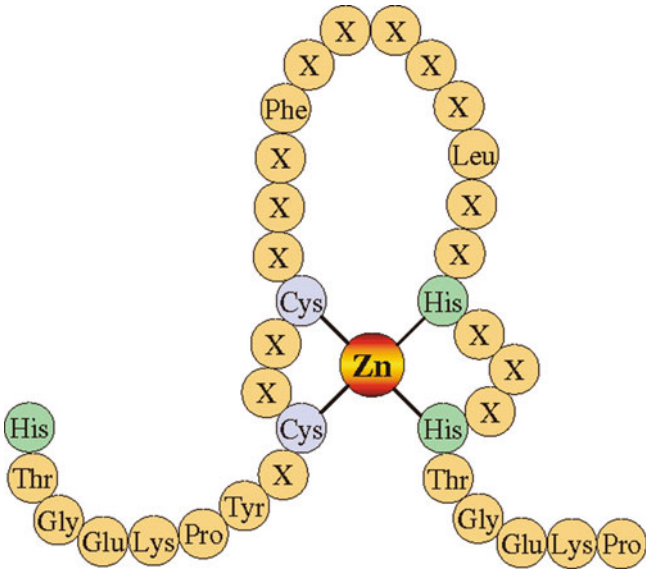


Fig. 7.22 General structure of zinc fingers

that plays a central role in controlling the transcription of 5S ribosomal RNA genes in the African toad *Xenopus laevis*. This protein is slightly unusual because not only does it recognize the internal control region of about 45 bp in the center of the 5S RNA gene, but also TFIIIA itself is bound to the product. The name "zinc fingers" was introduced because of the specific interaction between the amino acids cysteine and histidine and a zinc ion responsible for the formation of the characteristic loop structure (Fig. 7.22). Zinc fingers are generic protein motifs that can mediate DNA-binding and are both widespread and multifunctional. Since first being discovered in the early 1980s, several more zinc-finger proteins have been identified. More than 50 zinc-finger proteins are known today.

Zinc-finger domains are common, relatively small protein motifs that fold around one or more zinc ions. In addition to their role as DNA-binding modules, zinc-finger domains have recently been shown to mediate protein:protein and protein:lipid interactions (Ladomery and Dellaire 2002). This small zinc-ligating domain, often found in clusters containing fingers with different binding specificities, can facilitate multiple, often independent intermolecular interactions between nucleic acids and proteins. Classical zinc fingers, typified by TFIIIA, ligate zinc via pairs of cysteine and histidine residues, but there are at least 14 different classes of zinc fingers, which differ in the nature and arrangement of their zinc-binding residues. Some types of zinc fingers can bind to both DNA and a variety of proteins. Thus, proteins with multiple fingers can play a complex role in regulating transcription through the interplay of these different binding selectivities and affinities (Ladomery and Dellaire 2002). Other zinc fingers have

more specific functions, such as DNA-binding zinc-finger motifs in the nuclear hormone receptor proteins and small-molecule-binding zinc fingers in protein kinase C. Some classes of zinc fingers appear to act exclusively in protein-only interactions. It has also been suggested that zinc fingers, in addition to the functions described above, play a protective role through their prevention of chemical attack by, for example, radicals or reactive oxygen species (Dreosti 2001).

The fortuitous confluence of statistical genetics and molecular biology uncovered the role of Fickle fingers in the zinc-finger protein specification of meiotic recombination hotspots in humans. It has been known for quite some time that during mammalian meiosis, double-strand breaks are deliberately made throughout the genome, leading to the exchange of genetic material between copies of chromosomes. Especially, parts of the role of the PRDM9 DNA-binding protein in male infertility have been elucidated (Ségurel et al. 2011).

Zinc-finger nucleases (ZFNs) are a powerful tool that can be used to edit the human genome ad libitum. Phase I clinical trials have been initiated aimed at knocking out the CCR5 receptor in T cells isolated from HIV patients to protect the cells from infection (Rahman et al. 2011). These results point to promising possibilities for therapeutic genome engineering.

The plant homeodomain (PHD) zinc finger is a small protein domain of 50–80 amino acid residues of diverse sequence and the first structure was unveiled in Li et al. (2006). In comparison with other histone binding modules like bromodomains and chromodomains, it is surprising to note that PHD fingers, which are the smallest in size, are capable of exerting such complex and sophisticated functional versatility as epigenome readers (Sanchez and Zhou 2011).

7.5.6.3 Zinc and Metallothionein

Metallothionein is a generic name for a superfamily of ubiquitous proteins or polypeptides possessing sulfur-based metal clusters. These proteins consist of a single polypeptide chain of 60–68 amino acids that contains 20 highly conserved cysteine residues that, in turn, contain several bivalent cations such as zinc bound through metal–thiolate linkages. These clusters are usually formed through the preferential coordination of d^{10} metal ions by the cysteine thiolate ligands. Currently, four isoforms of metallothionein have been identified. Not all of them are expressed in all organs of mammals. The human metallothioneins are a family of at least 17 gene products. Quite substantial differences exist among species. Metallothionein isoforms I and II have a ubiquitous tissue distribution, with particular abundance in liver, pancreas, intestine, and kidney, whereas isoforms III and IV are found principally in brain and skin (Davis and Cousins 2000).

The zinc coordination in metallothionein is exceptional. The structure is a dumbbell-shaped molecule with two domains, in each of which zinc is bound in a cluster. In one domain, three zinc atoms are bound to nine cysteines; in the other domain, four zinc atoms are bound to 11 cysteines. In this way, each zinc atom is bound tetrahedrally to four cysteines, but overall there are fewer than the maximum number of possible ligands for the seven metals; consequently, some of the cysteines form ligand bridges that form an extensive zinc–sulfur network. The protein envelops the zinc atoms in a manner that effectively shields them from the environment and leaves only a few of the sulfur ligands partially exposed to solvent. Because both protein structure and tetrahedral zinc coordination preclude access of ligands to zinc, it would seem that a conformational change of the protein is necessary to release zinc (Maret 2000).

It is challenging that very few suggestions as to the biological functions of metallothionein have emerged, although the protein has been extensively studied for decades. The main consensus seems to be that it has a role in the detoxification of metal. The biological function of metallothionein is likely related to the physiologically relevant metals that it binds. Recent studies have produced strong evidence to support the idea that metallothionein functions as a metal chaperone for the regulation of gene expression and for the synthesis and functional activity of metalloproteins and metal-dependent transcription factors (Sun and Kang 2002). Vital roles for this pleiotropic protein in more primitive life forms involve the sequestration of environmental metals such as cadmium and mercury.

It has been suggested recently that a biological function of metallothionein is to provide redox functions to the cells. The association of zinc with only sulfur ligands and the biological significance of the peculiar cluster of metallothionein and its purpose have not been elucidated in detail; however, Maret (2000) provided a chemical solution to this challenge: The cluster unit operates via a novel mechanism that allows the cysteine sulfur ligands to zinc to be oxidized and reduced with the concomitant release and binding of zinc. This results in an oxidoreductive process exercised by the ligands of the otherwise redox-inert zinc atom. Thus, metallothionein can become a redox protein in which the redox chemistry originates not from the metal atom but rather from its coordination environment. In this principle of modulating zinc affinity lies the significance of the sulfur coordination chemistry of zinc in the cluster structure of metallothionein.

Although zinc, copper, cadmium, mercury, gold, and bismuth are all metals that induce metallothionein, zinc is the primary physiological inducer. Zinc and copper are essential trace elements, and the other metals are environmental toxicants; however, copper in nontoxic concentrations does not induce metallothionein (Coyle et al. 2002). Metal regulation of metallothionein genes is a complex process involving several steps. In short, the binding of zinc to metal transcription factor 1 (MTF-1) allows the protein to bind to metal-responsive elements (MREs) in the promotor region, which

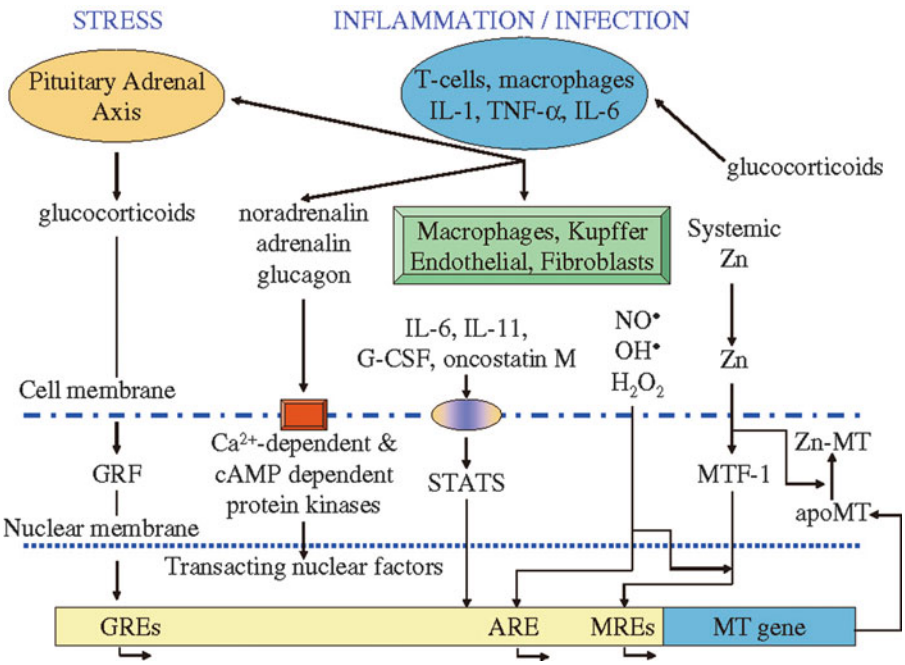


Fig. 7.23 A model for the complex interactions both extra- and intracellularly controlling the expressions of MT (Adapted from Coyle et al. 2002)

subsequently initiates metallothionein gene transcription. It has been proposed that MTF-1 regulates the free zinc concentration by controlling the expression of metallothionein as well as that of a zinc-transporter protein, ZnT-1. Basal expression of MTF-1 may be controlled by a zinc-sensitive inhibitor, which prevents MTF-1 binding to MREs. Zinc dissociates the inhibitor from MTF-1, thereby promoting transcription of metallothionein. MTF-1 is important in the regulation of a number of genes that play a role in cellular response to various stresses (Coyle et al. 2002). Figure 7.23 illustrates a model of the complex interactions controlling the expression of metallothionein.

In humans, there are four major metallothionein family members, the genes of which are tightly clustered on the q13 region of chromosome 16: metallothionein-1 with seven isoforms, metallothionein-2, metallothionein-3 and metallothionein-4. Metallothioneins 1 and 2 are expressed in most tissues, while the expression of metallothionein-3 is restricted to the central nervous system and form four to the squamous epithelia (West et al. 2008). Zinc-thiolate clusters are redox active and that make metallothioneins an attractive molecular sensor for oxidized dopamine. However, the events that result in mobilization of Zn^{2+} from intracellular metallothionein have yet to be characterized in detail. Understanding the process by which cellular defense mechanisms are activated may provide new strategies for prevention, intervention and therapeutic management of disease states in which dopamine neurotoxicity is implicated, including Parkinson's disease (Eibel et al. 2010).

7.5.6.4 Zinc and Inflammation

There is no single factor regulating metallothionein synthesis in inflammation; instead, a complex interrelationship exists between factors that, in combination and in different tissues, act synergistically on metallothionein gene transcription. Nucleotide sequences other than MREs respond to glucocorticoids, interleukin-6 (IL-6), phorbol esters, and hydrogen peroxide. Many of the acute-phase proteins appear to be regulated by combinations of the same factors that include catecholamines and glucocorticoids as well as the cytokines IL-6, IL-1, TNF-α, and γ-interferon. Unlike other acute-phase proteins, metallothionein induction by inflammatory mediators has been found to be conditional upon the presence of zinc. Reactive oxygen species generated during the inflammatory response may induce metallothionein through multiple pathways, including directly stimulating an antioxidant response element and specific MREs in the promoter region as well as by events associated with various second-messenger protein kinase pathways.

Alzheimer's disease (AD) is a neurodegenerative disease characterized by a progressive loss of memory and cognition. The hallmarks of AD are the deposition of beta-amyloid (Aβ) and neurofibrillary tangles. Oxidative stress and inflammation have been implicated. Aβ is neurotoxic and can evoke mitochondrial as well as endoplasmic reticulum stress. A recent study showed that metallothionein expression in reactive astrocytes in a mouse model of AD. Astrocyte-derived metallothionein seemed to be neuroprotective by modulating multiple events associated with Aβ pathology (Kim et al. 2012).

7.5.6.5 Zinc, Insulin, and Diabetes

Several forms of disordered glucose metabolism are collectively referred to as diabetes. Although all diabetes mellitus (mellitus = "sweet as honey") syndromes have some degree of hyperglycemia in common, this is a symptom rather than the metabolic error itself. In insulin-dependent diabetes mellitus (IDDM), there is a destruction of the beta cells of the islets of Langerhans in the pancreas, most often on an autoimmune basis, which results in no insulin being produced. Without insulin, muscle, fat, and liver cells cannot transport glucose from the blood to the intracellular space. Intracellular starvation ensues, with fats becoming the primary intracellular energy source. This form of energy generation results in the production of ketone bodies and organic acids, primarily acetoacetic and beta hydroxybutyric acids, resulting in the development of severe metabolic acidosis (Chausmer 1998).

With non-insulin-dependent diabetes mellitus (NIDDM), the pancreatic islet cells are capable of producing large quantities of insulin, at least at the beginning of the disease. In the healthy individual, insulin binds to a cell membrane receptor and, through several pathways, results in the transport of glucose across the membrane. The intracellular events associated with the activation of glucose transport, after the signal from the insulin–receptor complex is received, are referred to as postreceptor events. To a great extent, it is the failure of the postreceptor events that results in hyperglycemia. In response to the hyperglycemia, the pancreatic islets produce greater and greater quantities of insulin, resulting in downregulation of the number of insulin receptors on the cell membrane and compounding the problem. This results in both hyperglycemia and hyperinsulinemia (Chausmer 1998).

There is an intriguing relation between zinc and insulin regarding the storage of insulin in granulae of the pancreatic islets. Although insulin circulates in the blood and binds to its receptor as a monomer, it forms dimers at micromolar concentrations, and in the presence of zinc ions it further assembles to hexamers. The insulin monomer itself consists of two chains: an A chain of 21 amino acids and a B chain of 30 amino acids. Insulin is synthesized in the β cells of the islets of Langerhans. The β cells are characterized by two features associated with cells that export proteins: rough endoplasmatic reticular surfaces and well-defined storage vesicles. In this case, the vesicle typically contains microcrystals of packaged insulin. The presence of zinc is crucial for the stability of this storage system. On release into the blood, the insulin microcrystals experience a change of pH from about 5.5–7.4. This causes the hexamers and therefore the crystal to disintegrate rapidly (Dodson and Steiner 1998).

The predominant effect on zinc homeostasis of diabetes is hypozincemia, which may be the result of hyperzincuria or decreased gastrointestinal absorption of zinc or both. Whereas the evidence for increased zinc excretion is uniform, the data supporting decreased absorption of zinc are less clear cut. It appears that hyperzincuria is more a result of hyperglycemia than of any specific effect of endogenous or exogenous insulin on the renal tubule (Chausmer 1998).

The zinc–metallothionein complex in the islet cells may provide protection against radicals produced in the cell from any cause, and certainly the immune-mediated, cytokine-provoked oxidative stress would be a significant oxidative stress. The more depleted the intracellular zinc stores, the less able the cell is to defend itself against this oxidative load. This provides a potential mechanism for zinc deficiency to affect the progress of IDDM. With NIDDM, there is no good evidence for oxidative stress as a major factor in the development of either insulin deficiency or islet cell damage; however, there is clear evidence for increased secretion of insulin, at least early in the progress of the disease. Because zinc leaves the cell with insulin, the greater secretion of insulin causes depletion of zinc. The cell can make more insulin, but it cannot make more zinc. With hypozincuria and decreased retention, the zinc is more likely to be excreted and not available for reuptake into the cellular pool. Zinc deficiency may therefore negatively affect the progress of NIDDM (Chausmer 1998).

Recent results have implicated a member of the zinc transporter (ZnT/Slc30) family, ZnT8, in the development of type 2 diabetes (NIDDM) in man. In both pancreatic islet cell types, ZnT8 immunoreactivity is confined to the secretory granule, where ZnT8 is responsible for the accumulation of zinc. ZnT8 expression predominates over other ZnT family members. Hence changes in the ZnT8 abundance or activity seem likely to affect zinc accumulation and thereby insulin storage or release (Rutter 2010).

7.5.6.6 Zinc and the Immune System

It is by now well recognized that nutritional factors are important for the function of the immune system. In the case of zinc, the situation currently of interest is its deficiency and subsequent adverse effects on immune functions. Studies in young adult mice have shown greatly depressed responses to both T-lymphocyte-dependent and -independent antigens. Both primary and secondary antibody responses have been reported to be lowered in zinc-deficient mice. Declines in *in vivo*-generated cytotoxic T-killer activity as well as decreased natural killer (NK) cells have been reported in zinc-deficient mice. All these effects of zinc deficiency on immune functions in mice can be reversed with zinc supplementation (Dardenne 2002).

Malabsorption of zinc occurs in the hereditary disease acrodermatitis enteropathica, in which patients experience thymic atrophy, anergy, reduced lymphocyte proliferative response to mitogens, selective decrease in $T4^+$ helper cells, and deficient thymic hormone activity. All of these

symptoms may be corrected with zinc supplementation. Less severe cellular immune defects have been reported in patients who become zinc-deficient while receiving total parenteral nutrition. Controversial and partly contradictory results have been obtained when zinc intakes were high. In experimental models, high-zinc diets have been shown to reinforce immune functions above basal levels. Other studies have demonstrated the adverse effects of zinc excess; therefore, caution should be exercised when taking large zinc supplements for prolonged periods of time (Dardenne 2002).

Several possible hypotheses can be offered regarding the mode of action of zinc on immune function. Zinc may be necessary for the activity of some immune system mediators such as thymulin, a nonapeptidic hormone secreted by thymic epithelial cells that requires the presence of zinc for its activity. This peptide promotes T-lymphocyte maturation, cytotoxicity, and IL-2 production. Thymulin activity in both animals and humans is dependent on plasma zinc concentrations. Zinc could also be critical for the activity of some cytokines; for example, it has been demonstrated that the production or activity of IL-1, IL-2, IL-3, IL-4, IL-6, IFN-γ, and TNF-α are affected by zinc deficiency. Zinc could contribute to membrane stabilization by acting at the cytoskeletal level. Additionally, zinc is a major intracellular regulator of lymphocyte apoptosis. Thus, it is becoming evident that the thymic atrophy and lymphopenia that accompany zinc deficiency are primarily due to an alteration in the production of lymphocytes and the loss of precursor cells via an apoptotic mechanism (Dardenne 2002).

Vitamin C and zinc are required to support the functions of innate immunity. Zinc and vitamin C both support epithelial barriers, although by different mechanisms. They target different populations of phagocytic cells and thereby ensure an effective phagocytic response. Regarding adaptive immunity zinc is the key player. It is essential for the process that causes stems cells in the bone marrow to form lymphocytes and for the subsequent differentiation into B- and T-lymphocytes (Maggini et al. 2010).

Hematopoiesis is controlled by a number of regulatory networks, including a series of specific transcription factors. The Ikaros family of transcription factors is characterized by two sets of highly conserved zinc-finger motifs. Ikaros proteins can function like archetypal transcription factors to activate specific genes. In addition, they can also repress target genes by associating them to transcriptionally inactive pericentric heterochromatin (John and Ward 2011).

7.5.7 Selenium and Iodine: Young and Old Trace Elements

These trace elements are found in the non-metal area of the periodic table. Although they are neighbors, significant differences exist in their chemical behavior. Selenium is a nonmetal with semiconductor properties, and iodine is a halogen. The biological history of iodine can be traced back to the beginning of the nineteenth century when a physician named Jean-Francois Coindet (1821) used various iodine solutions to treat goiter. Another halogen, although more reactive than iodine, fluorine was also quite early connected to goiter. It was shown that feeding a dog with sodium fluoride caused goiter to appear (Maumené 1854). The human essentiality of iodine was established in 1850.

The scientific community's appreciation of the trace element selenium has more or less undergone a metamorphosis. The toxic effects of selenium were first discovered in the 1930s when livestock ate certain plants with unfortunate results. This problem, mistakenly called "alkali disease," occurred in an acute form following the consumption by range animals of some wild vetches of the genus *Astragalus*, which accumulated toxic amounts of selenium from the soil (Moxon 1937). An historical aside is that it is thought that General Custer might have survived his trip to the Little Bighorn if reinforcements had not been delayed by pack animals that were apparently suffering from selenium-induced lameness. In 1943, selenium was even considered to be a carcinogenic element (Nelson et al. 1943). It was some years before selenium was recognized as an essential trace element (Schwartz and Foltz 1957). In the late 1960s, research suggested an anticarcinogenic effect of selenium (Shamberger and Frost 1969).

7.5.7.1 Iodine Biochemistry

The thyroid, the largest endocrine gland in the body, is located in the neck. The normal gland consists of two lobes connected by a narrow isthmus and is composed of numerous functional units called *follicles*. Only one clearly established function has been demonstrated: to synthesize its hormones, thyroxine (T_4) and 3,5,3′-triiodothyronine (T_3). Iodine is an integral part of T_4 and T_3 that contributes 65 and 59% to their respective molecular sizes (Dunn and Dunn 2001). Figure 7.24 provides an overview of iodine metabolism, which includes thyroglobulin (Tg) iodination, hormone formation, hormone release, and factors that influence these processes. Details regarding iodine metabolism have been known for some time; however, new information still appears quite often in the literature. Of special interest, of course, are the cell trafficking and regulatory aspects.

Iodine provides the raw material for hormone synthesis. In order to accomplish this function, iodine must be transported across the basal membrane of the thyroid by the sodium iodide symporter (Shen et al. 2001). The major factors involved in this process are thyroperoxidase (TPO), hydrogen peroxide, nicotinamide adenine dinucleotide phosphate (NADPH), pendrin, cell-trafficking proteins (the molecular chaperones), and Tg itself (Dunn and Dunn 2001).

Iodine is ingested in a variety of chemical forms. Most ingested iodine is reduced in the gastrointestinal tract and

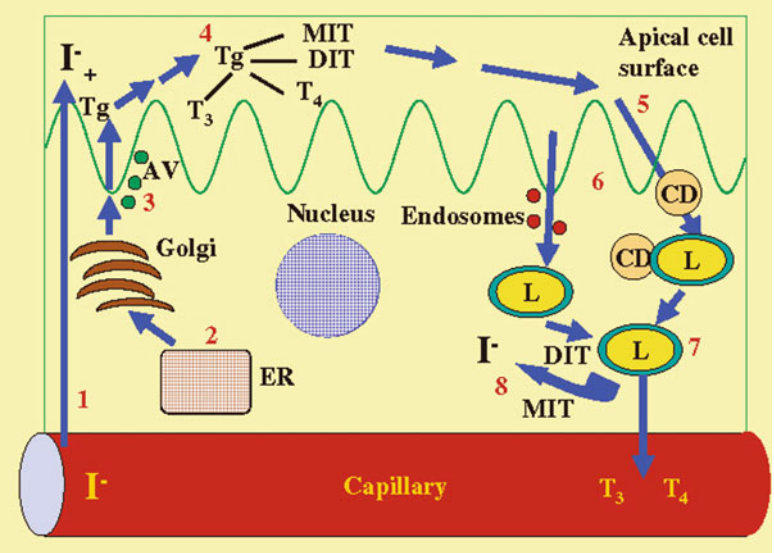


Fig. 7.24 An overview of iodine metabolism

absorbed almost completely. Some iodine-containing compounds (e.g., thyroid hormones) are absorbed intact. Iodate, widely used in many countries as an additive to salt, is rapidly reduced to iodide and completely absorbed. Once in the circulation, iodide is removed by the thyroid gland and the kidney. The thyroid selectively concentrates iodide (see above and Fig. 7.24) in amounts required for adequate thyroid hormone synthesis. Most of the remaining iodine is excreted in the urine. Several other tissues can also concentrate iodine, including salivary glands, breast tissue, choroid plexus, and gastric mucosa. Other than the lactating breast, these are minor pathways of uncertain significance.

The NIS sodium iodide symporter in the thyroidal basal membrane is responsible for iodine concentration. It transfers iodide from the circulation into the thyroid gland at a concentration gradient of about 20–50 times that of the plasma to ensure that the thyroid gland obtains adequate amounts of iodine for hormone synthesis.

Iodine in the thyroid gland participates in a complex series of reactions to produce thyroid hormones. Thyroglobulin, a large glucoprotein weighing 666 kDa, is synthesized within the thyroid cell and serves as a vehicle for iodination. Iodide and Tg meet at the apical surface of the thyroid cell. There, TPO and H_2O_2 promote the oxidation of the iodide and its simultaneous attachment to tyrosyl residues within the Tg molecule to produce the hormone precursors diiodotyrosine and monoiodotyrosine. Thyroperoxidase further catalyzes the intramolecular coupling of two molecules of diiodotyrosine to produce T_4. A similar coupling of one monoiodotyrosine and one diiodotyrosine molecule produces T_3. Mature iodinated Tg is stored extracellularly in the lumen of thyroid follicles, with each consisting of a central space rimmed by the apical membranes of thyrocytes (Dunn and Dunn 2001).

Thyroglobulin, which contains the thyroid hormones, is stored in the follicular lumen until needed. The endosomal and lysosomal proteases digest Tg and release the hormones into the circulation. About two-thirds of Tg iodine is in the form of the inactive precursors monoiodotyrosine and diiodotyrosine. The iodine is not released in the circulation but instead is removed from the tyrosine moiety by a specific deiodinase (see below) and then recycled within the thyroid gland. This process is an important mechanism for iodine conservation, and individuals with impaired or genetically absent deiodinase activity risk iodine deficiency.

Once in the circulation, T_4 and T_3 rapidly attach to several binding proteins synthesized in the liver, including thyroxine-binding globulin, transthyretin, and albumin. The bound hormone then migrates to target tissues where T_4 is deiodinated to T_3, which is the metabolically active form. The responsible deiodinase contains selenium, and selenium deficiency may impair T_4 conversion and hormone action. The iodine of T_4 returns to the serum iodine pool and follows the cycle of iodine again or is excreted in the urine.

Thyroid-stimulating hormone (TSH) is the major regulator of thyroid function. The pituitary secretes this protein hormone, which has a molecular weight of about 28 kDa, in response to circulating concentrations of thyroid hormone: TSH secretion increases when levels of circulating thyroid hormone decrease. TSH affects several sites within the thyrocyte. The principal actions are to increase thyroidal uptake of iodine and to break down Tg in order to release thyroid hormone into the circulation. An elevated serum TSH concentration indicates primary hypothyroidism and a decrease in TSH concentration reflects hyperthyroidism.

Although being known as essential for man since 1850, severe problems with iodine deficiency still exist. Globally, about 2000 million people are affected by iodine deficiency.

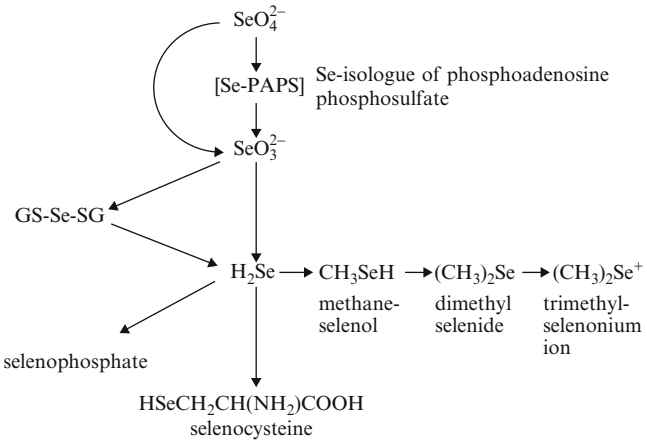


Fig. 7.25 The assimilation of selenate

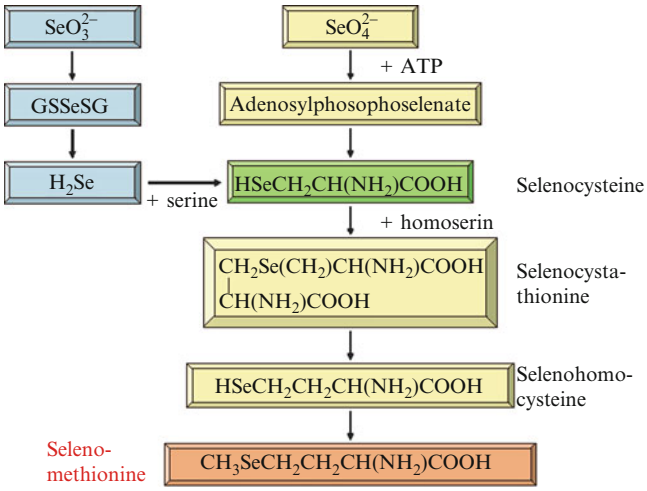


Fig. 7.26 The major pathway of selenomethionine synthesis by plants, marine algae, and yeast

Endemic goiter is perhaps the best known, and visible, sign of iodine deficiency. The most serious consequence of this deficiency, however, is brain damage causing mental retardation in children (Li and Eastman 2012).

7.5.7.2 Selenium Biochemistry

Selenium is primarily taken up from the soil by plants as selenate (SeO_4^{2-}) or selenite (SeO_3^{2-}). The assimilation of selenate appears to follow the sulfate reduction pathway common to higher plants (Fig. 7.25). Analogous to sulfur metabolism, selenate is presumed to be activated by ATP sulfurylase to adenosine phosphoselenate, which then is reduced to selenite. Selenite reacts nonenzymatically with glutathione to form selenodiglutathione, which is readily reduced to selenide by flavine-dependent disulfide reductases, such as glutathione or thioredoxin reductase. In mammals, selenite may also be directly reduced by thioredoxin reductase. Selenide is the central metabolite in the utilization and excretion of selenium. It serves as a substrate for the biosynthesis of selenocysteine by cystein synthases, and it is transformed into selenophosphate, which is required for selenoprotein biosynthesis. Alternatively, it can be methylated by *S*-adenosylmethionine-dependent methyl transferases, which leads to excretion or volatilization of selenium (Ellis and Salt 2003).

Cereals and forage crops convert selenium into mainly selenomethionine and incorporate it into protein in place of methionine because $tRNA^{Met}$ does not discriminate between methionine and selenomethionine. The major pathway of selenomethionine synthesis by plants, marine algae, and yeast is shown in Fig. 7.26. There are some doubts, however, that marine algae can reduce selenate. It is also unclear whether higher animals can make use of selenate without the support of intestinal flora (Birringer et al. 2002). Selenomethionine is not required for growth by plants but is produced along with methionine in quantities depending on the amount of selenium available (Schrauzer 2000). Practically all small organic selenium compounds in plants, bacteria, yeast, and animals are the isologs of corresponding sulfur compounds (Table 7.14). Most of the known selenoproteins contain one or more selenocysteine residues integrated into the main polypeptide chain. The incorporation of selenocysteine is a cotranslational process that makes use of the microsomal machinery of protein synthesis; however, the process is unique in several respects (Birringer et al. 2002).

An intriguing phenomenon, usually referred to as the hierarchy of selenoproteins, has been observed in mammals. It recognizes that the individual selenoproteins respond differently to selenium availability when it is limiting (Birringer et al. 2002).

It is generally assumed that most biological effects of selenium are mediated by selenoproteins. These proteins contain the rare amino acid selenocysteine, which is known as the 21st naturally occurring amino acid. This is cotranslationally incorporated into nascent polypeptide chains. Structurally selenocysteine is similar to cysteine but instead of a sulfur atom it contains a selenium atom. Translation of selenoproteins requires the re-coding of UGA triplets in order to insert the amino acid. The UGA triplet usually works as a stop codon. Although proteins with methionine replaced unspecifically incorporated selenomethionine contain selenium they are not regarded as selenoproteins. The human genome contains 25 genes encoding selenoproteins (Schweizer et al. 2011).

7.5.7.3 Selenoproteins

The selenium-containing proteins identified thus far can be divided into three groups: proteins into which selenium is incorporated nonspecifically, specific selenium-binding proteins, and specific proteins that contain selenium in the

Table 7.14 Distribution (%) of inorganic and organic selenium compounds in garlic and yeast extracts, determined by HPLC-ICP-MS

Compound	Garlic (296mgkg^{-1} Se)	Yeast (1922mgkg^{-1} Se)
Selenate	2	ND
Selenite	ND	1
Selenocystine	0.5	0.5
Selenocystathione	0.5	1
Se-methylated selenocysteine	3	0.5
γ-Glutamyl-Se—methylselenocysteine	73	0.5
Selenomethionine	13	85
γ-Glutamylselenomethionine	4	ND
Se-Adenosylselenohomocysteine	ND	3
Selenolanthionine	ND	1.5
Total selenium (%)	96	93

Source: From Kotrebai et al.(2000)
ND not determined

Table 7.15 Selenoproteins in Prokaryotes

Protein	Function
Glycine reductase	Formation of a selenoether
Glycine/sarcosine/betaine reductase	Redox function
Glycine reductase selenoprotein B	Formation of a selenoether
Sarcosine reductase selenoprotein B	Formation of a selenoether
Betaine reductase selenoprotein B	Formation of a selenoether
Proline reductase	Redox function, formation of a selenoether
Heterosulfide reductase	Redox function
Selenoperoxiredoxin	Redox function (peroxidase)
Putative redox active selenoprotein	Redox function
Formate dehydrogenase	Hydrogen donor
Formylmethanofuran dehydrogenase	Redox function
NiFeSe-hydrogenase	Hydrogen donor
F420 nonreducing hydrogenase	Redox function
F420 reducing hydrogenase	Redox function
Selenophosphate synthetase	Selenoprotein synthesis
CO dehydrogenase	Formation of a carbon oxide selenide
Nicotinic acid hydroxylase	Unknown
Xanthine dehydrogenase	Unknown

Source: Data from Köhrle (2000) and Birringer et al. (2002)

form of genetically encoded selenocysteine and that have been defined as selenoproteins. In addition, there are proteins in which selenium has been demonstrated but for which no information regarding function is yet available (Behne and Kyriakopoulos 2001).

Although there are intriguing differences between the selenoproteins found in prokaryotes and eukaryotes, especially mammals (Tables 7.15 and 7.16), this chapter deals exclusively with mammalian proteins in this context. In fact, at present it seems that selenophosphate synthetase is the only protein belonging both to prokarya and eukarya.

Selenium exerts its biological function through certain proteins containing the element. In this situation, it is in the form of covalently bound selenocysteine. The incorporation of selenium into these proteins requires a set of specific factors. Among others, incorporation of sulfur instead of selenium must be prevented. These elements share similar chemical and physical properties, and sulfur is much more abundant in the biosphere than selenium.

All selenoproteins identified thus far are enzymes in which the selenocysteine residues are responsible for their catalytic functions. Their metabolic importance is based on

Table 7.16 Mammalian selenoproteins

Glutathione peroxidases (GSH-Px)
Cytosolic or classical GSH-Px (cGSH-Px, GSH-Px-1)
Gastrointestinal GSH-Px (GI-GSH-Px, GSH-Px-GI, GSH-Px-2)
Plasma GSH-Px (pGSH-Px, GSH-Px-3)
Phospholipid hydroperoxide GSH-Px (PhGSH-Px, GSH-Px-4)
Iodothyronine deiodinases
5′-deiodinase, type I (5′DI)
5′-deiodinase, type II (5′DII)
5-deiodinase, type III (5-DIII)
Thioredoxin reductases
Thioredoxin reductase (TrxR)
Testicular thioredoxin reductase (TrxR-2)
Mitochondrial thioredoxin reductase (TrxR-3)
Thioredoxin reductase homologues
Selenophosphate synthetase-2
Selenoprotein H
Selenoprotein I
Selenoprotein K
Selenoprotein M and 15-kDa selenoprotein
Selenoprotein P
Selenoprotein R
Selenoprotein S
Selenoprotein W
Functionally undefined
Selenoprotein N
Selenoprotein O
Selenoprotein T
Selenoprotein U
Selenoprotein V

Source: Data from Köhrle (2000), Birringer et al. (2002) and Reeves and Hoffmann (2009)

the fact that in contrast to the thiol in cysteine-containing enzymes, the selenol is fully ionized at normal physiological pH and that under comparable conditions it is of much higher reactivity than the thiol group (Behne and Kyriakopoulos 2001).

7.5.7.4 Glutathione Peroxidases

Glutathione peroxidases (GSH-Px) catalyze the reduction of hydrogen peroxide and organic hydroperoxides. This is an important component of the cellular defense against reactive oxygen species. As can be expected, glutathione usually serves as the electron donor; however, in some cases other thiols are oxidized to fulfill a specific biological function. At present, four selenocysteine-containing GSH-Pxs have been identified (Table 7.16):

- **Cytosolic or classical glutathione peroxidase** – As the name indicates, it is found in the cytosol of cells. It is present in almost all tissues but is not homogeneously distributed. Cytosolic GSH-Px consists of four identical selenocysteine-containing subunits of about 22 kDa. It catalyzes the reduction of hydrogen peroxide and various soluble organic peroxides. In this way, it contributes to the antioxidant defense against reactive molecules and complements the effects of vitamin E. This system seems to be quite insensitive to low activities of cGSH-Px, and it seems that the protective effects of cGSH-Px are of particular importance when the system is exposed to additional stress factors (Behne and Kyriakopoulos 2001). Another important function of this peroxidase is redox control (Schweizer et al. 2011). There are also findings that it reduces retroviral virulence by preventing viral mutations (Rayman 2012).
- **Gastrointestinal glutathione peroxidase** – This enzyme is thought to protect mammals from the toxicity of ingested lipid hydroperoxides. It is similar to cGSH-Px in that it is a cytosolic enzyme consisting of four identical selenocysteine-containing subunits weighing slightly below 22 kDa. In animal studies, selenium deficiency decreases the enzyme activity; however, no effect of human gastrointestinal (GI) GSH-Px has been reported. Gastrointestinal glutathione peroxidase is the most important selenoprotein antioxidant in the colon. Oxidative stress is a critical event in tumorigenesis. It is therefore likely that the antioxidant function of GI-GSH-Px is important in the early defense against colon cancer (Brown and Arthur 2001). A novel aspect of this peroxidase is that it helps to maintain intestinal mucosal integrity by antiapoptotic function in colon crypts (Florian et al. 2010).
- **Phospholipid hydroperoxide glutathione peroxidase** – This protein has been shown to have a membrane-associated as well as a cytosolic location. It is responsible for the reductive destruction of lipid hydroperoxides, and it was the second mammalian selenoenzyme to be discovered. The enzyme is a monomer of 19.7 kDa. The activity of the enzyme is preserved in preference to cGSH-Px when dietary selenium is low. It reacts with phospholipid hydroperoxides as well as small soluble hydroperoxides. It is also capable of metabolizing cholesterol and cholesterol ester hydroperoxides in oxidized low-density lipoprotein. Consequently, it is well recognized as being essential to the destruction of fatty acid hydroperoxides, which, if not reduced to hydroxyl fatty acids, will lead to uncontrolled radical chain reactions that are deleterious to the integrity of membranes (Brown and Arthur 2001). It has also been suggested that the enzyme may have important functions in the redox regulation of a variety of processes, such as inflammation and apoptosis, although it is not known to what extent the other GSH-Pxs are involved in these reactions (Behne and Kyriakopoulos 2001). Of note is that a significant role in spermatogenesis is exclusively fulfilled by this enzyme. It is essential for sperm motility and viability (Noblanc et al. 2011).

- **Plasma glutathione peroxidase** – Extracellular glutathione peroxidase (pGSH-Px) is another protein with antioxidant potential. Again, we are confronted with a tetrameric GSH-Px with subunits of about 23 kDa. The significant difference between pGSH-Px and cGSH-Px, as well as GI-GSH-Px, is that pGSH-Px is a glucoprotein and is present in extracellular fluids. It is secreted into the extracellular fluids from tissues where it is expressed. The kidney is the main site of production for this enzyme. Similar to other tetrameric GSH-Pxs, pGSH-Px catalyzes the reduction of hydrogen peroxide and various organic peroxides when glutathione is used as a substrate. In addition to the antioxidant properties in plasma, pGSH-Px protects the thyroid from hydrogen peroxide in thymocytes and follicular lumen (Rayman 2012). The main source of pGSH-Px in plasma is the kidney, produced by the cells of the proximal tubular epithelium and in the parietal cells of Bowman's capsule and released into the blood (Reeves and Hoffmann 2009).

7.5.7.5 Thioredoxin Reductases

Thioredoxin reductases are a recently identified family of selenoproteins that catalyze the NADPH-dependent reduction of oxidezed thioredoxin and therefore play a regulatory role in its metabolic activity. Thioredoxin is in turn used by several enzymes a cofactor in dithiol-disulfide exchange reactions by which a reduced environment is maintained within cells. This is a family of homodimeric flavoenzymes present in various tissues. In addition to the flavin and the active site of the prokaryotic homologs with their redox-active disulfide, they also contain selenocysteine as the penultimate C-terminal amino acid residue, which is indispensable for their enzymatic activity (Gromer et al. 1998; Reeves and Hoffmann 2009). A description of the thioredoxin reductases identified so far follows:

- **Thioredoxin reductase 1** – TrxR1 was the first thioredoxin reductase to be identified. It is a dimer with two identical 50-kDa subunits. It is a ubiquitous cytosolic enzyme in contrast to the other types. It controls the activity of transcription, cell proliferation and apoptosis. A reduction of expression leads to slower tumor-cell growth (Reeves and Hoffmann 2009).
- **Thioredoxin reductase 2** – The second thioredoxin reductase to be discovered, it was named mitochondrial thioredoxin reductase 2 (TrxR2). The biological role of TrxR2 in the mitochondria is not known, but it seems to be involved primarily in the protection against mitochondria-mediated oxidative stress (Behne and Kyriakopoulos 2001). In addition to this housekeeping activity, it seems to be indispensable for cardiomyocyte viability (Reeves and Hoffmann 2009).
- **Thioredoxin reductase 3** – A third selenocysteine-containing thioredoxin reductase, here named thioredoxin reductase 3 (TrxR3), is preferentially expressed in the testes. The deduced sequence of the human enzyme shows 70% identity to that of TrxR1. It contains a long N-terminal extension and has a higher molecular mass (about 65 kDa) than the two other isozymes (Behne and Kyriakopoulos 2001). This protein is part of the thioredoxin system and works as an antioxidant, redox regulation and cell signaling (Pappas et al. 2008).

7.5.7.6 Selenophosphate Synthetase

Selenophosphate synthetase catalyzes the reaction of selenide with AMP. The product, selenophosphate, acts as the selenium donor for the biosynthesis of selenocysteine. In addition to selenophosphate synthetase 1, which contains threonine in its active center, a selenocysteine-containing homolog of about 50 kDa has also been identified in various human and mouse tissues. The detection of a selenoenzyme that is involved in the production of the selenoproteins is of special interest with regard to regulation of mammalian selenium metabolism (Behne and Kyriakopoulos 2001). The eukaryotic selenophosphate synthetases 1 and 2 were first identified in mammals and implicated as essential components in selenoprotein synthesis. Selenophosphate synthetase 2 is a selenoprotein and may thereby autoregulate its own production along with the production of other selenoproteins (Squires and Berry 2008).

7.5.7.7 Iodothyronine Deiodinases

Comparable to the prokaryotic and the methanococcus world, most of the eukaryotic selenocysteine-containing proteins with identified function are also involved in redox reactions (Köhrle 1999). The deiodinases were discovered in the 1990s. The iodothyronine deiodinases are a large group of selenoproteins. Three iodothyronine deiodinases regulate the conversion of thyroxine (T_4) to 3,3′,5-triiodothyronine (T_3): the active thyroid hormone or reverse triiodothyronine (rT_3), the inactive thyroid hormone. Type I deiodinase (IDI) is expressed in liver, kidney, brain, pituitary, and brown adipose tissue of ruminants. Type II deiodinase (IDII) has been present in the brain and pituitary of all species examined thus far and in brown adipose tissue of humans. IDII catalyzes conversion of T_4 to T_3 within tissues that cannot utilize circulating T3. Type III deiodinase (IDIII) converts T_4 to rT_3 and T_3 to diiodothyronine; is found in brain, skin, and placenta; and functions to deactivate thyroid hormones. The role of selenium in iodothyronine deiodinases implies that some of the consequences of selenium deficiency may be directly attributed to disturbances in thyroid hormone metabolism. The type I enzyme deiodinates 4′-*O*-sulfates of T_4 and T_3 at the tyrosyl ring by 5-deiodination. The type II enzyme acts as a heterotrimeric complex of about 200 kDa containing a 29-kDa subunit that interacts with filamentous

actin (Köhrle 2000). In contrast to IDI, IDII is strictly a phenolic-ring 5′-deiodinase that, as a substrate, prefers T_4 to rT_3. The type III enzyme inactivates T_4 and its metabolites by removal of iodine atoms at the tyrosyl ring. The holoenzyme structure of this enzyme, containing a 32-kDa substrate binding subunit, is not yet known (Köhrle 2000).

The content of selenium in the diet may affect human thyroid hormone metabolism. However, co-factors such as iodine deficiency and thiocyanate overload seem to be required for manifestation of diseases like Kashin-Beck disease or endemic myxodematous cretinism. A possible explanation is that optimal function of the deiodinases requires very little selenium intake (Reeves and Hoffmann 2009).

7.5.7.8 Selenoprotein H

This selenoprotein is a 14-kDa thioredoxin fold-like protein. Its expression is widely distributed throughout a variety of tissues. It is localized in the nucleus and a suggested function is to regulate expression levels of genes involved in the *de novo* glutathione synthesis and phase II detoxification in response to redox status (Panee et al. 2007).

7.5.7.9 Selenoprotein I

This protein contains sequence homology to enzymes involved in phospholipid synthesis. It is expressed in multiple tissues and might have an essential role in developed or viability (Horibata and Harabayashi 2007).

7.5.7.10 Selenoprotein K

Selenoprotein K is small 16-kDa protein and localized to the endoplasmatic reticulum membrane. Like several other selenoproteins, the function of selenoprotein K remains unclear. The localization to the endoplasmatic reticulum may suggest a function relevant to this organelle. However, it seems like the selenocysteine-containing portion is located in the cytoplasm (Chen et al. 2006).

7.5.7.11 Selenoprotein M and 15-kDa Selenoprotein

Both proteins are 15 kDa and share 31% sequence identity and are localized to the endoplasmatic reticulum. They have been suggested to play a role in protein-folding in the endoplasmatic reticulum. A link between 15-kDa selenoprotein and cancer has been made in several in vitro studies (Reeves and Hoffmann 2009).

7.5.7.12 Selenoprotein P

The existence of selenoprotein P was reported in 1973 (Burk 1973); however, it was not until ten years later that it was recognized as a selenoprotein. The P stands for its presence in blood plasma. Selenoprotein P has been purified from humans and rats by a procedure that includes immunoaffinity chromatography. Six typical glycosylation sites in the deduced amino acid sequence of the human proteins have been detected; however, no characterization of the bound carbohydrates has been reported. Selenoprotein P is the first and so far the only protein described to contain more than one selenium atom per polypeptide chain. Ten selenocysteine residues were predicted (Persson Moschos 2000) and have later been confirmed (Reeves and Hoffmann 2009). Selenoproteins with known enzymatic activity are redox enzymes that contain selenocysteine in their active sites. A key issue concerning selenoprotein P that remains to be revealed is its catalytic function. It has been suggested that it acts as an extracellular oxidant defense or as a transport protein. It seems less likely to be a transporting protein because the selenium is covalently bound in the protein (Persson Moschos 2000). It is currently unknown whether selenoprotein P is able to react with certain phospholipid hydroperoxides under physiological conditions (Saito and Takahashi 2000). Recent work suggests that selenoprotein P in plasma diminishes the oxidizing and nitrating reactivity of peroxynitrite, a reactive intermediate formed by the reaction of nitrogen monoxide and superoxide anion. Due to the association with endothelial membranes, it has been speculated that endothelial cells are protected against peroxynitrite toxicity by selenoprotein P (Saito and Takahashi 2000).

A lot of results point to that selenoprotein P is a selenium transport protein. However, transport is not the only function carried out by this protein. Other functions are glutathione peroxidase activity, heparin binding and heavy metal complexation. Even Alzheimer's disease seems to be associated with selenoprotein P; analysis of post-mortem tissue from patients with hallmark lesions of this disease demonstrated colocalization of the protein with amyloid plaques. Selenoprotein P seems also to be a good measure of selenium status (Reeves and Hoffmann 2009).

7.5.7.13 Selenoprotein R

Selenoprotein R is a small protein with a molecular mass of about 12 kDa. It contains selenocysteine. Homologs of this protein have been identified in bacteria, archaea, and eukaryotes, and, with the exception of vertebrate selenoprotein R, all homologs contain cysteine in place of selenocysteine (Kryukov et al. 2002). Bioinformatic analyses have suggested a functional linkage of selenoprotein R to a pathway of methionine sulfoxide reduction as well as a role of selenoprotein R in protection against oxidative stress and/or redox regulation of cellular processes (Kryukov et al. 2002). Methionine sulfoxide reduction is an important process by which cells regulate biological processes and cope with oxidative stress (Hoshi and Heinemann 2001). Methionine sulfoxide reductase is a protein that has been known for decades. It is involved in the reduction of methionine sulfoxides in proteins. It has been shown that methionine sulfoxide reductase is only specific for methionine-*S*-sulfoxides. The fact that oxidized methionines occur in a mixture of R and S isomers *in vivo* raised the question of how methionine sulfoxide reductase could be responsible for the

reduction of all protein methionine sulfoxides. The study of Kryukov and coworkers (2002) explained this puzzle. It appeared that a second methionine sulfoxide reductase exists specific for methionine-*R*-sulfoxides. This reductase was selenoprotein R, and, in addition, these researchers showed that it contains zinc.

Selenoprotein R is the only member of the methionine sulfoxide reductase family that contains selenocystein instead of cysteine. Future studies involving selenoprotein R may include its role not only in the mitigation of oxidative damage to methionine residues on proteins, but a role in regulating certain cell signaling molecules through alteration of redox status of specific methionine residues (Reeves and Hoffmann 2009).There are speculation that selenoprotein R may include its role not only in the mitigation of oxidative damage to methionine residues on proteins, but a role in regulating certain cell signaling molecules through alteration of redox status of specific methionine residues (Erickson et al. 2008).

7.5.7.14 Selenoprotein S

This protein is a transmembrane protein located in the endoplasmatic reticulum and plasma membranes and is widely distributed in a variety of tissues. It has been suggested to participate in the removal of misfolded proteins for degradation and to protect cells from oxidative damage and endoplasmatic reticulum stress-induced apoptosis (Kim et al. 2007a, b; Reeves and Hoffmann 2009).

7.5.7.15 Selenoprotein W

Selenoprotein W contains both a selenocysteine residue that is encoded by a UGA codon in the open reading frame of the mRNA as well as a bound glutathione molecule. The protein is localized predominantly in the cytoplasm (Whanger 2000; Dae-won et al. 2002); however, a small portion of the total selenoprotein W is associated with the cell membrane. Selenoprotein W was expressed in all tissues examined in selenium-supplemented animals including muscle, heart, testis, spleen, kidney, intestine, tongue, brain, lung, and liver (Gu et al. 1999). The loss of the protective effect of selenoprotein W against hydrogen-peroxide-induced cytotoxicity in cells treated with an inhibitor of glutathione synthesis indicates that the protein is a glutathione-dependent antioxidant *in vivo* (Dae-won et al. 2002). Glutathione and its redox cycle play a critical role in catabolizing hydrogen peroxide and other peroxides through enzymatic coupling reactions. Additionally, glutathione is important for the detoxification of electrophiles and for protection of the thiol groups from oxidation. Glutathione is also required for regeneration of the glutathione peroxidase and glutaredoxin system (Dae-won et al. 2002).

As mentioned above, selenoprotein W interacts with glutathione and probably serves as an antioxidant in cells. It has been found that the protein is highly expressed in proliferating myoblasts suggesting that it functions in muscle growth and differentiation by protecting the developing myeoloblast from oxidative stress (Reeves and Hoffmann 2009).

7.5.7.16 Selenium and Cancer

Chemoprevention is a recently introduced and strongly growing area within oncology. The number of studies of chemoprevention has increased drastically during recent years. The reason for this is, of course, that the possibility to prevent or hinder cancer is generally attractive. A great number of agents have been tested for prevention of different forms of cancer (Decensi and Costa 2000). Tamoxifen has been used as a chemoprevention agent against breast cancer as well as raloxifen and synthetic retinoides and combinations. Several nonsteroidal antiinflammatory preparations have been used against colorectal cancer, as have cyclooxygenase-2 and prostanoides. Even calcium has been used in this way, as well as α-difluoromethylornithine, which is an irreversible inhibitor of the enzyme ornithine decarboxylase; also, beta-carotene and retinol have been used against lung cancer.

The first intervention study of prevention of human cancer with selenium was performed in Qidong, an area north of Shanghai with a high incidence of primary liver cancer. In an urban area, 20,847 inhabitants received table salt supplemented with 15 ppm of selenium as sodium selenite. Those individuals were thus supplemented with about 30–50 μg of selenium per day over the course of 8 years. Supplementation resulted in a decrease of the incidence of primary liver cancer to 27.2 per 100,000 inhabitants, while it remained at 50.4 per 100,000 inhabitants in four surrounding areas. When selenium was no longer added to the table salt, the incidence began to increase (Yu et al. 1991, 1997). In another trial, 2,474 members of families with a high risk of primary liver cancer were administered 200 μg of selenium as yeast with a high concentration of selenium or placebo. During the 2-year trial, 1.26% of the controls developed primary liver cancer in contrast to 0.69% in the treated group ($p < 0.05$). Out of 226 bearers of HBVsAg, 7 of 113 individuals in the placebo group developed primary liver cancer, while none of the 113 individuals in the treated group developed cancer during the same period ($p < 0.05$).

The effects of supplementation of vitamins and trace elements on cancer incidence and mortality were investigated in an intervention study in Linxian, China (Li et al. 1993). This area is known to have the highest mortalities of esophagus and ventricular cancer. Among those who received supplementation with beta-carotene, vitamin E, and selenium, total mortality was reduced by 9% and cancer mortality by 13%. Mortality due to ventricular cancer decreased significantly (20%), while mortality in other forms of cancer decreased by 19%.

Within this context, prevention of skin cancer assumes a prominent position, although the outcome of chemoprevention

has not always been positive. Two excellently designed studies evaluated the preventive effects of selenium or retinol on skin cancer (Clark et al. 1996; Moon et al. 1997). The selenium study (Clark et al. 1996) showed negative results for the prevention of squamous cell cancer and basilioma. This study involved 1,312 patients that had a history of skin cancer of the non-melanoma type and lived in areas of the United States with a naturally low intake of selenium. The investigation, presented as a post hoc observation, showed a significant preventive effect on prostate, lung, and colon cancer. The cancer incidences were as follows: prostate cancer (selenium, $n = 13$; placebo, $n = 35$; RR (relative risk) = 0.37; 95% CI (confidence interval); 0.18–0.71; $p = 0.002$), lung cancer (selenium, $n = 17$; placebo, $n = 31$; RR = 0.54; 95% CI; 0.30–0.98; $p = 0.04$), and colorectal cancer (selenium, $n = 8$; placebo, $n = 19$; RR = 0.42; 95% CI; 0.18–0.95; $p = 0.03$). The study also showed a decreased total mortality without statistical significance, although the total mortality in cancer decreased significantly.

The relation between selenium and lung cancer was recently reviewed by Fritz et al. (2011). Lung cancer is a leading cause of cancer deaths worldwide. The American Cancer Society projected 159,390 deaths from lung cancer in 2009, accounting for about 28% of all cancer deaths. The conclusions of the review was that selenium may be effective for lung cancer prevention among individuals with lower selenium status, but at present should not be used as a general strategy for lung cancer prevention.

It was shown that female breast-cancer mortalities were significantly lower in populations whose blood selenium levels were in the order of 0.25–0.30 mg mL^{-1} and when adults of average with attained dietary selenium intakes of 200–300 mg per day (Schrauzer and Surai 2009).

There is a lot of controversy as to the preventive effect from selenium in various types of cancer. However, a large review conducted by the Cochrane Collaboration arrived at the following conclusions. There is little evidence, for women, that lower or higher nutritional intake of selenium exhibiting a major impact on cancer risk. For men, however, there is evidence for an inverse association between higher selenium biomarker levels and cancer risk (Dennert et al. 2012).

See also These Chapters. Chapter 5 (Uptake of Elements from a Chemical Point of View) • Chapter 9 (Biological Responses of Elements)

Further Reading

Abreu IA, Cabelli DE (2010) Superoxide dismutases – a review of the metal-associated mechanistic variations. Biochim Biophys Acta 1804:263–274

Andreesen JR, Makdessi K (2008) Tungsten, the surprisingly positively acting heavy metal element for prokaryotes. Ann N Y Acad Sci 1125:215–229

Andreini C, Bertini I (2012) A bioinformatics view of zinc enzymes. J Inorg Biochem 111:150–156

Andrews SC, Robinson AK, Rodriquez-Quinones F (2003) Bacterial iron homeostasis. FEMS Microbiol Rev 27:215–237

Arunan E, Desiraju GR, Klein RA, Sadlej J, Scheiner S, Alkorta I, Clary DC, Crabtree RH, Dannenberg JJ, Hobza PH, Kjaergaard G, Legon AC, Mennucci B, Nesbitt DJ (2011) Definition of the hydrogen bond. Pure Appl Chem 83:1637–1641

Aubin JE, Triffitt JT (2002) Mesenchyman stem cells and osteoblast differentiation. In: Bilezikian JP, Raisz LG, Rodan GA (eds) Principles of bone biology, 2nd edn. Academic, San Diego, pp 59–81

Baker HM, Anderson BF, Baker E (2003) Dealing with iron: common structural principles in proteins that transport iron and heme. Proc Natl Acad Sci U S A 100:3579–3583

Banci L, Bertini I, Cantini F, Ciofi-Baffoni S (2010) Cellular copper distribution: a mechanistic systems biology approach. Cell Mol Life Sci 67:2563–2589

Barañano DE, Snyder SH (2001) Neural roles for heme oxygenase: contrasts to nitric oxide synthase. Proc Natl Acad Sci U S A 98:10996–11002

Barber J (2003) Photosystem II: the engine of life. Q Rev Biophys 36:71–89

Barceloux DG, Bond GR, Krenzelok EP, Cooper H, Vale JA (2002) American academy of clinical toxicology practice guidelines on the treatment of methanol poisoning. J Toxicol Clin Toxicol 40:415–446

Bates DM, Popescu CV, Khoroshilova N, Vogt K, Beinert H, Münck E, Kiley PJ (2000) Substitution of leucine 28 with histidine in the *Escherichia coli* transcription factor FNR results in increased stability of the $[4Fe–4S]^{2+}$ cluster to oxygen. J Biol Chem 275:6234–6240

Beard JL (2001) Iron biology in immune function, muscle metabolism and neuronal functioning. J Nutr 131:568S–580S

Begley TP, Xi J, Kinsland C, Taylor S, MacLafferty F (1999) The enzymology of sulphur activation during thiamine and biotin biosynthesis. Curr Opin Chem Biol 3:623–629

Behne D, Kyriakopoulos A (2001) Mammalian selenium-containing proteins. Annu Rev Nutr 21:453–473

Beinert H (2000) A tribute to sulphur. Eur J Biochem 267:5657–5664

Beinert H, Kennedy MC, Stout CD (1996) Aconitase as iron–sulfur protein, enzyme and iron-regulatory protein. Chem Rev 96:2335–2373

Beninga J, Rock KL, Goldberg AL (1998) Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase. J Biol Chem 273:18734–18742

Berg D, Gerlach M, Youdim MBH, Double KL, Zecca L, Riederer P, Becker G (2001) Brain iron pathways and their relevance to Parkinson's disease. J Neurochem 79:225–236

Berridge MJ, Lipp P, Bootman MD (2000) The versatility and universatility of calcium signalling. Nat Rev Mol Cell Biol 1:11–21

Birringer M, Pilawa S, Flohé L (2002) Trends in selenium biochemistry. Nat Prod Rep 19:693–718

Brown KM, Arthur JR (2001) Selenium, selenoproteins and human health: a review. Public Health Nutr 4:593–599

Burk RF (1973) Effect of dietary selenium level on Se binding to rat plasma proteins. Proc Soc Exp Biol Med 143:719–722

Camarakis J, Voskoboinik I, Mercer JF (1999) Molecular mechanisms of copper homeostasis. Biochem Biophys Res Commun 261:225–232

Carafoli E, Klee C (1999) Calcium as a cellular regulator. Oxford University Press, Oxford

Chausmer AB (1998) Zinc, insulin and diabetes. J Am Coll Nutr 17:109–115

Chen CL, Shim MS, Chung J, Yoo HS, Ha JM, Kim JY, Choi J, Zang SL, Hou X, Carlson BA, Hatfield DL, Lee BJ (2006) G-rich, a

drosophila selenoprotein, is a Golgi-resident type III membrane. Biochem Biophys Res Commun 348:1296–1301
Chen TC, Sakaki T, Yamamoto K, Kittaka A (2012) The roles of cytochrome P450 enzymes in prostate cancer development and treatment. Anticancer Res 32:291–298
Cigic B, Pain RH (1999) Location of the binding site for chloride Ion activation of cathepsin C. Eur J Biochem 264:944–951
Ciobica A, Padurariu M, Dobrin I, Stefanescu C, Dobrin R (2011) Oxidative stress in schizophrenia – focusing on the main markers. Psychiatr Danub 23:237–245
Clark LC, Combs GF Jr, Trumbull BW, Slate EH, Chalker DK, Chow J, Davis LS, Glover RA, Graham GF, Gross EG, Krongrad A, Lesher JL Jr, Park HK, Sanders BB Jr, Smith CL, Taylor JR (1996) Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin. A randomized controlled trial. JAMA 276:1957–1963
Coindet JF (1821) Nouvelles recherches sur les effets d'iode, et sur les precautions à suivre dans le traitement du goitre par ce nouveau remède. Ann Chim Phys (Paris) 16:252–266
Conklin DS, McMaster JA, Culbertson MR, Kung C (1992) COT1: a gene involved in cobalt accumulation in *Saccharomyces cerevisiae*. Mol Cell Biol 12:3678–3688
Cousins RJ (1998) A role of zinc in the regulation of gene expression. Proc Nutr Soc 57:307–311
Coyle P, Phicox JC, Carey LC, Rofe AM (2002) Metallothionein: the multipurpose protein. Cell Mol Life Sci 59:627–647
D'Amico S, Gerday C, Feller G (2000) Structural similarities and evolutionary relationships in chloride-dependent alpha-amylases. Gene 253:95–105
Dae-won J, Tae SK, Youn WC, Byeong JL, Ick YK (2002) Selenoprotein W is a glutathione-dependent antioxidant *in vivo*. FEBS Lett 517:225–228
Dardenne M (2002) Zinc and immune function. Eur J Clin Nutr 56: S20–S23
Davis SR, Cousins RJ (2000) Metallothionein expression in animals: a physiological perspective on function. J Nutr 130:1085–1088
De Feo CJ, Aller SG, Siluvai GS, Blackburn NJ, Unger VM (2009) Three-dimensional structure of the human copper transporter hCTR1. Proc Natl Acad Sci U S A 106:4237–4242
Decensi A, Costa A (2000) Recent advances in cancer chemoprevention, with emphasis on breast and colorectal cancer. Eur J Cancer 36:694–709
Dennert G, Zwahlen M, Brinkman M, Vinceti M, Zeegers MPA, Horneber M (2012) Selenium for preventing cancer (review). Cochrane Lib 4:1–135
Dodson G, Steiner D (1998) The role of assembly in insulin's biosynthesis. Curr Opin Struct Biol 8:189–194
Dreosti IE (2001) Zinc and the gene. Mutat Res 475:161–167
Duester G (1998) Alcohol dehydrogenase as a critical mediator of retinoic acid synthesis from vitamin a in the mouse embryo. J Nutr 128:459S–462S
Dunn JT, Dunn AD (2001) Update on intrathyroidal iodine metabolism. Thyroid 11:407–414
Dupont CL, Neupane K, Shearer J, Palenik B (2008) Diversity, functioning and evolution of genes coding for putative Ni-containing superoxide dismutases. Environ Microbiol 10:1831–1843
Eibel JK, Abdallah Z, Ross GM (2010) Zinc-metallothionein: a potential mediator of antioxidant defence mechanisms in response to dopamine-induced stress. Can J Physiol Pharmacol 88:305–312
Eitinger T, Mandrand-Berthelot MA (2000) Nickel transport systems in microorganisms. Arch Microbiol 173:1–9
Ellis DR, Salt DE (2003) Plants, selenium and human health. Curr Opin Plant Biol 6:273–279
Erickson JR, Joiner ML, Guan X, Kutschke W, Yang J, Oddis CV, Bartlett RK, Lowe JS, O'Donnell SE, Aykin-Burns N, Zimmerman MC, Zimmerman K, Ham AJ, Weiss RM, Spitz DR, Shea MA, Colbran RJ, Mohler PJ, Anderson ME (2008) A dynamic pathway for calcium-independent activation of CaM_KII by methionine oxidation. Cell 133:462–474
Falchuk KH (1998) The molecular basis for the role of zinc in developmental biology. Mol Cell Biochem 188:41–48
Finazzo C, Harmer J, Jaun B, Duin EC, Mahlert F, Thauer RK, Van Doorslaer S, Schweiger A (2003) Characterization of the MCRred2 form of methyl-coenzyme M reductase: a pulse EPR and ENDOR study. J Biol Inorg Chem 8:586–593
Fletcher EL, Clark MJ, Furness JB (2002) Neuronal and glial localization of GABA transporter immunoreactivity in the myenteric plexus. Cell Tissue Res 308:339–346
Florian S, Krehl S, Loewinger M, Banning A, Esworthy S, Chu FF, Brigelius-Flohé R (2010) Loss of GPx2 increases apoptosis, mitosis, and GPx1 expression in the intestine of mice. Free Radic Biol Med 49:1694–1702
Fraústo da Silva JJR, Williams RJP (2001) The biological chemistry of the elements: the inorganic chemistry of life, 2nd edn. Oxford University Press, London
Frazzon J, Fick JR, Dean DR (2002) Biosynthesis of iron-sulphur clusters is a complex and highly conserved process. Biochem Soc Trans 30:60–685
Fridovich I (1995) Superoxide radical and superoxide dismutases. Annu Rev Biochem 64:97–112
Fritz H, Kennedy D, Fergusson D, Fernandez R, Cooley K, Seely A, Sagar S, Wong R, Seely D (2011) Selenium and lung cancer: a systematic review and meta analysis. PLoS One 6:1–10
Gaggelli E, Berti F, D'Amelio N, Gaggelli N, Valensin G, Bovalini L, Paffetti A, Trabalsini L (2002) Metabolic pathways of carcinogenic chromium. Environ Health Perspect 110(suppl 5):733–738
Gaxiola RA, Yuan DS, Klausner RD, Fink GR (1998) The yeast CLC chloride channel functions in cation homeostasis. Proc Natl Acad Sci U S A 95:4046–4050
Ghosh M, Grunden AM, Dunn DM, Weiss R, Adams MW (1998) Characterization of native and recombinant forms of an unusual cobalt-dependent proline dipeptidase (prolidase) from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 180:402–408
Gromer S, Wissing J, Behne D, Ashman K, Schirmer RH, Flohe L, Becker K (1998) A hypothesis on the catalytic mechanism of the selenoenzyme thioredoxin reductase. Biochem J 332:591–592
Gu QP, Beilstein MA, Barofsky E, Ream W, Whanger PD (1999) Purification, characterization, and glutathione binding to selenoprotein W from monkey muscle. Arch Biochem Biophys 361:25–33
Hagen H, Marzenell P, Jentzsch E, Wenz F, Veldwijk MR (2012) Aminoferrocene-based produgs activated by reactive oxygen species. J Med Chem 55:924–934
Haile DJ, Rouault TA, Harford JB, Kennedy MC, Blondin GA, Beinert H, Klausner RD (1992) Cellular regulation of the iron-responsive element binding protein: disassembly of the cubane iron–sulfur cluster results in high-affinity RNA binding. Proc Natl Acad Sci U S A 89:11735–11739
Handy RD, Eddy FB, Baines H (2002) Sodium-dependent copper uptake across epithelia: a review of rationale with experimental evidence from gill and intestine. Biochim Biophys Acta 1566:104–115
Harris ED (2000) Cellular copper transport and metabolism. Annu Rev Nutr 20:291–310
Hartmann GC, Klein AR, Linder M, Thauer RK (1996) Purification, properties and primary structure of H2-forming N5, N10-methylenetetrahydromethanopterin dehydrogenase from methanococcus thermolithotrophicus. Arch Microbiol 165:187–193
Hausinger RP (1993) Carbon monoxide dehydrogenase. In: Hausinger RP (ed) The biochemistry of nickel. Plenum Press, New York, pp 107–145
Hellman NE, Gitlin JD (2002) Ceruloplasmin metabolism and function. Annu Rev Nutr 22:439–458

Hille R (2002) Molybdenum and tungsten in biology. Trends Biochem Sci 27:360–367
Hilvo M, Tolvanen M, Clark A, Shen B, Shah GN, Wahed A, Halmi P, Hanninen M, Hamalainen JM, Vihinen M, Sly WS, Parkkila S (2005) Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase. Biochem J 392:83–92
Holley AK, Dhar SK, St. Clair DK (2010) Manganese superoxide dismutase versus p53: the mitochondrial center. Ann N Y Acad Sci 1201:72–78
Holmberg CG, Laurell CB (1948) Investigations in serum copper. II. Isolation of the copper-containing protein and a description of some of its properties. Acta Chem Scand 2:550–556
Horibata Y, Harabayashi Y (2007) Identification and characterization of human ethanolaminephosphotransferase1. J Lipid Res 48:503–505
Hoshi T, Heinemann S (2001) Regulation of cell function by methionine oxidation and reduction. J Physiol 531:1–11
Huffman DL, O'Halloran TV (2001) Function, structure, and mechanism of intracellular copper trafficking proteins. Annu Rev Biochem 70:677–701
Iyengar GV, Kollmer WE, Bowen HJM (1978) The elemental composition of human tissues and body fluids: a compilation of values for adults. Verlag Chemie, Weinheim
Jaiswal JK (2001) Calcium – how and why? J Biosci 26:357–363
John LB, Ward AC (2011) The ikaros gene family: transcriptional regulators of hematopoiesis and immunity. Mol Immunol 48:1272–1278
Kappler U, Dahl C (2001) Enzymology and molecular biology of prokaryotic sulfite oxidation. FEMS Microbiol Lett 203:1–9
Katagiri T, Takahasi N (2002) Regulatory mechanisms of osteoblast and osteoclast differentiation. Oral Dis 8:147–159
Kawakami N, Ueki T, Kanamori K, Matsuo K, Gekko K, Michibata H (2009) A novel vanadium reductase, Vanabin2, forms a possible cascade involved in electron transfer. Biochim Biophys Acta 1794:676–679
Kayne FJ (1971) Thallium (I) activation of pyruvate kinase. Arch Biochem Biophys 143:232–239
Kehres DG, Maguire ME (2002) Structure, properties and regulation of magnesium transport proteins. Biometals 15:261–270
Kehres DG, Maguire ME (2003) Emerging themes in manganese transport, biochemistry and pathogenesis in bacteria. FEMS Microbiol Rev 27:263–290
Kim JY, Tillison KS, Zhou S, Lee JH, Smas CM (2007a) Differentiation-dependent expression of Adhfe1 in adipogenesis. Arch Biochem Biophys 464:100–111
Kim KH, Gao Y, Walder K, Collier GR, Skelton J, Kissebah AH (2007b) SEPS1 Protects RAW264.7 Cells from pharmacological ER stress agent-induced apoptosis. Biochem Biophys Res Commun 354:127–132
Kim J-H, Nam Y-P, Jeon S-M, Han H-S, Suk K (2012) Amyloid neurotoxicity is attenuated by metallthionein: dual mechanisms at work. J Neurochem 121:751–762
Kletzin A, Adams MWW (1996) Tungsten in biological systems. FEMS Microbiol Rev 18:5–63
Kobayashi M, Shimizu S (1999) Cobalt proteins. Eur J Biochem 261:1–9
Köhrle J (1999) The trace element selenium and the thyroid gland. Biochimie 81:527–533
Köhrle J (2000) The deiodinase family: selenoenzymes regulating thyroid hormone availability and action. Cell Mol Life Sci 57:1853–1863
Kossiva L, Gourgiotis ID, Tsentidis C, Anastasiou T, Marmarinos A, Vasilenko H, Sdogou T, Gerogouli H (2012) Serum hepcidin and ferritin to iron ratio in evaluating of bacterial versus viral infections in children. A single center study. Pediatr Infect Dis J 31(8):795–798
Kotrebai M, Birringer M, Tyson JF, Block E, Uden PC (2000) Selenium speciation in enriched and natural samples by HPLC-ICP-MS and HPLC-ESI-MS with perfluorinated carboxylic acid Ion-pairing agents. Analyst 125:71–78
Kruzel ML, Zimecki M (2002) Lactoferrin and immunologic dissonance: clinical implications. Arch Immunol Ther Exp 50:399–410
Kryukov GV, Kuman RA, Koc A, Sun Z, Gladyshev VN (2002) Selenoprotein R is a zinc-containing stereo-specific methionine sulfoxide reductase. Proc Natl Acad Sci U S A 99:4245–4250
Ladomery M, Dellaire G (2002) Multifunctional zinc finger proteins in development and disease. Ann Hum Genet 66:331–342
Lau FC, Bagchi M, Sen CK, Bagchi D (2008) Nutrigenomic basis of beneficial effects of chromium(III) on obesity and diabetes. Mol Cell Biochem 317:1–10
Lebedeva MA, Eaton JS, Shadel G (2009) Loss of p53 causes mitochondrial DNA depletion and altered mitochondrial reactive oxygen species homeostasis. Biochim Biophys Acta 1787:328–334
Lebioda L (2000) The honorary enzyme haemoglobin turns out to be a real enzyme. Cell Mol Life Sci 57:1817–1819
Leroy J (1926) Necessité du magnesium pour la croissance de la Souris. C R Seances Soc Biol 94:431–433
Li YH (2000) A compendium of geochemistry: from solar nebula to the human brain. Princeton University Press, Princeton
Li M, Eastman CJ (2012) The changing epidemiology of iodine deficiency. Nat Rev Endocrinol 8(7):434–440
Li JY, Li B, Blot WJ, Taylor PR (1993) Preliminary report on the results of nutrition prevention trials of cancer and other common disease among residents in Linxian, China. Chin J Pathol 15:165–181
Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature 442:91–95
Linder MC, Hazegh-Azam M (1996) Copper biochemistry and molecular biology. Am J Clin Nutr 63:797S–811S
Lund PA, Brown NL (1987) Role of the merT and merP gene products of transposon Tn501 in the induction and expression of resistance to mercuric ions. Gene 52:207–214
Lyonnet B, Martz F, Martin E (1899) L'emploi Thérapeutique des derives du vanadium. Presse Med 1:191–192
Ma K, Adams WW (1999) An unusual oxygen-sensitive, iron- and zinc-containing alcohol dehydrogenase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 181:1163–1170
Macomber L, Hausinger RP (2011) Mechanisms of nickel toxicity in microorganisms. Metallomics 3:1153–1162
Maggini S, Wenzlaff S, Hornig D (2010) Essential role of vitamin C and zinc in child immunity and health. J Int Med Res 38:386–414
Maret W (2000) The function of zinc metallothionein: a link between cellular zinc and redox state. J Nutr 130:1455S–1458S
Mason KE (1979) A conspectus of research on copper metabolism and requirements of man. J Nutr 109:1079–2066
Matoso CMM, Pombeiro AJL, Fraústo da Silva JJR, Guedes da Silva MFC, Silva JAL, Baptista-Ferreria JL, Pinho-Almeida F (1998) A possible role for amavadine in some amanita fungi. In: Tracey AS, Craus DC (eds) Vanadium compounds – chemistry, biochemistry and therapeutic applications, vol 711, ACS symposium series. American Chemical Society, Washington, DC, pp 186–201
Mauk MR, Smith A, Mauk AG (2011) An alternative view of the proposed alternative activities of hemopexin. Protein Sci 20:791–805
Maumené E (1854) Experiencé pour determiner l'action des fluores sur l'economie animale. C R Acad Sci (Paris) 39:538–539
McCall KA, Huang C, Fierke CA (2000) Function and mechanism of zinc metalloenzymes. J Nutr 130:1437S–1446S
Mendel RR (1997) Molybdenum cofactor of higher plants: biosynthesis and molecular biology. Planta 203:399–405

Mertz W (1998) Review of the scientific basis for establishing the essentiality of trace elements. Biol Trace Elem Res 66:185–191
Mertz W, Toepfer EW, Roginiski EE, Polansky MM (1974) Present knowledge of the role of chromium. Fed Proc Fed Am Soc Exp Biol 33:2275–2280
Messerschmidt A, Rossi A, Ladenstein R, Huber R, Bolognesi M, Gatti G, Marchesini A, Petruzzelli R, Finazzi-Agro A (1989) X-Ray crystal structure of the blue oxidase ascorbate oxidase from zuccnini. Analysis of the polypeptide fold and a model of the copper sites and ligands. J Mol Biol 206:513–529
Michels AJ, Joisher N, Hagen TM (2003) Age-related decline of sodium-dependent ascorbic acid transport in isolated rat hepatocytes. Arch Biochem Biophys 410:112–120
Michibata H, Uyama T, Ueki T, Kanamore K (2002) Vanadocytes: cells hold the Key to resolving the highly selective accumulation and reduction of vanadium in ascidians. Microsc Res Tech 56:421–434
Mitchell AC, Peterson L, Reardon CL, Reed SB, Culley DE, Romine MR, Geesey G (2012) Role of outer membrane C-type cytochromes MtrC and OmcA in shewanella oneidensis MR-1 cell production, accumulation, and detachment during respiration on hematite. Geobiology 10(4):355–370
Mobley HLT, Island MD, Hausinger RP (1995) Molecular biology of microbial ureases. Microbiol Rev 59:451–480
Moon TE, Levine N, Cartmel B, Bangert JL, Rodney S, Dong Q, Peng YM, Alberts DS (1997) Effect of retinol in preventing squamous cell skin cancer in moderate-risk subjects: a randomized, double-blind, controlled trial. Southwest skin cancer prevention study group. Cancer Epidemiol Biomarkers Prev 6:949–956
Moxon AL (1937) Akali disease, or selenium poisoning, Bulletin 331. South Dakota Agricultural Experimental Station, Brookings
Mühlenhoff U, Molik S, Godoy JR, Uzarska MA, Richter N, Seubert A, Zhang Y, Stubbe J, Pierrel F, Herrero E, Horst Lillig C, Lill R (2010) Cytosolic monothiol glutaredoxins function in intracellular iron sensing and trafficking via their bound iron-sulfur cluster. Cell Metab 12:373–385
Nelson AA, Fitzhugh OG, Calvery HO (1943) Liver tumors following cirrhosis caused by selenium in rats. Cancer Res 3:230–236
Nishino T, Okamoto K (2000) The role of the [2Fe–S] cluster centers in xanthine oxidoreductase. J Inorg Biochem 82:43–49
Noblanc A, Kocer A, Chabory E, Vernet P, Saez F, Cadet R, Conrad M, Drevet JR (2011) Glutathione peroxidases at work on epidymal spermatozoa: an example of the dual effect of reactive oxygen species on mammalian male fertilizing ability. J Androl 32:641–650
Okamoto S, Eltis LD (2011) The biological occurrence and trafficking of cobalt. Metallomics 3:963–970
Ortiz de Montellano PR, De Voss JJ (2002) Oxidizing species in the mechanism of cytochrome P450. Nat Prod Rep 19:477–493
Osaki S, Johnson DA, Frieden E (1971) The mobilization of iron from the perfused mammalian liver by a serum copper enzyme, ferroxoidase 1. J Biol Chem 246:3018–3023
Panee J, Stoytcheva ZR, Liu W, Berry MJ (2007) Selenoprotein H is a redox-sensing high mobility group family DNA-binding protein that Up-regulates genes involved in glutathione synthesis and phase II detoxification. J Biol Chem 282:23759–23765
Pappas AC, Zoidis E, Surai PF, Zervas G (2008) Seleoproteins and maternal nutrition. Comp Biochem Physiol B Biochem Mol Biol 151:361–372
Peña MMO, Lee J, Thiele DJ (1999) A delicate balance: homeostatic control of copper uptake and distribution. J Nutr 129:1251–1260
Persson Moschos M (2000) Selenoprotein P. Cell Mol Life Sci 57:1836–1845
Prasad AS (1995) Zinc: an overview. Nutrition 11:93–99
Purcarea C, Hervé G, Ladjimi MM, Cunin R (1997) Aspartate transcarbamylase from the deep-Sea hyperthermophilic archaeon *Pyrococcus abyssi:* genetic, organisation, structure, and expression in *Escherichia coli*. J Bacteriol 179:4143–4157
Putney JW, McKay RR (1999) Capacitive calcium entry channels. Bioassays 21:38–46
Ragsdale SW (1998) Nickel biochemistry. Curr Opin Chem Biol 2:208–215
Rahman SH, Maeder ML, Joung JK, Cathomen T (2011) Zinc-finger nucleases for somatic gene therapy: The next frontier. Hum Gene Ther 22:925–933
Ramirez-Silva L, Ferreira ST, Nowak T, Tuena de Gomez-Puyou M, Gomez-Puyou A (2001) Dimethylsulfoxide promotes K+ -independent activity of pyruvate kinase and the acquisition of the active catalytic conformation. Eur J Biochem 268:3267–3274
Rayman M (2012) Selenium and human health. Lancet 379:1256–1268
Reeves MA, Hoffmann PR (2009) The human selenoproteome: recent insights into functions and regulation. Cell Mol Life Sci 66:2457–2478
Richardson DR, Lane DJ, Becker EM, Whitnall M, Rahmanto YS, Sheftel AD, Ponka P (2010) Mitochondrial iron trafficking and the integration of iron metabolism between the mitochondrion and cytosol. Proc Natl Acad Sci U S A 107:10775–10782
Romani AMP, Maguire ME (2002) Hormonal regulation of Mg^{2+} transport and homeostasis in eukaryotic cells. Biometals 15:271–283
Romão MJ (2009) Molybdenum and tungsten enzymes: a crystallographic and mechanistic overview. Dalton Trans 21:4053–4068
Roy R, Adams MWW (2002) Characterization of a fourth tungsten-containing enzyme from the hyperthermofilic archaeon *Pyrococcus furiosus*. J Bacteriol 184:6952–6956
Rutter GA (2010) Think zinc. New roles for zinc in the control of insulin secretion. Islets 2:49–50
Sager TN, Thomsen C, Valsborg JS, Laursen H, Hansen AJ (1999) Astroglia contain a specific transport mechanism for *N*-acetyl-L-aspartate. J Neurochem 73:807–811
Saito Y, Takahashi K (2000) Selenoprotein P: its structure and functions. J Health Sci 44:409–413
Sakurai H (2002) A new concept: the use of vanadium complexes in the treatment of diabetes mellitus. Chem Rec 2:237–248
Sakurai T, Kataoka K (2007) Structure and function of type I copper in multicopper oxidases. Cell Mol Life Sci 64:2642–2656
Sanchez R, Zhou M-M (2011) The PHD finger: a versatile epigenome reader. Trends Biochem Sci 36:364–372
Schipper HM (1999) Glial HO-1 expression, iron deposition and oxidative stress in neurodegenerative disease. Neurotox Res 1:57–70
Schmitt C, Strazielle N, Richaud P, Bouron A, Ghersi-Egea J-F (2011) Active transport at the blood-CSF barrier contributes to manganese influx into the brain. J Neurochem 117:747–756
Schrauzer GN (2000) Selenmethionine: a review of its nutritional significance, metabolism, and toxicity. J Nutr 130:1653–1656
Schrauzer GN, Surai PF (2009) Selenium in human and animal nutrition: resolved and unresolved issues. A partly historical treatise in commemoration of the fiftieth anniversary of the discovery of the biological essentiality o selenium, dedicated to the memory of Klaus Schwarz (1914–1978) on the occasion of the thirtieth anniversary of his death. Crit Rev Biotechnol 29:2–9
Schwartz K, Foltz CM (1957) Selenium as an integral part of factor 3 against dietary necrotic liver degeneration. J Am Chem Soc 79:3292–3293
Schweizer U, Dehina N, Schomburg L (2011) Disorders of selenium metabolism and selenoprotein function. Curr Opin Pediatr 23:429–435
Ségurel L, Leffler EM, Przeworski M (2011) The case of the fickle fingers: How the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. PLoS Biol 9:e1001211. doi:10.1371/journal.pbio.1001211

Shamberger RJ, Frost DV (1969) Possible protective effect of selenium against human cancer. Can Med Assoc J 100:682
Shen DHY, Kloos RT, Mazzaferri EL, Jhiang SM (2001) Sodium iodide symporter in health and disease. Thyroid 11:415–425
Siddique T, Ajroud-Driss S (2011) Familial amyotrophic lateral sclerosis, a historical perspective. Acta Myol 30:117–120
Sorrentino V, Rizzuto R (2001) Molecular genetics of Ca^{2+} stores and intracellular Ca^{2+} signalling. Trends Pharmacol Sci 22:459–464
Sotiriou S, Gispert S, Cheng J, Wang Y, Chen A, Hoogstraten-Miller S, Miller GF, Kwon O, Levine M, Guttentag SH, Nussbaum RL (2002) Ascorbic-acid transporter Slc23a1 is essential for vitamin C transport into the brain and for perinatal survival. Nat Med 8:514–517
Squires JE, Berry MJ (2008) Eukaryotic selenoprotein synthesis: mechanistic insight incorporating new factors and new functions for old factors. IUBMB Life 60:232–235
Sreedhara A, Cowan JA (2002) Structural and catalytic roles for divalent magnesium in nucleic acid biochemistry. Biometals 15:211–223
Stanely JS, Mock DM, Griffin JB, Zempleni J (2002) Biotin uptake into human peripheral blood mononuclear cells increases early in the cell cycle, increasing carboxylate activities. J Nutr 132:1854–1859
Sterling D, Reithmeier RAF, Casey JR (2001) Carbonic anhydrase: in the Driver's seat for bicarbonate transport. J Pancreas (Online) 2:165–170
Sträter N, Sun L, Kantrowitz ER, Lipscomb WN (1999) A bicarbonate ion as a general base in the mechanisms of peptide hydrolysis by dizinc leucine aminopeptidase. Proc Natl Acad Sci U S A 96:11151–11155
Sun X, Kang YJ (2002) Prior increase in metallothionein levels is required to prevent doxorubicin cardiotoxicity. Exp Biol Med 227:652–657
Supuran CT, Scozzofava A (2007) Carbonic anhydrases as targets for medicinal chemistry. Bioorg Med Chem 15:4336–4350
Tamarit J, Cabiscol E, Aguilar J, Ros J (1997) Differential inactivation of alcohol dehydrogenase isoenzymes in *Zymomonas mobilis* by oxygen. J Bacteriol 179:1102–1104
Tolosano E, Fagoonee S, Morello N, Vinchi F, Fiorito V (2010) Heme scavenging and the other facets of hemopexin. Antioxid Redox Signal 12:305–320
Torsdottir G, Kristinsson J, Snaedal J, Sveinbjörnsdottir S, Gudmundsson G, Hreidarsson S, Jóhannesson T (2010) Case-control studies on ceruloplasmin and superoxide dismutase (SOD1) in neurodegenerative diseases: a short review. J Neurol Sci 299:51–54
Trinder D, Macey DJ, Olynyk JK (2000) The new iron age. Int J Mol Med 6:607–612
Ueki T, Uyama Y, Yamamoto K, Kanamori K, Michibata H (2000) Exclusive expression of transketolase in vanadocytes of the vanadium-rich ascidian, *Ascidia syneiensis samea*. Biochim Biophys Acta 1494:83–90
Ueki T, Furuno N, Michibata H (2011) A novel vanadium transporter of the Nramp family expressed at the vacuole of vanadium-accumulating cells of the ascidian *Ascidia sydneiensis samea*. Biochim Biophys Acta 1810:457–464
Väänänen K, Zhao H (2001) Osteoclast function. In: Bilezikan JP, Raisz LG, Rodan GA (eds) Principles of bone biology, 2nd edn. Academic, San Diego, pp 127–139
Vasilyev VB (2010) Interactions of ceruloplasmin with other proteins participating in inflammation. Biochem Soc Trans 38:947–951
Venkatchalam K, van Rossum DB, Patterson RL, Ma HT, Gill DL (2002) The cellular and molecular basis of store-operated calcium entry. Nat Cell Biol 4:E263–E272
Vincent JB (2000) The biochemistry of chromium. J Nutr 130:715–718
Vulpe CD, Kuo YM, Murphy TL, Cowley L, Askwith C, Libina N, Gitschier J, Anderson GJ (1999) Hephaestin, a ceruloplasmin homologue implicated in intestinal iron transport, is defective in the sla mouse. Nat Genet 21:195–199
Wang J, Pantopoulos K (2011) Regulation of cellular iron metabolism. Biochem J 434:365–381
Watt RK, Ludden PW (1999) Nickel-binding proteins. Cell Mol Life Sci 56:604–625
West AK, Hidalgo J, Eddins D, Levin ED, Aschner M (2008) Metallothionein in the central nervous system: roles in protection, regeneration and cognition. Neurotoxicology 29:489–503
Whanger PD (2000) Selenoprotein W: a review. Cell Mol Life Sci 57:1846–1852
Whittaker JW (2010) Metal uptake by manganese superoxide dismutase. Biochim Biophys Acta 1804:298–307
Wilks A (2002) Heme oxygenase: evolution, structure, and mechanism. Antioxid Redox Signal 4:603–614
Williams RJP (2012) Zinc in evolution. J Inorg Biochem 111:104–109
Williamson A, Conian B, Hillier W, Wydrzynski T (2011) The evolution of photosystem II: insights into the past and future. Photosynth Res 107:71–86
Yamamoto A, Wada O, Ono T (1987) Isolation of a biologically active Low-molecular-mass chromium compound from rabbit liver. Eur J Biochem 165:627–631
Yang Y, Zhou HM (2001) Effect of zinc ions on conformational stability of yeast alcohol dehydrogenase. Biochemistry (Moscow) 66:47–54
Yocum CF, Pecoraro VL (1999) Recent advances in the understanding of the biological chemistry of manganese. Curr Opin Chem Biol 3:182–187
Youn HD, Kim EJ, Roe JH, Hah YC, Kang SO (1996) A novel nickel-containing superoxide dismutase from Streptomyces spp. Biochem J 318:889–896
Yu SY, Zhu YJ, Li WG, Huang QS, Zhi-Huang C, Zhang QN, Hou C (1991) A preliminary report of the intervention trials of liver cancer in high risk populations with nutritional supplementation in China. Biol Trace Elem Res 29:289–294
Yu SY, Zhu YJ, Li WG (1997) Protective role of selenium against hepatitis B virus and primary liver cancer in qidong. Biol Trace Elem Res 56:117–124
Yusuf M, Fariduddin Q, Hayat S, Ahmad A (2011) Nickel: an overview of uptake, essentiality and toxicity in plants. Bull Environ Contam Toxicol 86:1–17
Zinchuk VX, Kobayashi T, Garcia del Saz E, Seguchi H (1997) Biochemical properties and cytochemical localization of ouabain-insensitive, potassium-dependent *p*-nitrophenylphosphatase activity in rat atrial myocytes. J Histochem Cytochem 45:177–187
Zoroddu MA, Schinocca L, Jankowska-Kowalik T, Koslowski H, Salnikow K, Costa M (2002) Molecular mechanisms in nickel carcinogenesis: modelling Ni(II) binding site in histone 4. Environ Health Perspect 101(suppl 5):719–723

Geological Impacts on Nutrition

8

Gerald F. Combs Jr.

Contents

8.1 Geological Sources of Nutrients

Humans, like all living organisms, biosynthesize the proteins, nucleic acids, phospholipids, and many of the smaller molecules on which they depend for life functions. Their health and well-being also depend on their ability to obtain from their external chemical environments a number of compounds that they cannot synthesize, or that they cannot produce at rates sufficient to support vital functions. Thus, of the large set of bioactive compounds and metabolites called "nutrients," some are referred to as "essential" because they must be obtained from the air (oxygen), diet and water. These include vitamins, some fatty acids, some amino acids, and several mineral elements.[1] Foods contain essential nutrients as a result of the capacities of plants and, in some cases, food animals to synthesize and/or store them. The human body, therefore, consists of substantial amounts of mineral elements (see Table 8.1) obtained mostly from such foods. These elements ultimately come from soils and, in turn, from the parent materials from which soils are derived. Therefore, good mineral nutrition is, in part, a geologic issue.

About 22 mineral elements are known or suspected to be essential for humans and other animals; 16 are accepted as being indispensible from diets. (see Table 8.2). Some are required in fairly large amounts, e.g., grams per kilogram of diet, and are, therefore, referred to as "macronutrients". Others are required in much smaller amounts, e.g., microgram-to-milligrams per kilogram of diet and are referred to as "micronutrients." At least eight mineral elements function physiologically in their simple cationic forms (Ca^{+2}, Mg^{+}, Na^{+}, K^{+}, Fe^{+2}, Cu^{+2}, Zn^{+2}, and Mn^{+}) and can, therefore, be subject to chelation by either intact proteins or a variety of small, organic molecules. Some chelates (e.g., the heme moieties of hemoglobin and myoglobin) are essential in

G.F. Combs Jr. (✉)
Grand Forks Human Nutrition Center, USDA Agricultural Research Service, Grand Forks, ND, USA
e-mail: Gerald.Combs@ars.usda.gov

[1] This term refers to bioactive elements other than C, H and O that can be obtained from the earth.

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_8,

Table 8.1 Nutritionally essential mineral elements in the human body

Element	Typical amount[a]
Ca	1,000 g
P	700 g
Mg	20–28 g
Na	1.3 g
K	110–150 g
Mg	20–28 g
Zn	2–2.5 g
Cu	120 mg
Se	20 mg

[a]70 kg reference

metabolism; some (e.g., amino acids, EDTA) facilitate the absorption, transport and tissue storage of mineral ions. Others (e.g., phytic acid, oxalic acid) can interfere with the enteric absorption of certain essential mineral cations. For example, the transition metal ions (Fe^{+2}, Cu^{+2} and Zn^{+2}) form coordinate covalent bonds with ligands containing the electron-donor atoms N, S and O; Cu^{+2} complexes with the histidinyl imidazole-N; Zn^{+2} binds the cysteinyl sulfhydryl-S; and Fe^{+2}, Cu^{+2} and Zn^{+2} can bind the aspartyl and glutamyl carboxyl-O. Three mineral elements function as anions or in anionic groupings: Cl^{-}, PO_4^{-3}, MoO_4^{-2}. Unlike the other nutritionally important mineral elements, two non-metals, iodine [I] and Se, function in organic metabolites in which they are covalently bound (e.g., iodothyronine, selenocysteine). The biologic significance of these elements is a property of the particular organic species in which the element is present.

Nutritionally important mineral elements occur in soils as silicates (e.g., manganese, zinc, selenium), sulfides (e.g., zinc, selenium, copper, iron), and as the native elemental form (iron). The most abundant of these is iron, the fourth most abundant element in the Earth's crust (see also Chap. 2, this volume). These are taken up by plants, which are consumed directly or indirectly as tissues of animals feeding of those plants. The grazing practices of many livestock allow them to harvest mineral elements over relatively large geographic areas. Nevertheless, for several mineral elements (e.g., selenium, zinc), the contents of both plant- and animal-base foods ten to reflect the mineral contents of the soils of their respective areas of production.

Water supplies are variable in their mineral contents. Most provide only small amounts of essential minerals; however, some can contribute appreciable amounts of certain mineral elements due to natural conditions (e.g., calcium, magnesium, selenium, fluoride, zinc), to intentional addition (e.g., flouride) and/or to leaching from pipes (e.g.,copper). In general, those water-borne mineral elements of greatest health import include calcium, magnesium, sodium, potassium, copper and selenium. Epidemiological studies conducted in several parts of the world have indicated an inverse relationship of water hardness and cardiovascular disease risk (WHO 2005). These benefits are thought to relate to magnesium and/or calcium, as clinical studies have found that these minerals can be effective in reducing hypertension. There is evidence that consumption of drinking water containing flouride reduces risk of dental carries, but that consumption of drinking water containing iron and copper may increase risks of heart attack.

8.2 Mineral Elements Needed for Good Health

Sixteen mineral elements are established as being essential for good health (Table 8.2). These, collectively, have five general physiological roles:

1. Bone and membrane structure: calcium, phosphorus, magnesium, fluoride
2. Water and electrolyte balance: sodium, potassium, chloride
3. Metabolic catalysis: zinc, copper, selenium, magnesium, molybdenum
4. Oxygen binding: iron
5. Hormone effects: iodine, chromium

Although some of these functions are effected by ionic forms of minerals, many are effected by macromolecules in which minerals are bound, either covalently or otherwise. Because these are all critical life functions, the tissue levels of many nutritionally essential mineral elements tend to be regulated within certain ranges despite varying levels of intake by homeostatic control of enteric absorption and tissue storage and/or excretion. For mineral cations such as Cu^{+2} and Zn^{+2}, regulation occurs primarily at the level of enteric absorption. For mineral elements that tend to be highly absorbed (e.g., selenium, boron), homeostasis is achieved by control at the level of excretion, i.e., through the urine, bile, sweat and breath. In the case of iron, access to active forms is regulated by altering the storage of the element in an inactive form, e.g., ferritin. The ability to orchestrate these physiological processes to achieve homeostatic control of cellular access to such mineral elements is an important factor in ameliorating the effects of short-term dietary deficiencies or excesses.

Calcium—The human body contains more than 1 kg of calcium 99% of which is in the skeleton where it serves as the dominant cationic component (26% of dry weight). Bone mineral consists of a complex matrix of plate-like crystals laid down by osteoblasts in or along collagen fibrils in several solid phases: hydroxyapatite ($Ca_{10}[PO_4]_6[OH]_2$), whitlockite ($[Ca,Mg]_3[PO_4]_2$), amorphous $Ca_9(PO_4)_6X$, octacalcium phosphate ($Ca_8H_2[PO_4]_6$-H_2O), and brushite ($CaHPO_4$-$2[H_2O]$). Bone calcium is in constant turnover, with mineralization and mobilization of bone minerals

Table 8.2 Mineral elements known and suspected to be essential for optimal health

Accepted essentials[a]	Suspected essentials[b]	Known or implicated functions
"macronutrient" elements		
Ca		Bone structure; nervous transduction
P		Bone structure; membrane structure; metabolic regulation
Mg		Bone structure; electrochemical regulation; enzyme catalysis
Na		Electochemical regulation; acid–base balance, osmotic control of water distribution
K		Electochemical regulation; acid–base balance, osmotic control of water distribution
Cl		Electochemical regulation; acid–base balance, osmotic control of water distribution
"micronutrient" elements		
Fe		Oxygen transport; electron transport
Cu		Enzyme catalysis
Zn		Enzyme catalysis; protein structure
I		Metabolic regulation
Se		Enzyme catalysis; antioxidant protection, redox regulation; anti-tumorigenic metabolites
Mn		Enzyme catalysis
Mo		Potentiation of insulin action in the maintenance of glucose tolerance
Cr		Enzyme catalysis
F		Protects against dental caries
Co[c]		Single carbon metabolism as active center of the vitamin B_{12} molecule
	Ni	Fetal survival and anemia in experimental animals
	Pb	Anemia in experimental animals
	As	Reproductive function and growth in experimental animals
	B	Bone mineralization in experimental animals
	V	Growth in experimental animals
	Si	Reproductive function and fetal development in animals, calcification in cell culture

[a]Essentiality demonstrated on the basis of specific biochemical functions
[b]Essentiality indicated by physiological impairment correctable by supplementation
[c]The element itself can be used only by ruminants with foregut microflora capable of synthesizing that vitamin. For this reason, Co is considered essential only for ruminants, while the essential form for all non-ruminants including humans is vitamin B_{12}

occurring continually in the healthy bones of both children and adults. This "remodeling" allows the bone to serve as a source of calcium for noncalcified tissues, which mitigates against irregularities in day-to-day calcium intakes. Such homeostatic control maintains calcium in the ranges of 8.5–10.5 $mgdl^{-1}$ in plasma and of 45–225 mg kg^{-1} intracellularly. These levels are regulated by vitamin D metabolites (e.g., 1,25-dihyroxycholecalciferol), which affect the active transport of calcium across the gut, the recovery of calcium by the renal tubule, and the remodeling of bone in processes also involving parathyroid hormone, calcitonin, and estrogen (Hollick 1994; Jones et al. 1998). Only a small fraction of intracellular calcium exists in the ionic form, Ca^{+2}, which functions as a second messenger to signal many key cellular events, e.g., cell volume regulation, fertilization, growth-factor-induced cell proliferation, secretion, platelet activation, and muscle contraction. Therefore, a key aspect of cell regulation is the control of the release of the Ca^{+2} signal. In various cells, this process is thought to involve 1,4-inositoltriphosphate or nervous stimulation as triggers and protein kinase C and cyclic AMP as inhibitors (Bronner 1997). Impaired bone mineralization in young children results in deformities of the growing bones and is called rickets; in adults with formed bone it is called osteomalacia and is characterized by increase fracture risk and loss of stature. Conditions of either type can be caused by insufficient intakes of calcium, vitamin D (or exposure to sunlight which is necessary for the biosynthesis of the vitamin), phosphorus and/or magnesium. Only in very severe deficiency, when bone mineral has largely been exhausted, does calcium deficiency result in impaired nervous conduction or muscular contraction. Excessive calcium intake, which typically occurs due to the inappropriate use of calcium supplements, can lead to renal stone formation, hypercalcemia, and renal insufficiency as well as impaired utilization of iron, zinc, magnesium, and phosphorus.

Phosphorus—The human body contains approximately 700 g of phosphorus, about 85% of which is in bones where it serves a structural function in bone minerals (see also Chap. 30, this volume). Much of the nearly 14% of body phosphorus in noncalcified tissues also serves a structural function in the phospholipids that comprise plasma and subcellular membranes (Berner 1997). As phosphate, the element phosphorus is also important in metabolism because it is incorporated into nucleic acids, RNA, DNA, proteins (including transcription factors), ATP, and numerous other

high-energy substrates. Intracellular phosphate serves as a regulator of glycolysis, a key pathway for rendering oxygen available to the tissues; and phosphoproteins play essential roles in the electron-transport system of mitochondria, which generates metabolically useful energy from carbohydrates and lipids. Other phosphoproteins serve as cellular growth factors and cytokines. Like calcium, phosphorus homeostasis is affected by the vitamin D hormone system. Deficiency of phosphorus (hypophosphatemia) can result in tissue hypoxia due to the loss of erythrocyte 2,3-diphosphoglucose and ATP, which leads to nervous signs (convulsions, confusion), renal dysfunction, and smooth muscle problems (e.g., dysphagia, gastric atony). Hypophosphatemia can also result in rickets and osteomalacia. Chronic, excessive intakes of phosphorus can cause hyperphosphatemia which leads to interference with calcium homeostasis, bone demineralization, and ectopic calcification of the kidney.

Magnesium—The human body typically contains 20–28 g of magnesium, which is widely distributed: 60–65% of that amount is in bone, 25–30% in muscle, and the balance is in other tissue and extracellular fluid. In fact, Mg^{+2} is second only to K^+ as the most abundant intracellular inorganic cation. Normal plasma magnesium concentrations are in the range of 0.65–0.88 mmol L^{-1}. Magnesium tends to be well absorbed (i.e., at rates of 67–70%) by a saturable process as well as simple diffusion. Amounts of the element not retained for tissue growth/turnover are excreted in the urine. The cation, Mg^{+2}, functions as a cofactor in at least 300 enzymatic reactions (Shils 1997). These include virtually all kinase reactions (in which Mg^{+2} complexes with the negatively charged ATP^{-4} to form the substrate), pyrophosphotransferases, acyl-CoA synthetases, and adenylate cyclase. The cation is also involved in the regulation of ion movements within cells. Magnesium deficiency (hypomagnesemia) is characterized by neuromuscular signs (hyperactivity, muscle spasms, tremor, weakness) and gastrointestinal symptoms (anorexia, nausea, vomiting). Magnesium has a cathartic effect, but adverse effects have been identified only for magnesium ingested from non-food sources. Epidemiological studies conducted in several parts of the world have indicated an inverse relationship of water hardness and cardiovascular disease risk (WHO 2005). These benefits are thought to relate to the associated intakes of magnesium and, perhaps, calcium, which have been found in clinical studies to be effective in reducing hypertension presumably through its effects on the function of smooth muscle.

Sodium—The human body contains approximately 1.3 g of sodium, 90% of which is in the extracellular space where it serves as one of the three (with K^+ and Cl^-) osmotically active solutes in extracellular fluid. Sodium is freely and quantitatively absorbed, but the element is homeostatically regulated within a normal range of 135–145 mmol L^{-1} at the level of renal reabsorption effected by renin, angiotensin, aldosterone, antidiuretic hormone, atrial natriuretic peptide, and other factors affecting renal blood flow (Harper et al. 1997). Intracellular Na^+ is normally maintained at relatively low levels. The Na^+ gradient is used as an energy source for the uphill transport of a variety of solutes (e.g., amino acids, Ca^{+2}, Mg^{+2}, H^+) into the cell. The maintenance of the Na^+ gradient is maintained by several transport systems including the Na^+–K^+ pump which effects the ATP-dependent antitransport of Na^+ (in) and K^+ (out), the Na^+–H^+ exchanger, Na^+–K^+–Cl^-co-transporters, the Na^+–Ca^{+2} exchanger, Na^+–Mg^{+2} exchangers, and the voltage-regulated Na^+ channel. Sodium deficiency results in muscle cramps, headache, poor appetite, and dehydration, but the main sign is fatigue.

Potassium—Potassium is the most abundant cation in the human body, with total body stores typically in the range of 110–150 g. In contrast to Na^+, K^+ is found primarily (98%) in the intracellular compartment; most cells contain about 150mMK^+, while the level in extracellular fluid is only about 4 mM. Potassium is freely absorbed, with homeostasis affected by rapid renal excretion. Potassium passes the plasma membrane into cells by the Na^+, K^+-ATPase, the H^+,K^+-ATPase, the Na^+–2Cl^-–K^+ co-transporter, and K^+ conductance channels. Increases in extracellular K^+ concentrations can be caused by vigorous exercise leading to K^+ efflux from myocytes and mediating vasodilation and increased blood flow. Such increases stimulate the release of catecholamines and insulin, which stimulates K^+ uptake via the Na^+, K^+-ATPase (Peterson 1997). Potassium deficiency (hypokalemia) can be caused by insufficient intake and/or excessive excretion (e.g., due to diarrhea, bulimia) of the element. This is characterized by skeletal muscular weakness; smooth muscle paralysis resulting in anorexia, nausea, vomiting, and constipation; cardiac arrhythmias; carbohydrate intolerance due to diminished insulin secretion; impaired renal function due to reduced blood flow; and altered water balance involving increased water consumption secondary to elevated angiotensin II levels.

Chloride—Chloride is the major extracellular anion maintained at a concentration of 100–110 mmol L^{-1} in that fluid. Like Na^+, there is no control over Cl^-absorption, and homeostasis is affected by renal reabsorption/elimination. The transport and cellular uptake of $Cl^-$is effected by a number of transporters including a K^+–Cl^-co-transporter, a Na^+–K^+–2Cl^-co-transporter, Cl^-–HCO_3 exchangers, cystic fibrosis transmembrane conductance regulator (mutation of this causes cystic fibrosis), Ca^{+2}-activated Cl^-channels, voltage-regulated Cl^-channels, and mechanically activated Cl^-channels (Harper et al. 1997).

Iron—The human body typically contains approximately 5 g of iron, and its metabolic function is to transport oxygen and electrons. Iron serves as the redox agent in a large

number of enzymatic reactions involving substrate oxidation and reduction. These include oxidoreductases (e.g., xanthine oxidase/dehydrogenase), monooxygenases (e.g., amino acid oxidases, cytochrome P-450), and dioxygenases (e.g., amino acid dioxygenases, lipoxygenases, peroxidases, fatty acid desaturases, nitric oxide synthases) (Beard and Dawson 1997). Iron homeostasis is effected at the level of enteric absorption. Dietary iron generally exists in either heme (from hemoglobin and myoglobin in animal products) or non-heme (i.e., organic and inorganic salts in plant-based and iron-fortified foods) forms, each of which is absorbed by a different mechanism. Heme iron is much better absorbed and less affected by enhancers and inhibitors of absorption than non-heme iron, which is strongly regulated by the intestinal mucosal cells in response to iron stores and blood hemoglobin status. Thus, iron-adequate men and women typically absorb about 6 and 13% of dietary iron, respectively, with non-heme iron absorption as great as 50% under conditions of severe iron-deficiency anemia. Excess absorbed iron is stored as ferritin and hemosiderin in the liver, reticuloendothelial cells, and bone marrow. The loss of iron from the body is very low, about 0.6 mg per day, and is primarily due to losses in the bile and exfoliated mucosal cells eliminated in the feces. Menstrual losses can be significant, as can nonphysiological losses resulting from parasitism, diarrhea, and enteritis, which are thought to account for half of the cases of global iron-deficiency anemia. Iron deficiency is manifested as hypochromic, normocytic anemia; lethargy; apathy; listlessness; fatigue; impaired non-shivering thermogenesis; impaired immune function; impaired cognitive development; and reduced physical performance. In pregnancy, iron-deficiency increases the risk of premature delivery, low birth weight, and infant and maternal mortality.

Epidemiologic observations have linked high dietary iron intakes or high iron stores with increased risk of coronary heart disease (Salonen et al. 1992). The toxic potential of iron arises from its pro-oxidative effects, which yield reactive oxygen species that attack polyunsaturated membrane lipids, proteins, and nucleic acids. An iron-overload disease, hereditary hemochromatosis, is caused by a defect in the regulation of iron absorption, which leads to very high circulating transferrin iron. Clinical signs appear when body iron accumulates to about tenfold excess of normal: these include hepatic cirrhosis, diabetes, heart failure, arthritis, and sexual dysfunction.

Zinc—The human body contains 2–2.5 g of zinc: just over half (55%) of which is in muscle, 30% in bone, with the balance distributed in other tissues. Zinc is absorbed by both carrier-mediated and simple diffusion processes which render the element only moderately absorbed (about 30%) at nutritionally adequate intakes and more efficiently absorbed under deficient conditions (Chesters 1997). Both processes can be effected by the presence of chelating substances that may promote (e.g., meats) or impair (e.g., phytic acid) zinc absorption. Zinc homeostasis is also affected by the regulation of zinc excretion/reabsorption in and from pancreatic and intestinal secretions; urinary losses of zinc are low and not generally responsive to changes in zinc intake. Plasma zinc levels, comprising 0.1% of total body zinc, are not regulated and are therefore not indicative of overall zinc status except under conditions of marked deficiency. Zinc has been shown to function in at least 50 widely varied enzymes. In each, the element serves either in a catalytic (i.e., at the active site), a co-catalytic (i.e., near the active site), or a structural role bound most commonly to histidinyl, glutamyl, or aspartyl residues. Zinc deficiency is manifested as losses in activities of at least some zinc enzymes (e.g., some dehydrogenases, alkaline phosphatase, superoxide dismutase), although direct links between such losses and the physiologic manifestations of zinc deficiency have not been established. Zinc deficiency is characterized by poor growth and dwarfism, anorexia, parakeratotic skin lesions, diarrhea, impaired testicular development, impaired immune function (including wound healing), and impaired cognitive function. Low zinc status is also thought to increase risk to osteoporosis and susceptibility to oxidative stress. Very high intakes of zinc, which have occurred due to inappropriate use of zinc supplements, can interfere with copper metabolism and deplete the body of copper. Chronic exposure to excess zinc (>100 mg d^{-1}) can reduce immune function and HDL cholesterol.

Copper—The human body contains approximately 120 mg of copper, which is widely distributed in many tissues and fluids at mg kg^{-1} or μg kg^{-1} concentrations. Copper serves as a cofactor for a number of oxidase enzymes including lysyl oxidase, ferroxidase (ceruloplasmin), dopamine beta-monooxygenase, tyrosinase, alpha-amidating monooxygenase, cytochrome *c* oxidase, and superoxide dismutase (Harris 1997). These enzymes are involved in generating oxidative energy, stabilizing connective tissue matrices, maintaining iron in the ferrous (Fe^{+2}) state, synthesizing neurotransmitters, pigmenting hair and skin, supporting immune competence, and protecting the body from reactive oxygen species. Copper also functions non-enzymatically in angiogenesis, neurohormone release, oxygen transport, and the regulation of genetic expression. Copper homeostasis is effected at the level of enteric absorption. It is absorbed by facilitated diffusion (involving either specific transporters or nonspecific divalent metal ion transporters on the brush-border surface) and is transported to the liver where it is re-secreted into the plasma bound to ceruloplasmin. The element is excreted in the bile; only very small amounts are lost in the urine. Copper absorption/retention varies inversely with the level of copper intake, and tends to be moderate (e.g., 50–60%) even at low copper intakes. Copper

deficiency is manifested as hypochromic, normocytic or macrocytic anemia; bone abnormalities resembling osteoporosis or scurvy; increased susceptibility to infection; and poor growth. The ingestion of high amounts of copper can cause nausea. Chronic high copper intake can lead to the hepatic accumulation of copper, which has been suspected in juvenile cases of hepatic cirrhosis in India. There is evidence that consumption of drinking water containing copper may increase risks of heart attack.

Iodine—The human body contains approximately 5 mg of iodine, which functions only in the iodine-containing thyroid hormones. These include the tetraiodinated protein thyroxine (T_4) that is converted by a single deiodination to yield the active thyroid hormone triiodothyronine (T_3). The latter functions as a regulator of growth and development by increasing energy (ATP) production and activating or inhibiting the synthesis of various proteins (Hetzel and Wellby 1997). Organic forms of iodine are converted in the upper gastrointestinal tract to the iodide anion (I^-), which is rapidly and almost completely absorbed. In contrast, when T_3 is ingested, about 80% is absorbed intact. Absorbed I^-circulates in the plasma in the free ionic form and is rapidly removed by the thyroid and kidney. Iodine homeostasis is effected at the level of the kidney, and urinary excretion is the major route of loss (comprising 90% of iodine absorbed by iodine-adequate individuals). The deiodination of T_4 occurs in the thyroid, skeletal muscles, and brain, but the thyroid gland is the only storage site for iodine where it appears mostly as mono- and diiodotyrosine and T_4, with a small amount of T_3. Iodine deficiency in adults is characterized as thyroid hypertrophy or goiter and in children as myxedematous cretinism. Collectively, these iodine deficiency diseases comprise a global health problem, and cretinism is the greatest source of preventable mental retardation (see also Chap. 17, this volume).

Selenium—The human body typically contains approximately 20 mg of selenium. This is widely distributed in all tissues in each of which the element present almost exclusively bound to proteins. Genomic analyses have revealed selenium to be an essential constituent of a relatively small number of proteins of which contain the element in the form of selenocysteine (Kryukov et al. 2003). That amino acid is found only in the selenproteins, as it is incorporated co-translationally by a process signaled by TGA in DNA (UGA in mRNA), which in the synthesis of other proteins functions as a stop codon. Humans have 25 selenoproteins. These include multiple isoforms of glutathione peroxidase, thioredoxin reductase and iodothyronine 5′-deiodinase; the selenium-transporter, selenoprotein P; an enzyme involved in SeCys synthesis, selenophosphate synthase; and others of uncharacterized function. In these selenoproteins, selenium functions in antioxidant protections, redox regulation, and thyroid hormone regulation. When selenium is consumed as selenomethionine (SeMet), the dominant form in plant foods, other proteins can incorporate that form nonspecifically, due to its mimicry with methionine. The non-specific presence of selenium in proteins comprises the dominant portion of tissue selenium, although apparently without physiological significance. Because foods contain both of these selenoaminoacids (selenomethionine in plant tissues; both selenocyteine and selemethionine in animal tissues), in human tissues typically contain both the specific and non-specific selenoproteins. It is not clear whether selenium deficiency not accompanied by other deficiencies or oxidative stress results in significant physiological impairment. Severely low selenium intakes (greater than 20 μg d^{-1}) have been associated with juvenile cardiomyopathy in China (Keshan disease); but that disease also appears to have a viral component to its etiology.

Recent interest in selenium centers around the apparent efficacy of supranutritional intakes to reduce cancer risk. A decade-long, randomized, double-blind, placebo-controlled clinical intervention trial found that supplementation of free-living American adults with 200 μg d^{-1} selenium (in addition to their normal diets) reduced major cancer incidence by half or more (Clark et al. 1996) if their baseline intakes did not exceed ca. 100 μg d^{-1}. Studies with animal models have demonstrated similar protective effects of selenium intakes greater than those required to support maximal expression of selenoenzymes. Current thinking is that these effects are due to a methylated metabolite of selenium (Jackson and Combs 2011). Selenium intakes greater than 1 mg d^{-1} can induce dermatological changes, including brittle hair and nails. Chronic intakes approaching 5 mg has been reported to lead to skin rash, paresthesia, weakness, and diarrhea (see also Chap. 16, this volume).

Manganese—Manganese functions as a cofactor for enzymes in antioxidant defense (mitochondrial superoxide dismutase), gluconeogenesis (pyruvate carboxylase, phosphoenol–pyruvate carboxykinase), glycoprotein biosynthesis (glycosyl transferases), nitrogen metabolism (arginase, glutamine synthase), and cholesterol biosynthesis (farnesyl pyrophosphate synthetase). Little is known about the mechanisms of absorption, transport, or cellular uptake of manganese, although the element is widely distributed in noncalcified tissues with the greatest concentrations in the liver (Leach and Harris 1997). The greatest route of manganese excretion appears to be the bile in which it is released in bound form. There is little available evidence of manganese deficiency in humans, although studies in experimental animals have shown effects on fetal survival, normal skeletal development (i.e., shortened limbs, twisted legs, lameness), ataxia, glucose tolerance, and hepatic steatosis.

Molybdenum—Molybdenum functions as the active center of three enzymes that catalyze oxidative hydroxylations: sulfite oxidase (the last step in the degradation of

sulfur amino acids), xanthine dehydrogenase, and aldehyde oxidase (which transfers electrons to other redox cofactors and, ultimately, to cytochrome *c*, molecular oxygen, or NAD^+). In these enzymes, the element is found in a pterin-containing, molybdenum cofactor, the synthesis of which in eukaryotes remains poorly understood (Johnson 1997). Molybdenum appears to be efficiently absorbed at all levels of intake, apparently by a passive process. It is transported in the blood attached to proteins in erythrocytes. Whole blood molybdenum levels vary directly with dietary molybdenum intake, although plasma Mo^+ levels are maintained at about 5 $nmolL^{-1}$. Molybdenum is widely distributed in the body, with greatest concentrations in liver, kidney, adrenal gland, and bone. One human case of molybdenum deficiency has been described; signs (all of which responded to molybdenum therapy) included tachycardia, tachypnea, severe headache, night blindness, nausea, and vomiting.

Chromium—Chromium potentiates the action of insulin and has been shown to restore glucose tolerance in malnourished infants. Several studies have shown that chromium supplementation lowers circulating glucose levels, increases plasma insulin, and produces a favorable profile of plasma lipids (Offenbacher et al. 1997). It has been suggested that these effects may be due to a low molecular weight chromium-binding substance that may amplify insulin receptor tyrosine kinase activity in response to insulin. Chromium can also bind to one of the binding sites of transferrin, and it has been proposed that excessive iron storage in hemochromatosis may interfere with the transport of chromium to contribute to the diabetes associated with that disorder. The enteric absorption of the element, which is absorbed as Cr^{+3}, is very low (usually no more than 2%) and appears to be regulated. Absorbed chromium accumulates in the liver, kidney, spleen, and bone.

Fluoride—Fluoride is the ionic form of fluorine. It is very highly electronegative and reacts reversibly to hydrogen to form hydrogen fluoride which freely diffuses across the intestine, dissolves in the blood, and is taken up by the tissues where its high affinity for calcium causes it to accumulate in calcified tissues. Fluoride can stimulate new bone formation; when present in oral fluids, F^-exerts cariostatic effects due to enhanced remineralization of dental enamel and reduced acid production of plaque bacteria (Chow 1990; Cerklewski 1997). Prior to the widespread use of F^-in dental products and water supplies, most studies showed that the incidence of dental caries (in both children and adults) was 40–60% lower in areas with drinking water F^- concentrations of at least 0.7 mgL^{-1} when compared to communities with lower F^-levels. Excessive F^- intake can cause fluorosis of the enamel and bone. Although the former is a largely cosmetic effect involving the mottling of the teeth, skeletal fluorosis is associated with joint stiffness, calcification of ligaments, and some osteosclerosis of the pelvis and vertebrae (see also Chap. 13, this volume).

8.3 Dietary Sources of Essential Mineral Elements

Mineral elements are metabolized and, to varying degrees, stored by plants and animals, some of which constitute important sources of those elements in human diets (see Table 8.3). That the mineral elements are not homogeneously distributed among various types of foods is clear: few foods other than dairy products are rich in calcium; sea foods constitute the best sources of iodine and chloride; meats are the most important sources of iron; and

Table 8.3 Important dietary sources of essential mineral elements

Element	Sources
Ca	Dairy products, fortified juices, kale, collards, mustard greens, broccoli, sardines, oysters, clams, canned salmon
P	Meats, fish, eggs, dairy products, nuts, beans, peas, lentils, grains
Mg	Seeds, nuts, beans, peas, lentils, whole grains, dark green vegetables
Na	Common table salt, seafoods, dairy products, meats, eggs
K	Fruits, dairy products, meats, cereals, vegetables, beans, peas, lentils
Cl	Common table salt, seafoods, dairy products, meats, eggs
Fe	Meats, seafoods
Cu	Beans, peas, lentils, whole grains, nuts, organ meats, seafood (e.g., oysters, crab), peanut products, chocolate, mushrooms
Zn	Meats, organ meats, shellfish, nuts, whole grains, beans, peas, lentils, fortified breakfast cereals
Se	Meats from Se-fed livestock; sea fish; grain products, nuts, garlic, broccoli grown on high-Se soils
I	Iodized salt; sea fish; kelp
Mn	Whole grains, beans, peas, lentils, nuts, tea
Mo	Beans, peas, lentils, dark green leafy vegetables, organ meats
F	Fluoridated water

protein-rich foods comprise the best sources of zinc, copper, and selenium. Therefore, optimal mineral nutrition, like optimal nutrition in general, is most likely to be obtained from mixed diets based on a diverse selection of foods. Conversely, the monotonous, non-diverse, grain-based diets accessible to the poor of the developing world are likely to provide insufficient energy, protein, and minerals, especially calcium, copper, selenium, and biologically available iron and zinc. At the same time, the increasing use in industrialized countries of non-diverse eating habits is associated with prevalent insufficient intake of such minerals as calcium.

For many people in industrialized countries, fortified foods and nutritional supplements constitute important sources of several of the mineral elements. Various forms of copper, zinc, iron, and selenium are offered in over-the-counter formations, both as individual supplements as well as compounded in multivitamin mineral supplements. Calcium, typically as the carbonate or gluconate salts, is now commonly used to fortify orange and other fruit juices. Consumer response to such nutrient-fortified foods has been very strong, and this aspect of consumer retailing is expected to continue to grow.

Soils can contribute to the total dietary intake of mineral elements. This can occur through adherent soil particles on foods and suspended soil particles in drinking and cooking water, as well as through the direct consumption of soil. The latter practice of geophagia can be deliberate in some communities in which the eating of clays occurs (see Chap. 18 "Geophagy and the Involuntary Ingestion of Soil", this volume). Consumption of clays with high cation-exchange capacities can provide substantial supplements of calcium, iron, copper, zinc, and manganese; other clays can interfere with the enteric absorption of iron and zinc. Consumption of iron-rich lateritic soils can provide enough iron to impair the utilization of copper and zinc.

Water supplies are variable in their mineral contents. In some areas, fresh water supplies can provide nutritionally important amounts of such minerals as calcium, magnesium, iron and manganese, and industrialized countries have used municipal water as a vehicle for providing fluoride. Municipal systems handling "soft", i.e., low-mineral water, can have appreciable amounts of copper due to leaching from pipes. In a few locales surface runoff from selenium-rich soils has been found to contain biologically significant amounts of selenium, but such cases are rare and most water supplies are very low in that nutrient. The contributions of water-borne minerals to nutritional status depend also on the rate of water consumption, which is also highly variable, depending on behavior and environmental conditions. Individuals with the greatest intakes of water include those engaged in strenuous physical activity, residents of hot climates, and (relatively) infants.

8.4 Mineral Element Bioavailability

For several nutrients only a portion of the ingested amount is absorbed and utilized metabolically. Therefore, it is necessary to consider this when evaluating the nutritional adequacy of foods and diets. This concept, bioavailability, is particularly important in mineral nutrition, because some foods are less useful sources of essential minerals than might be expected from their absolute mineral content.

Mineral bioavailability depends on both physiological and exogenous factors. Physiological determinants of mineral bioavailability include:

1. Age-related declines in the efficiency of enteric absorption of copper and zinc
2. Early postnatal lack of regulation of absorption of iron, zinc, and chromium
3. Adaptive increases in the absorptive efficiencies of iron and zinc, copper, manganese, and chromium by receptor upregulation during periods of deficiency
4. Dependence on other nutrients for the physiological functions of selenium and iodine in thyroid hormone metabolism, and copper and iron in catecholamine metabolism
5. Anabolic effects on tissue sequestration of zinc and selenium
6. Catabolic effects on zinc, selenium, and chromium losses

The absorption and post-absorptive utilization of ingested minerals can be affected by other factors in foods and diets. For example, 25–30% of the heme iron in animal tissues can be absorbed, while only 2–5% of the iron in plant foods is absorbed. The utilization of plant sources of iron can be markedly improved by including in the diet sources of ascorbic acid (e.g., oranges) or meats, both of which promote the utilization of non-heme iron. Similarly, citrate and/or histidine can enhance the absorption of dietary zinc. Dietary ascorbate (vitamin C) can also enhance the antagonistic effect of iron on copper utilization.

Mineral bioavailability can be reduced by dietary factors that reduce enteric absorption. For example, phytate, phosphorus, and triglycerides can each reduce the luminal solubility and, hence, the absorption of calcium. Phytate and other non-fermentable fiber components can bind zinc and magnesium, reducing the absorption of each. Sulfides can reduce the absorption of copper by similar means. Minerals that share transporters can be mutually inhibitory for absorption, e.g., sulfite and selenite, cadmium and zinc, and zinc and copper.

In general, problems related to poor bioavailability are greatest for iron in plant-based containing phytates and/or polyphenols but there are few problems with promotor substances. For calcium there are problems with bioavailability when poorly soluble forms are consumed with

vegetables (spinach, rhubarb, beet greens, chard) containing inhibitory oxalates without others (artichokes) containing fructose oligosaccharide promoters; for zinc in diets high in unrefined (>90% extraction), unfermented cereal grains or high-phytate soy products, especially those fortified with inorganic calcium salts; and for selenium consumed as plant foods (containing SeMet much of which is diverted to protein synthesis as a methionine analog). Accordingly, the utilization of these minerals as consumed in most diets tends to be moderate at best, though in each case it can be markedly enhanced through appropriate dietary choices.

8.5 Quantitative Estimates of Mineral Needs and Safe Exposures

Dietary standards have been set for several, but not all, of the nutritionally essential mineral elements. International standards have been developed for only some minerals (FAO-WHO 2002) (Table 8.4). The most current and extensive standards are the Dietary Reference Intakes (DRSs) published by the U.S. National Academy of Science (NAS) (Food and Nutrition Board 1997, 2000, 2001) (Table 8.5). It is important to note that the expert panels of the respective organizations used the same primary data, i.e., the published scientific literature. Also, each based its recommendations on estimates of individual physiological need (i.e., the World Health Organization's "basal requirement" and the "recommended dietary allowance," RDA, from NAS) which was then inflated to accommodate estimated inter-individual variation. This approach produced the WHO "normative requirement" and NAS "estimated average requirement" (EAR). The NAS process went further to include estimates of "average intakes" (AIs) in cases where data were not sufficient to support EARs or RDAs. Both groups also estimated safe limits of exposure: WHO created "upper limits of safe ranges of population mean intakes" (Table 8.6), and NAS created "upper tolerable limits" (ULs) (Table 8.7).

8.6 Clinical Assessment of Mineral Status

The status assessment of the essential minerals, which vary so much in metabolic function, homeostatic regulation, and tissue distribution, calls for a mixed approach. This approach includes elemental analyses of tissues and/or body fluids, assays of mineral-dependent enzyme activities, and measurement of functional and/or morphological indices. A battery of such tests may be feasible in research settings, but in clinical settings practicality and timeliness dictate approaches based on analyses of a single specimen of blood.

Beyond the obvious issues pertaining to sampling (i.e., number, bias, amount, homogeneity, interindividual variability, etc.), the analysis of minerals, particularly those present in only trace amounts in foods and tissues, calls for special attention to sample integrity and freedom from contamination (Milne 2000). For example, the iron and zinc contents of plasma or serum can be affected by hemolysis; rubber stoppers and borosilicate glass can contaminate blood with zinc and boron, respectively; and some anticoagulants can produce osmotic shifts that release several elements from erythrocytes. The laboratory, too, can be a significant source of contamination: poorly treated water can contaminate with iron, calcium, magnesium, manganese, zinc, or copper; stainless steel surfaces can contaminate with chromium and nickel; and dust, paper products, wood, skin, hair, and dandruff can also be sources of contamination. For these reasons, a well-monitored laboratory designed for mineral/trace element analyses is a prerequisite for the generation of useful data.

The available methods for the clinical assessment of mineral status are presented in Table 8.8, with normative values for the most practically useful of these presented in Table 8.9.

Status with respect to mineral elements that are active or highly regulated in circulating tissues can be assessed by analyzing their amounts in plasma/serum or blood cells (Sauberlich 1999). For example, knowledge of plasma/serum potassium or erythrocyte iron levels can be highly informative, because those elements exert their physiological functions in those respective compartments. This is not the case for mineral elements that function in other compartments and/or chemical forms. For example, analyses of chromium, copper, or selenium in serum/plasma have inferential value for assessing status only to the extent that those values correlate with the sizes/activities of other physiologically relevant pools. For elements that are not highly regulated in the blood, such as zinc, that parameter has limited, if any, value in assessing status in all except severely deficient individuals.

For mineral elements such as selenium, iodine, zinc, and copper, which exert their physiological functions as essential constituents of macromolecules, assessment of status calls for measurement of the levels/activities of their respective functional forms or metabolite profiles. Thus, zinc adequacy can be determined on the basis of the cytosolic superoxide dismutase, and iodine adequacy can be determined on the basis of circulating levels of triiodothyronine (T_3), thyroid hormone (T_4), and thyroid-stimulating hormone (TSH). Similarly, plasma selenium, because it consists of several components including nonfunctional selenium bound nonspecifically in albumin and other proteins, is best assessed in cases of subadequacy by determining the selenoproteins-extracellular glutathione peroxidase and selenoprotein P.

Table 8.4 International dietary recommendations (units per day) for essential mineral elements[a,b]

Life Stage	Ca (mg)	P (mg)	Na (mg)	K (mg)	Cl (mg)	Mg (mg)	Fe[c] (mg)	Cu (mg)	Zn[d] (mg)	Se (μg)	I (μg)	Mn (mg)	Mo (mg)	Cr (μg)	F (mg)
Children															
0–3 month								0.33–0.55[f]							
3–6 month	300[e]					26[e]		0.37–0.62[f]	2.8[e]	6	[15][g]				
	400[f]					36[f]									
7–12 month	400					53	[9][g]	0.60	4.1	10	130				
1–3 year	500					60	6	0.58	4.8	17	75				
3–6 year	600					73	6	0.57	5.1	21	115				
6–9 year	700					100	9	0.75	5.6	21	110				
10–11 year											140f				
											135m				
10–12 year								0.77f							
								0.73m							
10–14 year							14f				140f				
							15m				135m				
12–15 year								1.00							
12–18 year											110f				
											100m				
15–18 year							31f	1.33f							
							19m	1.15m							
10–18 year	1,300					230f			7.8f	26f					
						250m			9.7m	32m					
Adults															
19–50 year	1,000					220f	29f	1.35f	4.9f	26f	110f				
						260m	14m	1.15m	7.0m	34m	130m				
51–65 year	1,300f					220f	11f	1.35f	4.9f	26f	110f				
	1,000m					260m	14m	1.15m	7.0m	34m	130m				
65+ year	1,300					190f	11f	1.35f	4.9f	26f	110f				
						230m	14m	1.15m	7.0m	34m	130m				
Pregnancy															
1st trimest.r						220		1.15	5.5		200				
2nd trimest.						220		1.15	7.0	28	200				
3rd trimest.	1,200					220		1.15	10.0	30	200				
Lactation															
0–3 month	1,000					270	15	1.25	9.5	35	200				
3–6 month	1,000					270	15	1.25	8.8	35	200				
6–12 month	1,000					270	15	1.25	7.2	42	200				

[a]Recommendations for copper are normative dietary requirements: WHO (1996)
[b]Recommendations for calcium, magnesium, iron, zinc, selenium and iodine: FAO-WHO (2002)
[c]10% bioavailability conditions
[d]Moderate (30–35%) bioavailability conditions
[e]Breast fed infants
[f]Formula fed infants
[g]Value expressed in units kg^{-1} day^{-1}

8.7 Ecological Aspects of Mineral Nutrition

Because the mineral elements are ultimately derived from soils, the mineral status of humans and other animals depends on the minerals available in the soils upon which their foods were grown and through which their drinking and cooking waters drained (see Table 8.3). Therefore, it is not surprising that mineral nutritional status can vary geographically, particularly in cases where the soil-water-plant-animal linkages are fairly direct as in the cases of grazing animals and people in highly localized food systems. Such cases have been described for iodine, copper, zinc, selenium, molybdenum, manganese, iron, boron, and cobalt (see Table 8.10). Soil mineral deficiencies can involve intrinsically low mineral contents of soils

Table 8.5 Dietary recommendations for essential mineral elements (units/day), (Food and Nutrition Board 1997, 2000, 2001)

Life Stage	Ca (mg)	P (mg)	Na (mg)	K (mg)	Cl (mg)	Mg (mg)	Fe (mg)	Cu (μg)	Zn (mg)	Se (μg)	I (μg)	Mn (mg)	Mo (mg)	Cr (μg)	cF (mg)
Children															
0–6 month	210[a]	100[a]	120[c]	500[c]	180[c]	30[a]	.27[d]	200[d]	2[d]	15[e]	110[d]	.003[d]	.2[d]	.2[d]	.01[a]
7–12 month	270[a]	275[a]	200[c]	700[c]	300[c]	75[a]	11[f]	220[f]	3[f]	20[e]	130[d]	.6[d]	3[d]	5.5[d]	.5[a]
1–3 year	500[a]	460[b]	225–300[c]	1,000–1,400[c]	350–500[c]	80[b]	7[f]	340[f]	3[f]	20[g]	90[f]	1.2[d]	17[f]	11[d]	.7[a]
4–8 year	800[a]	500[b]	300–400[c]	1,400–1,600[c]	500–600[c]	130[b]	10[f]	440[f]	5[f]	30[g]	90[f]	1.5[d]	22[f]	15[d]	1[a]
9–13 year	1,300[a]	1,250[b]	400–500[c]	1,600–2,000[c]	600–750[c]	240[b]	8[f]	700[f]	8[f]	40[g]	120[f]	1.9[d](m)	34[f]	25[d](m)	2[a]
												1.6[d](f)		21[d](f)	
14–18 year	1,300[a]	1,250[b]	500[c]	2,000[c]	750[c]	410 (m)[b]	11[f](m)	890[f]	11[f](m)	55[g]	150[f]	2.2[d](m)	43[f]	35[d](m)	3[a]
						360 (f)[b]	15[f](f)		9[f](f)			1.6[d](f)		24[d](f)	
Adults															
19–30 year	1,000[a]	700[b]	500[c]	2,000[c]	750[c]	400 (m)[b]	8[f](m)	900[f]	11[f](m)	55[g]	150[f]	2.3[d](m)	45[f]	35[d](m)	4(m)[a]
						310 (f)[b]	11[f](f)		8[f](f)			1.8[d](f)		25[d](f)	3(f)[a]
31–50 year	1,000[a]	700[b]	500[c]	2,000[c]	750[c]	420 (m)[b]	8[f](m)	900[f]	11[f](m)	55[g]	150[f]	2.3[d](m)	45[f]	35[d](m)	4(m)[a]
						320 (f)[b]	11[f](f)		8[f](f)			1.8[d](f)		25[d](f)	3(f)[a]
51+ year	1,200[a]	700[b]	500[c]	2,000[c]	750[c]	420 (m)[b]	8[f]	900[f]	11[f](m)	55[g]	150[f]	2.3[d](m)	45[f]	30[d](m)	4(m)[a]
						320 (f)[b]			8[f](f)			1.8[d](f)		20[d](f)	3(f)[a]
>70 year	1,200[a]	700[b]	500[c]	2,000[c]	750[c]	420 (m)[b]	8[f]	900[f]	11[f](m)	55[g]	150[f]	2.3[d](m)	45[f]	30[d](m)	4(m)[a]
						320 (f)[b]			8[f](f)			1.8[d](f)		20[d](f)	3(f)[a]
Pregnancy															
≥18 year	1,300[a]	1250[b]				400[b]	27[f]	1,000[f]	13[f]	60[g]	220[f]	2[d]	50[f]	19[d]	3[a]
19–30 year	1,000[a]	700[b]				350[b]	27[f]	1,000[f]	11[f]	60[g]	220[f]	2[d]	50[f]	30[d]	3[a]
31–50 year	1,000[a]	700[b]				360[b]	27[f]	1,000[f]	11[f]	60[g]	220[f]	2[d]	50[f]	30[d]	3[a]
Lactation															
≥18 year	1,300[a]	1250[b]				360[b]	10[f]	985[f]	14[f]	70[g]	290[f]	2.6[d]	50[f]	44[d]	3[a]
19–30 year	1,000[a]	700[b]				310[b]	9[f]	1,000[f]	12[f]	70[g]	290[f]	2.6[d]	50[f]	45[d]	3[a]
31–50 year	1,000[a]	700[b]				320[b]	9[f]	1,000[f]	12[f]	70[g]	290[f]	2.6[d]	50[f]	45[d]	3[a]

[a]Average Intake (*AI*) value (Food and Nutrition Board 1997)
[b]Recommended Dietary Allowance (*RDA*) (Food and Nutrition Board 1997)
[c]Estimated Minimum Requirement (Food and Nutrition Board 1989)
[d]Average Intake (*AI*) value (Food and Nutrition Board 2001)
[e]Average Intake (*AI*) value (Food and Nutrition Board 2000)
[f]Recommended Dietary Allowance (*RDA*) (Food and Nutrition Board 2001)
[g]Recommended Dietary Allowance (*RDA*) (Food and Nutrition Board 2000)

(e.g., selenium), inefficient uptake by crops (e.g., zinc deficiency in calcareous soils), and excessive leaching (e.g., iodine, zinc). In at least two general cases, Keshan disease and the iodine deficiency diseases goiter and myxedematous cretinism, endemic distributions of a disease are directly related to the geographic patterns of soil deficiencies in selenium and iodine, respectively. Interregional and international transshipment of foods can be expected to mitigate against such local soil effects, particularly in industrialized countries (see also Chap. 16, this volume).

Nutritionally important mineral elements occur in soils as silicates (e.g., manganese, zinc, selenium), sulfides (e.g., zinc, selenium, copper, iron), and as the native elemental form (iron). The most abundant of these is iron, the fourth most abundant element in the Earth's crust (see also Chap. 2,

Table 8.6 International estimates of upper limits (units per day) of safe intakes of essential mineral elements, WHO (1996)

Life Stage	Ca (mg)	P (mg)	Na (mg)	K (mg)	Cl (mg)	Mg (mg)	Fe (mg)	Cu (mg)	Zn[a] (mg)	Se (μg)	I (μg)	Mn (mg)	Mo (mg)	Cr (μg)	F (mg)
Children															
0–6 month									13						
0–12 month								(150)[b]							
1–6 year								1.5	23						
6–10 year								3.0	28						
10–12 year								6.0	32f 34m						
12–15 year								8.0	36f 40m						
15–18 year									38f 48m						
Adults															
15–60 year								12.0		400					
18–60 year									55f 45m						
Pregnancy															
general								10.0							
Lactation															
general								10.0							

[a]Moderate (30–35%) bioavailability conditions
[b]mcg kg^{-1}

Table 8.7 Estimated upper tolerable intakes of essential mineral elements (units/day) (Food and Nutrition Board 1997, 2000, 2001)

Life Stage	Ca (mg)	P (mg)	Na (mg)	K (mg)	Cl (mg)	Mg (mg)	Fe (mg)	Cu (mg)	Zn (mg)	Se (μg)	I (mg)	Mn (mg)	Mo (mg)	Cr (μg)	F (mg)
Children															
0–6 month									4[c]	45[b]					.7[a]
7–12 month									5[c]	60[b]					.9[a]
1–3 year	2.5[a]	3[a]				65[a]		1[c]	7[c]	90[b]	.2[c]	2[c]	.3[c]		1.3[a]
4–8 year	2.5[a]	3[a]				110[a]		3[c]	12[c]	150[b]	.3[c]	3[c]	.6[c]		2.2[a]
9–13 year	2.5[a]	4[a]				350[a]		5[c]	23[c]	280[b]	.6[c]	6[c]	1.1[c]		10[a]
14–18 year	2.5[a]	4[a]				350[a]		8[c]	34[c]	400[b]	.9[c]	9[c]	1.7[c]		10[a]
Adults															
19–30 year	2.5[a]	4[a]				350[a]		10[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
31–50 year	2.5[a]	4[a]				350[a]		10[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
51+ year	2.5[a]	4[a]				350[a]		10[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
>70 year	2.5[a]	4[a]				350[a]		10[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
Pregnancy															
≥18 year	2.5[a]	3.5[a]				350[a]		8[c]	40[c]	400[b]	.9[c]	9[c]	1.7[c]		10[a]
19–30 year	2.5[a]	3.5[a]				350[a]		8[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
31–50 year	2.5[a]	3.5[a]				350[a]		8[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
Lactation															
≥18 year	2.5[a]	4[a]				350[a]		8[c]	34[c]	400[b]	.9[c]	9[c]	1.7[c]		10[a]
19–30 year	2.5[a]	4[a]				350[a]		10[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]
31–50 year	2.5[a]	4[a]				350[a]		10[c]	40[c]	400[b]	1.1[c]	11[c]	2[c]		10[a]

[a]Food and Nutrition Board (1997)
[b]Food and Nutrition Board (2000)
[c]Food and Nutrition Board (2001)

Table 8.8 Clinical assessment of mineral element status

	Most useful parameters of status, by general type			
Element	Elemental analysis[a]	Indicator enzymes/ proteins	Indicator metabolites	Physiological indices
Ca	Serum total Ca–AAS[b]. ES[c], MS[d], NAA[e], ICP-MS[f]			
	Serum Ca^{++} - EC[g]			
P	Serum P–AAS, C[g], EC, ES, MS, GFAAS, NAA			
Mg	Serum/plasma Mg–AAS, MS, ICP-M			
	S, NAA			
	Serum Mg^{++} - EC			
	Muscle Mg–AAS, MS, ICP-MS			
	Erythrocyte Mg–AAS, MS, ICP-MS			
Na	Serum/plasma Na–Aas, C, EC, NAA			
	Urinary Na–AAS, C, EC, NAA			
K	Serum/plasma K–AAS, C, EC, NAA			
	Urinary K–AAS, C, EC, NAA			
Cl	Serum/plasma Cl–C, EC, NAA			
	Urinary Cl–C, EC, NAA			
Fe	Serum Fe–AAS, GFAAS, C, EC, ES, MS, NAA, PIXE[I]	Erythrocyte hemoglobin	Free erythrocyte protoporpphyrin	Hematrocrit
		Serum ferritin		
		Seum transferrin	Zn-protoporphorin	Serum total Fe-binding capacity
		Serum transferrin receptor		
Zn	Serum/plasma Zn–AAS, GFAAS, ES, MS	Metallothionine I		
		Alkaline phosphatase		
	Hair/nail Zn–AAS, GFAAS, ES, MS	Carbonic anhydrase		
	Urinary Zn–AAS, GFAAS, ES, MS	Nucleoside Phosphorylase		
	Leukocyte Zn–AAS, GFAAS, ES, MS	Ribonuclease		
Cu	Serum/plasma Cu–AAS, GFAAS, ES, MS, PIXE	Ceruloplasmin activity		
		Superoxide dismutase activity		
		Cytochrome *c* Oxidase activity		
Se	Serum/plasma Se–GFAAS, HGAAS[j]	Glutathione peroxidase isoforms		
	Hair/nail Se–EAAS	Serum selenoprotein P		
I	Urinary I–C, POT[k], NAA		Tetraiodothyronine (T_4)	
			Thryoid hormone (T_3)	
			Thyroid-stimulating hormone (TSH)	
Mn	Serum/plasma Mn–AAS, GFAAS, ES, MS, NAA			
Mo	Serum/plasma Mo–GFAAS, MS, NAA			
Cr	Serum/plasma Cr–GFAAS, NAA			
F	Serum/plasma F–NAA			

[a]Acceptable analytical method (typically yield CV $< 10\%$)
[b]*AAS* atomic absorption spectrophotometry
[c]*ES* emission spectroscopy
[d]*MS* mass spectrometry
[e]*NAA* neutron activation analysis
[f]*GFAAS* AAS with electrothermal atomization using a graphite furnace
[g]*EC* electrochemistry
[h]Chemical methods
[i]*PIXE* proton-induced X-ray emission
[j]*HGAAS* hydride generation AAS
[k]*POT* potentiometry

Table 8.9 Reference values for key clinical parameters of mineral element status

Element	Parameter	Reference value[a]
Ca	Serum total Ca	86–102 μg L^{-1}
	Serum Ca^{++}	46.4–52.8 μg L^{-1}
P	Serum P	Adult: 25–45 μg L^{-1}
		Children: 40–70 μg L^{-1}
Mg	Serum/plasma Mg	16–26 μg L^{-1}
Na	Serum/plasma Na	300–350 μg L^{-1}
	Urinary Na	1,950–3,400 mg/day
K	Serum/plasma K	130–200 μg L^{-1}
	Urinary K	90–450 mg/day
Cl	Serum/plasma Cl	350–400 μg L^{-1}
	Urinary Cl	>40 mg/L
Fe	Seum Fe	Women: 500–1,700 μg L^{-1}
	Serum total Fe-binding capacity	Men: 650–1,650 μg L^{-1}
	Serum ferritin	2,500–4,250 μg L^{-1}
		Women: 100–1,200 μg L^{-1}
		Men: 200–2,500 μg L^{-1}
Zn	n.a.[b]	Plasma Zn normal range: 700–1,500 μg L^{-1}
Cu	Serum/plasma Cu	Women: 800–1,900 μg L^{-1}
		Pregnant women: 1,180–3,020 μg L^{-1}
		Men: 700–1,400 μg L^{-1}
		Infants: 200–700 μg L^{-1}
		Children, 60–120 y: 80–190 μg L^{-1}
Se	Plasma/serum Se	800–2,000 μg L^{-1}
I	Urinary I	>,1,000 μg L^{-1}
	Serum T_4	60–100 μg L^{-1}
	Serum TSH	1–50 μg L^{-1}
Mn	n.a.[b]	Serum/plasma Mn normal range: 4–11 μg L^{-1}
		Whole blood Mn normal range: 77–121 μg L^{-1}
Mo	n.a.[b]	Serum/plasma Mo normal range: 1–30 μg L^{-1}
		Whole blood Mo normal range: 8–33 μg L^{-1}
		Urine Mo normal range: 80–340 μg L^{-1}
Cr	Serum/plasma Cr	1–1.6 μg L^{-1}
F	n.a.[b]	n.a.[b]

[a]Normal range for healthy adults, unless otherwise indicates
[b]Validated method not available

Table 8.10 Examples of geographic patterns of deficiencies of nutritionally important mineral elements

Element	Known deficient areas
I	A wide range of soils in areas remote from sea coasts; in the US, the Northwestern mountains and upper Midwest lake areas
Cu	Acid histosols in the eastern US; acid sands in Florida; podzolic soils in Wisconsin; some sandy alkaline soils
Zn	An estimated half of the world's soils with small deficient spots in many areas; most likely in calcareous or leached, acid, sandy soils
Se	Mountainous belt of northeast China to the Tibetan plateau; parts of Africa; Pacific northwest, northeast and lower eastern seaboard of the US
Mo	Acid soils; eastern seaboard, Great Lakes and Pacific coast areas of the US
Mn	Humid, organic soils of the eastern US
Fe	Seldom a problem for food plants except in arid regions
B	Many US states, particularly in neutral-to-alkaline soils
Co	US lower Atlantic coastal plain and lower Maine coast; parts of Australia

this volume). These are taken up by plants, which are consumed directly or indirectly as tissues of animals feeding of those plants. The grazing practices of many livestock allow them to harvest mineral elements over relatively large geographic areas. Nevertheless, for several mineral elements (e.g., selenium, zinc), the contents of both plant- and animal-base foods ten to reflect the mineral contents of the soils of their respective areas of production.

Water supplies are variable in their mineral contents. Most provide only small amounts of essential minerals; however, some can contribute appreciable amounts of certain mineral elements due to natural conditions (e.g., calcium, magnesium, selenium, fluoride, zinc), to intentional addition (e.g., flouride) and/or to leaching from pipes (e.g., copper). The contributions of water-borne minerals to nutritional status depend also on the rate of water consumption, which is also highly variable, depending on behavior and environmental conditions. Individuals with the greatest intakes of water include those engaged in strenuous physical activity, residents of hot climates, and (relatively) infants. In general, those water-borne mineral elements of greatest health import include calcium, magnesium, sodium, potassium, copper and selenium. Epidemiological studies conducted in several parts of the world have indicated an inverse relationship of water hardness and cardiovascular disease risk (WHO 2005). These benefits are thought to relate to magnesium and/or calcium, as clinical studies have found that these minerals can be effective in reducing hypertension. There is evidence that consumption of drinking water containing flouride reduces risk of dental carries, but that consumption of drinking water containing iron and copper may increase risks of heart attack.

8.8 Summary

Minerals play essential roles in the normal metabolism and physiological functions of animals and humans. Some (calcium, phosphorus, magnesium, fluoride) are required for structural functions in bone and membranes. Some (sodium, potassium, chloride) are required for the maintenance of water and electrolyte balance in cells. Some (zinc, copper, selenium, manganese, molybdenum) are essential constituents of enzymes or serve as carriers (iron) for ligands essential in metabolism. Some serve as essential components of a hormone (iodine) or hormone-like factor (chromium).

Unlike other essential nutrients, the mineral elements cannot be derived from the biosynthesis of food plants or animals–they must be obtained from soils and pass through food systems to humans in food forms. For this reason, local deficiencies of minerals in soils can produce deficiencies in local food systems which clinically impact the people dependent on those systems. The development of international trade and interregional transportation of foods has ameliorated the impact of such local mineral deficiencies. However, cases still occur in areas where the transshipment of food and, thus, the diversity of the diet are limited.

See Also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements) • Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 9 (Biological Responses of Elements) • Chapter 16 (Selenium Deficiency and Toxicity in the Environment) • Chapter 17 (Soils and Iodine Deficiency) • Chapter 27 (Speciation of Trace Elements) • Chapter 28 (Mineralogy of Bones)

Further Reading

Beard JL, Dawson HD (1997) Phosphorus. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 275–334, chap. 3

Berner YN (1997) Phosphorus. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 63–92, chap. 3

Bogden JD, Klevay LM (eds) (2000) Clinical nutrition of the essential trace elements and minerals. The guide for health professionals. Humana Press, Totowa

Bronner F (1997) Calcium. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral element. Marcel Dekker, New York, pp 13–61, chap. 2

Cerklewski FL (1997) Fluorine. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 583–602, chap. 20

Chesters JK (1997) Zinc. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 185–230, chap. 7

Chow LC (1990) Tooth-bound fluoride and dental caries. J Dent Res 69:595–600

Clark LC, Combs GF Jr, Turnbull BW, Slate E, Alberts D, Abele D, Allison R, Bradshaw R, Chalker D, Chow J, Curtis D, Dalen J, Davis L, Deal R, Dellasega M, Glover R, Graham G, Gross E, Hendrix J, Herlong J, Knight F, Krongrad A, Lesher J, Moore J, Park K, Rice J, Rogers A, Sanders B, Schurman B, Smith C, Smith E, Taylor J, Woodward J (1996) The nutritional prevention of cancer with selenium 1983–1993: a randomized clinical trial. J Am Med Assoc 276:1957–1963

Combs GF Jr (2001) Selenium in global food systems. Br J Nutr 85:517–547

FAO-WHO (2002) Human vitamin and mineral requirements: report of a joint FAO/WHO expert consultation. Food and Agricultural Organization of the United Nations, World Health Organization, Rome

Food and Nutrition Board (1989) Recommended dietary allowances, 10th edn. National Academy Press, Washington, DC

Food and Nutrition Board (1997) Dietary reference intakes for calcium, phosphorus, magnesium, vitamin D and fluoride. National Academy Press, Washington, DC

Food and Nutrition Board (2000) Dietary reference intakes for vitamin C, vitamin E, selenium and carotenoids. National Academy Press, Washington, DC

Food and Nutrition Board (2001) Dietary reference intakes for vitamin a, vitamin K, arsenic, boron, chromium, copper, iodine, iron, manganese, molybdenum, nickel, silicon, vanadium and zinc. National Academy Press, Washington, DC

Harper ME, Willis JS, Patrick J (1997) Sodium and chloride in nutrition. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 93–116, chap. 4

Harris ED (1997) Copper. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 231–273, chap. 3

Hetzel BS, Wellby ML (1997) Iodine. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 557–581, chap. 19

Hollick MF (1994) Vitamin D–New horizons for the 21st century. Am J Clin Nutr 60:619–630

Jackson MI, Combs GF Jr (2008) Selenium and anticarcinogenesis: underlying mechanisms. Curr Opin Clin Nutr Metab Care 11: 718–726

Jackson MI, Combs GF Jr (2011) Selenium as a cancer preventive agent. In: Hatfield DL, Berry MJ, Gladyshev N (eds) Selenium: its molecular biology and role in human health. Springer, New York, pp 313–324, chapter 24

Johnson JL (1997) Molybdenum. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 413–438, chap. 13

Jones G, Strugnell RA, DeLuca HF (1998) Current understanding of the molecular actions of vitamin D. Physiol Rev 78:1193–1231

Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN (2003) Characterization of mammalian selenoproteins. Science 300:1439–1443

Kubota J, Allaway WH (1972) Geographic distribution of trace element problems. In: Mortvedt JJ, Giodano PM, Lindsay WL (eds) Micronutrients in agriculture. Soil Science Society of America, Madison, pp 525–554

Leach RM Jr, Harris ED (1997) Manganese. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 335–355, chap. 10

Milne DB (2000) Laboratory assessment of trace element and mineral status. In: Clinical nutrition of the essential trace elements and minerals: the guide for health professionals. Humana Press, Totowa, pp 69–90

Offenbacher EG, Pi-Sunyer FX, Stoeker BJ (1997) Chromium. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 389–411, chap. 12

Peterson LN (1997) Potassium in nutrition. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 153–183, chap. 6

Salonen JT, Nyyssonen K, Korpela H, Tuomilehto J, Seppanen R, Salonen R (1992) High stored iron levels are associated with excess risk of myocardial infarction in eastern Finnish Men. Circulation 86:803–811

Sauberlich HE (1999) Laboratory tests for the assessment of nutritional status, 2nd edn. CRC Press, Boca Raton

Shils ME (1997) Magnesium. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 117–152, chap. 5

Sunde RA (1997) Selenium. In: O'Dell BL, Sunde RA (eds) Handbook of nutritionally essential mineral elements. Marcel Dekker, New York, pp 493–556, chap. 18

WHO (1996) Trace elements in human nutrition and health. World Health Organization, Geneva

WHO (2005) Nutrients in drinking water. World Health Organization, Geneva, 186 pp

Biological Responses of Elements

9

Monica Nordberg and M. George Cherian

Contents

9.1 An Introduction to Biological Responses of Certain Elements

Medical geology is defined as the science dealing with the relationship between natural geological factors and health problems in man and animals. The geographical distribution of trace elements and metals in nature can explain "natural deficiency or toxicity," which includes the occurrence of health problems, diseases, and adverse health effects endemically seen. It is a broad subject and it attracts interdisciplinary collaboration and contributions. A relationship between humans and the ecosystem, environment and life, and environment and illness exists, and geological factors are of major interest in the environment in this respect (see also Chaps. 8 and 24, this volume).

Biological effects in response to environmental exposure of elements constitute two parts, deficiency and toxicity, which are the themes of this chapter. Under normal conditions human nutritional deficiencies may occur due to environmental, sociological, and genetic influences. Thus deficiency syndromes are an expression of multiple simultaneous effects. Functional abnormalities related to specific nutrients may be identified by the biological role of the nutrient, in controlled physiological experiments and by multivariate analysis of outcomes. Model systems and dose response experiments allow measurements of deficiency effects at different levels of physiological states and at different stages in the life cycle. Such experiments provide a basis for understanding the roles of nutrients in human physiology and the biological consequence of deficiencies. Nutritional deficiencies of certain minerals in soil have resulted in diseases in populations. For example, zinc deficiency in Iran and selenium deficiency (Keshan disease) in China were due to mineral deficiencies in soil. In most western countries, farmers are aware of this problem, and they test the soil every year and add the deficient nutrients before planting their crops. Thus they get better crops and also take care of the nutritional status of the population.

M. Nordberg (✉)
Institute Environmental Medicine Karolinska Institutet, Stockholm, Sweden
e-mail: Monica.Nordberg@ki.se

M.G. Cherian
Department of Pathology, University of Western Ontario, London, ON, Canada

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_9,

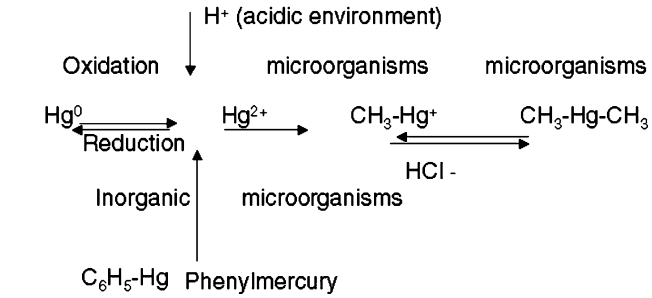


Fig. 9.1 The cycle of mercury

Primary nutritional deficiencies are caused by adverse economic conditions, customs, and food choices that limit dietary variety and thus nutrient availability. Conditioned (secondary) deficiencies are caused by non-dietary factors, especially illnesses and iatrogenic conditions that interfere with absorption, utilization, or retention of nutrients. In many instances, especially among the poor, combined effects of low intakes and conditioning factors result in deficiency disease. The conditioning factors can also give rise to toxicity.

Toxicity can occur either due to high bioavailability of elements or due to interactions of trace elements in the environment. To follow changes in the eco- and biological systems and their relationship to exposure conditions, different forms of monitoring, e.g., environmental and biological, are performed as illustrated in Table 9.1. These may involve mode of exposure, dose, and effects (see also Chap. 26, this volume).

Essentiality and toxicity of trace elements need to be in balance. Otherwise, adverse health effects to the organism will develop with cellular functional changes as the first sign of toxicity, followed by organ damage, and subsequent manifestation of illness as the end point. Most studies on toxicity of metals have been performed with the assumption that intake of essential trace elements are on a "normal basal" level. When the dietary level of essential trace elements is low, toxicity can occur at much lower concentration levels of the non-essential elements than previously regarded as toxic dose levels for these elements. The recommendation for supplementation of daily intake of essential metals in certain cases can exceed the dose that may cause toxicity. With that background a more intense collaboration between researchers in toxicology and nutrition was initiated on an international level (WHO 2002).

Toxicity of metals depends on bioavailability, species, and exposure conditions. Metals that are essential for life can also cause adverse health effects if exposure is excessive or if the chemical species or exposure route is different from the physiological route. Low iron status, and other nutritional factors such as low intake of protein, calcium, and zinc, can increase the toxicity of toxic metals such as cadmium and lead. Although an element is classified as essential to the organism it can cause toxicity with excessive exposures and interactions. The environmental level of metals in the ecosystem is influenced by industrial emissions and other human activities. Recycling of batteries and modern electronic equipment materials increases concentrations of metals such as cadmium, mercury, and semi- and superconductor materials like gallium and indium in the ecosystem. The interactions between ecosystem and these human activities may result in excessive exposure to metals (see also Chap. 4, this volume).

The acidification of lakes and soils due to emissions of sulfur and nitrogen oxides results in changes in the mobility and promotes chemical conversion of metals and may result in chemical species of certain metals with greater toxicity than the parent compound. Aluminum, arsenic, cadmium, and mercury constitute examples of elements with increased mobility whereas selenium has decreased mobility in an acidified ecosystem. With increased acidification of lakes, increased mercury concentration in the fish has been reported due to increased bioavailability of methylmercury. It is reported that the concentration of metals in fish and crops can vary with change of pH in lakes, sediments, and soils.

Environmental biochemistry of metals provides special emphasis on chemical reactions that transform metals in the environment into species or compounds of the element that are either more toxic or less toxic. For example, chemical reactions like methylation and demethylation can alter the toxicity of mercury and arsenic. It is known that both bacteria and vitamin B_{12} can methylate mercury in sediments.

Medical geology is of global concern not only for humans but also for plants and animals. Tissue samples from penguins and seals from the Antarctic that were analyzed for cadmium and copper showed increased concentrations in samples from areas that were regarded as non-polluted. This might be explained by the geology of the Antarctic. Oysters from New Zealand can be high in cadmium but binding to various bioligands can vary their intestinal absorption pattern, and thus affect toxicity.

The adverse health effects of mercury depend on the species of mercury. Figure 9.1 shows the cycle of mercury in the ecosystem. Inorganic mercury, mercury vapor, methylmercury, and various other mercury compounds exist in the ecosystem due to bioconversion, and can be transformed to a stable form of methylmercury by bacteria and vitamin B_{12}. This is the most toxic form of mercury for humans because it is easily absorbed (about 99%) from the gastrointestinal tract and retained in the body with high distribution to the brain.

Table 9.1 Different forms of monitoring of metals and their relationship to exposure, dose and effects

Emission	Exposure	Internal dose indicators	Biological effects
Sources/rates/patterns	Air	Absorbed dose	Bioindicators of effects
	Water	Body burden	Early health effects
	Soil	Target tissue concentration	Overt health impairment
Source characterization and emissions monitoring	Environmental monitoring	Biological monitoring	Health monitoring

From Clarkson et al. (1988)

Table 9.2 Factors influencing uptake of trace elements and metals

Concentration of agent
Chemical and physiological properties of the agent and the soil
Particle size
Physiological factors: age and nutritional status
Any existing disease
Genetic make up

The aim of this chapter is to illustrate the relationship between deficiency and toxicity of trace elements and to give examples of diseases related to geology-medical geology or medical geology. Special attention is paid to iodine, iron, selenium, cadmium, and mercury and this chapter also includes a few other trace elements. These elements are selected for discussion because of their known role in certain diseases or their interactions that can lead to diseases (see also Chaps. 5, 6, 22, and 25, this volume).

9.2 Metals and Geo-Environment

Human exposure to metals generally occurs through food, drinking water, and air. The drainage of contaminants from farmland is a major source of toxic metals in well water in rural areas. In several cities in developing countries, air pollution with metals has become a major problem. Occupational exposure mostly takes place via inhalation of metal fumes and dust where the trace elements exist as oxides, sulfides, or in the elemental form. Exposure to metals can be monitored in various ways: the exposure sources, the type of indicators, and the biological effects in general can be all documented in such monitoring methods (Table 9.1).

9.2.1 Toxicity

Paracelsus (1493–1541) stated that everything is toxic, it is just a matter of dose. Trace elements are important for life. Both deficiency and toxicity can cause adverse health effects. Many metals constitute part of the structure or cofactor of the enzymes and play important roles for the biological activity of enzymes and vitamins. Zinc plays an essential role in zinc-dependent enzymes such as alcohol dehydrogenase and cobalt in vitamin B_{12}.

Upon exposure to high concentrations of non-essential elements clinical disease or death can occur in exposed individuals, and in most acute toxic cases, the effect is directly related to the dose. At a lower dose range, less severe effects might occur after a long latent period. In most cases, the higher the dose, the shorter the latency period for appearance of toxic effects.

Risk assessment in environmental medicine serves as a basis of preventive action in order to avoid adverse health effects in general populations exposed to chemicals. To discuss toxicity, it is important to define concepts such as factors influencing uptake (Table 9.2), critical organ, critical dose, critical effects, and biological half-time. The organ or tissue where the exposed chemical can give rise to a critical effect is known as the critical organ. Normally, the critical organ is the organ where the earliest adverse effect occurs in an individual. Critical concentration is the concentration of the chemical, which can cause the earliest adverse effect in the critical organ in an individual or population (Nordberg and Nordberg 2002) (see also Chaps. 22 and 26, this volume).

In order to prevent toxicity and maintain quality of life, it is important to find various sensitive chemical or biological indicators for early detection of effects of metals in the critical organ. Each metal and sometimes each species of metal is unique with regard to its critical organ and toxicity. A typical example for this is the different forms of mercury and their different target organs for toxicity. Mercury vapor and methylmercury exposures can affect the brain while inorganic forms of mercury can affect the kidney. The cornerstone to develop a metabolic model constitutes knowledge of absorption, distribution, excretion, biotransformation, and biological half-time for specific species of the metal.

In order to understand the toxicity of metals and the role of geology (Tables 9.3 and 9.4), it is important to have information on concentration levels of metals in water, air, and soils in a particular geographical area (see also Part II, this volume).

Geological factors such as availability of trace elements and metals can be reflected in the occurrence of diseases in man and animals. If plants actively take up a metal, it can

Table 9.3 Health risk evaluation of agents in soil

Agents
Concentration and quantities
Exposure situation
Type of soil/rock
Mining and volcanic activity
Bioavailability
Health effects

Table 9.5 Human metal/metalloid carcinogens

As and inorganic compounds	1980
Chromium (VI)	1990
Nickel and compounds	1990
Cadmium and compounds,	1993
Beryllium and compounds	1993

Table 9.4 Examples of relationship between elements, geology and health effects as expression of medical geology

Metal	Activity	Disease
Arsenic and drinking water	Well-drilling in Bangladesh and cattles in US	Cancer
		Keracytoses
Cadmium	Japan	Itai-Itai
Cesium as 137	Radioactive release chernobyl	Cancer
Copper	Genetic defects	Menke's disease placenta
		Wilson disease liver
		Indian liver cirrhosis liver
	Copper in kitchen utensils	
Selenium	Farming in China	Kashin-Beck; Keshan
	Farming in Finland	cancer & heart disease

result in high exposures to people living in that area. The increased amount of cadmium in rice was the cause of "itai itai" disease in Japan, and also in Shipham in the UK. Nutritional deficiencies (iron, calcium, and vitamin D) also might have played a role in itai itai disease.

9.2.2 Neurotoxicity

Certain forms of metals such as methylmercury can cause neurotoxicity. Certain forms of mercury such as mercury vapor and methylmercury can easily cross the blood–brain barrier. Mercury can bind with critical ligands and cause a direct toxic effect in the cell. However, for cadmium neurotoxicity, the uptake in brain and the mechanism behind neurotoxicity are more obscure. Lactating pups of dams exposed to cadmium showed changes in the serotonin levels. Because cadmium does not pass the blood–brain barrier, this observation remains controversial. Other factors like interference of cadmium with zinc or calcium metabolism and the presence of a specific form of metallothionein (MT-3) in the brain should also be considered in the neurotoxicity of metals. Lead is known to cause central nervous system (CNS) toxic effects in children, but it may cause damage to peripheral nerves in adults because of the differences in permeability to the blood–brain barrier.

9.2.3 Carcinogenicity

A few metals or metalloids (arsenic) have been classified by the International Agency for Research on Cancer/World Health Organization, (IARC/WHO) (IARC 1980, 1990, 1994) to be carcinogenic in humans, i.e., group 1 as presented in Table 9.5. These classifications are based on epidemiological studies and detection of tumors in certain organs depending on the mode of exposures. A number of metals have been shown to be carcinogenic in animals, especially after injection of metal salts, but only a few have been classified as human carcinogens, i.e., group 1.

9.3 Protective Mechanisms

There are several binding sites for metals in the cells, and therefore, a number of mechanisms may be involved in the cell to protect it from developing toxic effects, as shown in Table 9.6.

9.3.1 Importance of Various Proteins

New families of proteins, important for protecting the cell from toxic insults by reactive oxygen species (ROS) and also

Table 9.6 Examples of protective mechanisms in metal toxicology

Metal binding proteins
Localization in the cell
Metallothionein in cytoplasm/nucleus
Lead-binding inclusion bodies in kidney
Detoxification by binding to glutathione or amino acids
Change of pH in lysosomes
Interaction between metals e.g. selenium-mercury, cadmium-zinc
Protein intake
Iron status
Genetic polymorphism
Methylation
Demethylation

heat shock proteins, have been identified after exposure to metals. Although these proteins are found in low basal levels in cells, they are inducible at a transcriptional level when exposed to metals (Piscator 1964). The genes coding for metallothioneins (MTs) and heat shock proteins are present in most of the organisms, and their induction after exposure to metals plays an important role in protection against metal toxicity. There are several review articles published on metals and stress proteins (Goyer and Cherian 1978; Goering and Fisher 1995; DelRazo et al. 2001), and they describe the specificity of certain metals to induce various heat shock proteins, and some of the similarities on gene expression for these proteins and MTs. The metal regulatory elements (MRE) present in the metallothionein gene and metal transcription factor (MTF-1) are important factors that may control the mechanisms of induction of these proteins after exposure to certain metals (Seguin 1991; Radtke et al. 1993). Both essential (zinc and copper) and non-essential (cadmium and mercury) metals can induce the synthesis of MTs and also bind to them. Thus, these proteins may have a role in the metabolism of essential metals and protection against the toxicity of metals (Cherian and Nordberg 1983; Cherian 1995, 1997).

9.3.2 Metallothionein

The structure, chemicophysical properties, and the physiological/biological functions of MTs in several organisms have been investigated for the last 55years. Metallothionein is a family of proteins with low molecular weight (6,500Da) with a unique capacity to bind seven metal ions through the 20 cysteinyl groups. Of the 61 amino acids of this protein, 20 of them are cysteines. In MT, divalent zinc and cadmium are bound tetrahedrally, with both primary and bridging sulfur bonds. It may bind with 12 monovalent copper trigonally. Two distinct metal binding domains, the α- and β-clusters have been characterized in MT by both NMR solution structure and crystallization. The purified MTs generally contain 5–10% metals as zinc, cadmium, mercury, or copper. The characteristic properties of MT (Kägi and Nordberg 1979) are listed in Table 9.7.

The definition of the MT super family follows the criteria set for polypeptides that have common features with equine renal MT. The four major forms of groups of MT consist of MT-1, -2, -3, and -4 isoforms. Mammalian MT-1 and -2 forms are present and expressed in almost all tissues. However, the number of isoforms differs, as MT-1 exists in many subisoforms, whereas only one isoform for MT-2 has been identified. MT-3 has seven additional amino acids and contains a total of 68 amino acids with differences in charge characteristics when compared to isoforms MT-1 and -2. The MT-3 isoform was identified years after characterization of MT-1 and -2 as a growth inhibitory factor (GIF) in the brain. At the N terminal region of MT-3 an additional threonine is inserted and 6 additional amino acids consisting of glutamic acid and alanine are present as a loop at the C terminal region. Thus MT-3 differs from MT-1 and -2 in its amino acid sequence. Another difference in the sequences of the MT-3 form is the presence of a proline close to the N-terminal region, which contributes to the growth inhibitory effects of MT-3. The threonine in MT-3 increases the acidity and the charge surface facilitates the interaction of MT-3 with other biological constituents. MT-4 consists of 62 amino acids with one glutamate inserted and is specific for squamous epithelium and expressed in keratinocytes. There are 14 human MT genes that are localized on chromosome 16q13–22. Of these six are functional, two are not, and six have not been characterized.

MT is often related to toxicokinetics and biochemistry of essential and non-essential metals such as zinc, cadmium, mercury, and copper. *In vivo* binding to other metals/metalloids such as selenium and bismuth is not yet understood. Although it is mainly an intracellular protein, MT has been detected in small amounts in blood and urine.

Several functions have been proposed for MT, and they are listed in Table 9.8. They may play an important role in the homeostasis of essential metals such as zinc and copper. It has been demonstrated that mammalian fetal liver contains very high levels of MT bound to these metals, and it may act as a metal storage protein during gestation similarly to ferritin as an iron storage protein. In fetal and newborn liver, MT is localized in the cell nucleus and cytoplasm, but during growth, MT is degraded and the levels of zinc and copper are maintained at a low basal level. In adults, the low levels of MT in the hepatocytes are detected in areas around the wound after partial hepatectomy and surgery. Thus, MT may serve as a storage protein for zinc during development and other conditions when the requirement for zinc is high. In addition, because of its high affinity for

Table 9.7 Characteristics of metallothionein

1. Molecular weight 6,000–7,000, 61 amino acids (aa)
2. 20 cysteine (30%), N-acetylmethionine, C-alanine, no aromatics, no histidine
3. Metal content (cadmium, zinc, copper, mercury) 5–10% w.w.
4. Absorption 250nm (cadmium), 225nm (zinc), 275nm (copper), 300nm (mercury)
5. Induced synthesis by cadmium, zinc, copper and mercury
6. No disulfide bonds
7. Heat stability
8. Cytoplasmic and nuclear localization
9. Unique amino acid sequence

Table 9.8 Functions of metallothionein

1. Metabolism of essential metals
2. Detoxification of metals
3. Protection from metal toxicity
4. Storage of metals
5. Protection from oxidative stress
6. Cellular proliferation and differentiation

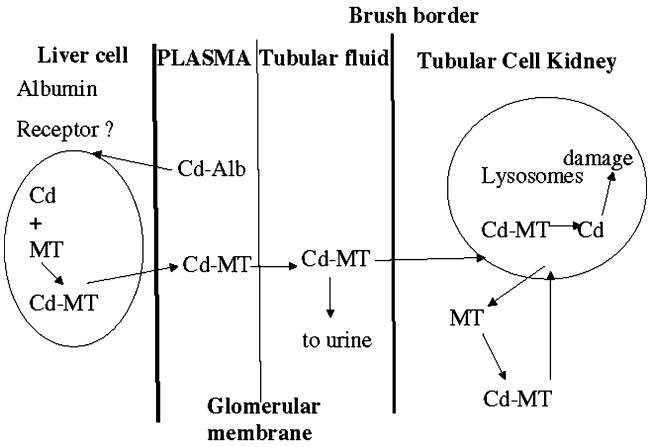


Fig. 9.2 Model of transport of cadmium in blood

metals, MT can detoxify the toxic effects of certain non-essential metals such as cadmium and mercury.

Cellular membranes are targets for metals, and MT may play a role in protection of the cell from metal toxicity. However, the release of Cd-MT from liver and other tissues can cause toxic effects. Membrane damage caused by Cd-MT in the renal tubule is most likely explained by a direct interference of cadmium on calcium transport in renal membranes and might explain chronic toxicity of cadmium on the kidney. Metals like mercury, copper, and cadmium are continuously accumulated in liver and kidneys with a major part bound to MT. Cd-MT is efficiently transported through the glomerular membrane and actively taken up by the renal tubular cells, which causes damage to the cell.

Expression of MT in different tissues is influenced by the exposure to metals and their accumulation. Therefore these proteins could be used as a biomarker of toxic metal contamination at an environmental level and as a biomarker of toxic effects in an individual.

A number of different techniques and methodologies have been used for the quantification of MTs. The detection is based on the redox properties of the thiol groups in the molecule, for electrochemical methods, on saturation with metal ions, for indirect quantification methods, and the immunoreactivity with antibodies for immunochemical methods. These methods are widely used to determine the levels of MT in tissues and biological fluids, depending on their detection limit. Still, certain technical problems exist with the determination of MT in certain biological samples.

When MT is isolated from various tissue samples, the content of metal ions may differ depending on tissue factors and the exposure levels of metals. In most cases, isolated MT is completely saturated with metals. It is well known that atmospheric oxygen can easily oxidize MTs, which results in the formation of disulfide bonds with "free" thiol groups. In isolation methods, a reduction agent, such as 2-mercaptoethanol or dithiothreitol, is used to prevent the disulfide bond (S–S) formation. The immunoassays of MT in biological samples may provide limited values because of the lack of antibodies, which can cross-react with all the forms.

Increased tissue concentration of MT-1 and -2 may indicate increased exposure to certain elements. MT-3 has not been shown to be inducible with metal exposure, but it may play a role in zinc metabolism and other elements in the brain. A mechanistic model for cadmium and MT illustrates the chronic toxic action of cadmium in renal tubular cells and the development of adverse health effects after low-dose chronic exposure.

Because of its low molecular weight and related efficient glomerular filtration in the kidney, the binding of Cd to MT in blood plasma serves as one of the modes (Fig. 9.2) of transport of cadmium from liver and other tissues to the kidney. The binding of cadmium to MT and other low molecular ligands in plasma and tissues may play a role in the tissue distribution of cadmium after uptake from the intestine. The absorption of cadmium in the gastrointestinal tract may involve divalent metal transporters (DMT-1), and it could also be influenced by MT synthesis in the intestine. Because both cadmium and iron are taken up by DMT-1, a direct competition by these metals in the intestinal level cannot be ruled out. Previous studies have shown that the gastrointestinal (GI) absorption of cadmium in humans may depend on the iron status (Flanagan et al. 1978). Uptake and distribution of cadmium occurs mainly in the initial phase in a form where cadmium is bound to albumin in plasma. A report suggests that iron deficiency can increase levels of MT-1 in bone marrow of

rats with hemolytic anemia with unchanged hepatic and reduced renal MT, and it may indicate that MT-1 levels in blood reflect erythropoietic activity (Robertson et al. 1989).

MT levels in the liver and kidneys may serve as potential indicators of environmental exposure to cadmium, zinc, and copper. The extremely long biological half-time of cadmium in the kidney (15–20years in humans) may be due to its binding to intracellular MT. Thus, MT can immobilize cadmium intracellularly, and can play a role in the kinetics of cadmium and protection from metal toxicity (Nordberg 1998; Nordberg and Nordberg 2000).

9.3.3 Metabolism and Kinetics of Trace Elements and Toxic Metals

The metabolism and kinetics of trace elements involve binding to various proteins. The specific role of MT in the binding of cadmium and zinc has been described in the previous sections.

A similar role for MT in the kinetics for copper also has been shown (Bremner 1987). In certain diseases like Wilson's disease, the excretion of copper is impaired and results in accumulation of copper in the liver. Initially, the toxicity of copper may be prevented by induced synthesis of MT, and its binding to copper. At a later stage, the cells are unable to synthesize MT, and binding capacity of MT may get saturated. The excess copper ions or copper-saturated MT may cause toxicity to liver cells.

9.4 Biological Monitoring of Metals and Trace Elements

To study the toxicity of metals in a population and protect public health, biological monitoring of metals has been performed in various populations such as lead in children and pregnant women and mercury in people with dental amalgam.

The use of MT as a potential biochemical indicator for environmental cadmium exposure has been tried by analyses of tissue samples from moose and reindeer from the northern part of Sweden and penguins and seals from the Antarctic. Cadmium was found to be a major metal component in MT even from this area that is regarded as non-polluted. Human tissue samples from autopsy have been used as potential indicators of environmental exposure to cadmium in Canadians. For bioindicators for metals and their potential use in risk assessment see Clarkson et al. (1988).

The quality of the analyses in biological monitoring depends on a number of factors. The time of collection and storage of specimens for analysis of metals or species of metal (inorganic or organic) should be controlled. The difficulty in interpretation of such results and accuracy of the method also should be considered in such studies. Among the sources of error in biological monitoring of environmental exposure are physiological factors such as variations in the biological material, age, diet, smoking, alcohol intake, and concurrent exposure to other compounds and drugs that can affect metabolic changes. Other factors may involve species-dependent kinetics of elements and simultaneous exposure to mixed species of an element.

The external contamination is a major problem in trace element analysis in biological samples during collection, and some of the special techniques can be successfully performed only by trained personnel. The contamination might arise from both the collection procedure and the storage container itself. Good hygiene should be practiced when collecting specimens, i.e., smokers should not collect samples for cadmium analyses. Contamination might also occur in sampling for zinc and aluminum where powder used as lubricant in disposable gloves can contribute to contamination. The material of the containers also can contribute to errors. Specific care has to be taken upon monitoring of aluminum. Quartz knifes are, for example, recommended for the handling of tissue samples in order to avoid contamination with chromium or nickel. Metal analysis is usually performed by atomic absorption spectrometry (flame, graphite furnace, and electrothermal), and also by X-ray fluorescence spectrometry.

9.5 Specific Metals

Several trace elements are found to be essential to human health, and they are iron, zinc, copper, chromium, iodine, cobalt, molybdenum, and selenium (WHO 2002). Other elements that might have a beneficial effect and probably are essential for humans are silicon, manganese, nickel, boron, and vanadium, and they may constitute a second group of trace elements of human concern. If they are accepted as essential trace elements, a biological requirement should be demonstrated at appropriate dose levels (WHO 2002). Some of these elements can be toxic, and those that can interact with others are discussed here.

9.5.1 Aluminum

9.5.1.1 Aluminum—Plasma

Aluminum is found in several human organs, including lungs. Pulmonary fibrosis has been reported after inhalation of aluminum in certain industrial operations and in mining. Aluminum is mainly excreted through urine, and kidney damage can increase its retention.

Aluminum administration to experimental animals can produce osteomalacia and cause neurotoxicity in certain animals with neurofibrillary tangle formation. The role of aluminum in Alzheimer's disease is still controversial. Reference values (Commission of the European Community, CEC, recommendations for patients receiving dialysis):

- Normal = $<10\mu gL^{-1}$ (no history of chronic renal failure; CRF)
- $<60\mu gL^{-1}$ = desirable in CRF patients
- $>60\mu gL^{-1}$ = excessive accumulation
- $>100\mu gL^{-1}$ = cause for concern; high risk of toxicity in children
- $>200\mu gL^{-1}$ = urgent action required; high risk of toxicity in all.

9.5.1.2 Aluminum—Water, Dialysis Fluid

Treatment of CRF by dialysis can give rise to increased aluminum concentration in the patient. Reference values (CEC Recommendations):

- Maximum allowable concentration (MAC) for potable water: $200\mu gL^{-1}$
- Guideline concentration (GLC) for potable water: $50\mu gL^{-1}$
- MAC for water for preparation of dialysis fluid: $30\mu gL^{-1}$.

9.5.1.3 Aluminum—Urine

CEC reference range shows that excretion is usually $<25\mu g/24h$.

9.5.2 Antimony

Antimony exists as V and III valence status with similar chemical characteristics as arsenic. In the geo-environment they occur simultaneously. The daily intake via food is estimated to be around 10μg/day. Exposure to antimony is rare. For biological monitoring concentration in blood and urine is an indicator of exposure and internal dose.

9.5.3 Arsenic

Arsenic is distributed in the Earth's crust with an average concentration of $2mgkg^{-1}$. Rock, soil, water, and air contain trace amounts of arsenic. Arsenic compounds are found in close association with gold as arsenopyrites in mining areas. Wines made from grapes sprayed with arsenic-containing insecticides may also contain high levels of arsenic. The chemistry of arsenic focuses on inorganic forms (−3, 0, +3, and +5) and on organic arsenic compounds of which arsenobetaine occurring in fish is the most prominent one. Inorganic arsenic of geological origin is found in groundwater and used as drinking water it gives rise to adverse health effects in several parts of the world, e.g., in Bangladesh, Hungary, and China.

The development of a variety of instrumental techniques, i.e., atomic absorption spectroscopy (AAS) and inductively coupled plasma mass spectrometry (ICP-MS) with hyphenated methods and element-specific detectors coupled to chromatographic separation techniques makes it possible to study chemical species of arsenic. A test kit that is based on the color reaction of arsine with mercuric bromide allows arsenic to be determined under field conditions, e.g., in the groundwater in Bangladesh with a detection limit of $50–100\mu gL^{-1}$.

Biochemical condition in the geo-environment enhances biotransformation between arsenite and arsenate reduction and methylation of arsenic including organoarsenic compounds. Similar to other cases, biotransformation follows pH. This is an important matter because toxicity of arsenic depends on the chemical species. Acute poisoning occurs when the arsenic blood level is above the normal range of $<10\mu gL^{-1}$ as inorganic arsenic and metabolites. Acceptable occupational exposure (ACGIH value) is $<50\mu gg^{-1}$ creatinine as inorganic arsenic and metabolites.

9.5.4 Cadmium

Cadmium, discovered in 1817, is a soft, silver-white metal and is similar in appearance to zinc. Cadmium does not have a defined taste or odor. Many radioactive isotopes of cadmium, e.g., 109 and 115m, are well recognized in experimental toxicology. Cadmium is an element with an average distribution of $0.1mgkg^{-1}$ in the Earth's crust. Cadmium is usually found associated with zinc. Particularly high concentrations of cadmium occur in some sulfide ores, but many soils, rocks, coal, and mineral fertilizers contain some cadmium. Cadmium is widely dispersed in the environment. Human exposure to low levels occurs as a result of natural processes and of human activities, e.g., mining, smelting, fossil fuel combustion, and industrial use. Due to the natural occurrence in the geo-environment and its active uptake by plants, some farming products such as tobacco could be high in cadmium content. The metal also remains strongly bound to other compounds in the soil and water (WHO 1992a, b).

Cadmium causes kidney damage with proteinuria and calciuria. Bone effects have been reported as itai itai disease in Japan and also as low mineral density of the skeleton in studies from Belgium and China. Cadmium is classified as a human carcinogen (IARC 1994).

A specific environmental exposure to cadmium occurred in Japan. Widespread exposure with both subclinical and clinical effects was found in itai itai patients in areas where water from the Jinzu river was used for irrigation. The cause of this

disease was confirmed to be due to excessive exposure to cadmium through rice. A large number of people living in the cadmium-polluted area have renal tubular dysfunction, and people in other polluted areas have the same health effects. Itai itai disease is characterized by osteomalacia and renal tubular dysfunction and is an unusual disease. The opinion is that cadmium can give rise to both osteomalacia and osteoporosis. The age, sex, nutrition, exercise, and number of pregnancies influences the type of bone effect. The kidney damage can also lead to bone damage. The subclinical bone and renal effects in cadmium-exposed populations may be common and undetected, and could contribute to more severe effects such as fractures and kidney stones.

For biological monitoring of environmental exposure, cadmium is measured in blood and urine. It should be noted that smokers have increased cadmium concentrations by a factor of 2 due to the high content of cadmium in tobacco. The reference range for blood is $<0.2\mu gL^{-1}$ for nonsmokers and $<1.4\mu gL^{-1}$ for smokers. Cadmium concentration in urine for non-smokers is $<1\mu gg^{-1}$ creatinine, and for smokers it is $<3\mu gg^{-1}$ creatinine. The high urinary cadmium may be detected only after renal damage, and by then it may be too late to treat these patients.

9.5.5 Chromium

Trivalent chromium is essential for man and animals, and it plays a role in carbohydrate metabolism as a glucose tolerance factor. The daily chromium requirement for adults is estimated to be 0.5–2μg of absorbable chromium(III). However, this is based on the calculation that if a fractional absorption value of 25% for "biologically incorporated" chromium(III) in food is assumed, then this is provided by a daily dietary intake of 2–8μg of chromium(III). Both acute and chronic toxic effects of chromium are caused by hexavalent compounds that are very toxic.

Ingestion of 1–5g of "chromate" results in severe acute toxic effects such as gastrointestinal disorders, hemorrhagic diathesis, and convulsions. Chromium concentration in serum/plasma and urine has a toxic reference range value in serum/plasma of $<0.5\mu gL^{-1}$ and for urine $<1\mu gg^{-1}$ creatinine.

9.5.6 Cobalt

Cobalt is an essential metal for vitamin B_{12}, which is involved in various methyl transfer reactions. About 25% of cobalt is absorbed from the GI tract, and it is mainly excreted in urine. Addition of cobalt to beer has caused endemic outbreaks of cardiomyopathy among beer drinkers resulting in fatalities.

9.5.7 Copper

In the general environment humans are exposed to copper via food and drinking water. Copper in drinking water has for some time been regarded as the cause of diarrhea and stomach problems in certain countries. Copper is found naturally in a wide variety of mineral salts and organic compounds and in the metallic form. It is essential for all biota. It is widely used in cooking utensils, water distribution systems, and in fertilizers. Bactericides, fungicides, algicides, antifouling paints, and animal feed additives and growth promoters also contain copper. It is also used in a number of industrial applications. Major sources for copper distribution in the environment are mining operations, agriculture, solid waste, and sludge from treatment factories. Copper that is biologically available can accumulate in tissues and give rise to high body burdens in certain animals and terrestrial plants (see also Chaps. 15 and 21, this volume).

The highest concentration of copper for the human diet is found in animal flesh, liver, oysters, fish, whole grains, nuts, seeds, and chocolate. Concentrations of approximately $10mgkg^{-1}$ have been reported. The adult daily intake of copper via food is estimated to be between 1 and 2mg and for children 2years of age it is 0.6–0.8mg. The contribution from drinking water is usually not included in these estimates. Thus the contribution of copper from drinking water can in some cases be high and populations should be alerted.

Also breastfeeding can contribute to an infant's exposure to copper. Concentrations of copper in drinking water exceeding $1–2mgL^{-1}$ give rise to staining of sanitation porcelain. Also hair turns blue-green as described for people exposed to water in, for example, swimming pools. The taste from copper influences the quality of water as reported for water of $0.3–12.7mgL^{-1}$. The taste from copper might be hidden due to the content of other additives of the drink and thus a misleading impression of copper content in water can occur. WHO recommends that the daily essential need for copper is 0.6mg for children from 6months to 6years.

The risks of copper exposure to human health for the general population are through food and drinking water when contaminated with copper. Thus, the major route of excessive copper exposure is oral. In Sweden, most of the water pipeline system consists of copper. Copper pipes were also used in several other countries. Copper is generally thought to be a good material to use for pipeline systems but under certain circumstances copper is released from the pipes, for example, acid rain can increase the bioavailability of copper.

About 30–40% of intake of copper is absorbed in the intestine and transported by albumin to the liver. After hepatic uptake, copper can be incorporated into copper-containing enzymes or into ceruloplasmin, which is then exported into the blood. Copper in the cytoplasm is

predominantly bound to MT, and any excess of copper is excreted into the bile mainly through a lysosome-to-bile pathway, which results in fecal excretion. Normally copper concentration in tissues is regulated by homeostatic control. Under abnormal conditions such as genetic diseases (e.g., Wilson's disease), biliary excretion of copper is impaired and it is accumulated in the liver, which is the critical organ of copper toxicity.

Copper is an essential metal that can be toxic when homeostatic control fails. Adverse health effects of copper can develop both from deficient and excessive intake. Copper deficiency can cause heart diseases. In the general population, copper toxicity occurs due to consumption of contaminated beverages, including drinking water. Copper can catalyze the production of hydroxyradicals and oxyradicals when it is available in Cu^{+}, a redox-active form. Both these radicals are extremely active and can attack many cell constituents, including lipids, nucleic acids, and proteins. Thus the occurrence of apoptotic bodies in the livers of copper-loaded animals is indicative of copper-induced DNA damage.

Special attention should be paid to a sensitive population with genetic disorders such as Menkes' or Wilson's disease. These are two specific inherent genetic disorders that give rise to disturbed copper metabolism. The copper transport protein, P-type ATPase is mutated in Wilson's disease, and copper accumulates in the liver, causing hepatic damage. Menkes' syndrome is characterized by disruption of copper transport from the intestine to the blood, which gives rise to copper deficiency and low activity of copper-dependent enzymes. Increased copper concentration in livers is seen in subjects suffering from Indian liver cirrhosis.

A relationship between MT and copper has been detected in patients with Wilson's disease. Induction of MT in the GI tract by oral zinc administration has been used to treat Wilson's disease by blocking the intestinal uptake of copper and decreasing the toxic tissue accumulation of copper.

MT induction in the intestine by feeding high levels of zinc can decrease copper uptake and tissue concentration. In Wilson's disease, the high hepatic copper is mainly bound to MT. This disease is due to a genetic defect in the transport of Cu into bile. Metallothionein has also been shown to be present in the placenta in patients suffering from Menkes; disease. The cause of Indian liver cirrhosis is still unknown, but high levels of copper have been detected in the liver of these patients.

Another disease is Indian childhood cirrhosis (ICC) reported mainly from India due to cooking utensils that are rich in copper. Nutritional health effects due to low and insufficient copper intake also exist and can affect the heart.

Limit values for health effects are weakly estimated. WHO has a recommended value of copper in drinking water of $2 mgL^{-1}$ based on an assumption that 10% of copper intake originates from drinking water. A provisional tolerable daily intake of copper exists since 1982; however, it is very close to the dose that causes vomiting. Tap water of $0.05 mgL^{-1}$ may be caused by corrosion of pipelines. Concentrations exceeding $0.2 mgL^{-1}$ can cause staining of sanitation porcelain (toilets and bath tubs) and hair. At 10mg copper L^{-1} there is an increased risk for taste and odor. Health effects start to occur at $2 mgL^{-1}$. This concentration increases the risk for diarrhea in children. The practical general recommendation is to allow water to run free from the tap for 1min before using the water for food preparation.

9.5.8 Iodine

In 1990, the United Nations and WHO estimated that about one billion people are at risk for iodine deficiency disorders (IDD), 211 million with goiter (enlargement of the thyroid gland), 5.1 million with severe cognitive and neuromotor deficiencies (cretinism), and many more with less severe neuropsychological defects. The loss of human capital contributes to the perpetuation of poverty and its associated social ills. Therefore, the elimination of IDD is a priority for the WHO (see also Chap. 17, this volume).

Goiter is a recognized disease of great antiquity. Some more recent advances in the understanding of iodine deficiency are listed in Table 9.9. Iodine deficiency may be understood through the context of thyroid function. The normal adult thyroid weighs 20–25g and contains 8–10mg of iodine. Goiter is the earliest and most common manifestation of IDD (Table 9.10).

Features of neurological cretinism include a wide range of mental retardation (nearly normal to severe) with associated hearing and speech disorders—the most severe is deaf-mutism—and abnormal neuromotor functions, e.g., proximal spastic rigidity of muscles of the leg with shuffling gait. Disposition is usually equable—most are able to carry out simple tasks of daily living and primitive farming and may have families and children. Growth is similar to that of the indigenous population, plasma T_4 is normal or low-normal, hypothyroidism is infrequent, and goiter may be present. Myxedematous cretinism is characterized by mental retardation, fetal hypothyroidism, persistent myxedema, growth stunting, and musculoskeletal disorders such as scoliosis, and atrophic thyroid gland. Mixed endemic cretinism is characterized by a combination of the above manifestations. The two types of cretinism may occur in the same endemia. The prevalence of cretinism in IDD endemias is 0–15% with 5–8% common. Factors that determine the prevalence and type of cretinism are incompletely understood. High-dietary goitrogens and selenium deficiency, noted below, have been postulated.

Table 9.9 Some historical advances in knowledge of iodine deficiency

Date	Name	Observation
	Chinese	The ancient Chinese were treating goitre with powdered seaweed and sea urchins several 1,000years ago
BCE	Greeks	Burnt sponge used to treat goiter
1811		The discovery of iodine by adding conc. Sulphuric acid to a seaweed of the type that was used to treat goitre
1819	Fyfe	Iodine identified in sponge
1820	Coindet	Treated goiter with iodine
1854	Chatin	Suggested low iodine in soil, water and food caused goiter
1896	Baumann	Thyroid rich in iodine
1896	Halsted	Maternal thyroid removal caused fetal thyroid hyperplasia (dog)
1908	McCarrison	Endemic cretinism characterized
1909	Marine	Maternal iodine deficiency caused goiter in fetus (dog)
1915	Kendall	Discovery of thyroxin
1917	Smith	Maternal iodine deficiency caused "cretinism" (swine)
1921	Marine	Prevention of goiter by iodide
1927	Harrington	Synthesis of thyroxin
1941	Mackenzie	Sulfanilguanidine inhibits iodide concentration by thyroid (rat)
1943	Mackenzie	Aminobenzene and thiourea inhibit iodine concentration by thyroid (rat)
1943	Mackenzie	Hyperplasia of pituitary gland in hypothyroid state (rat)
1947	Vanderlaan	Thiocyanate inhibits iodide concentration by thyroid (rat)

Table 9.10 Iodine deficiency disorders

Fetus	Malformations	
	Abortion	
	Perinatal death	
	Infant death	
	Neurological cretinism	Severe mental deficiency
		Deaf-mutism
		Spastic diplegia
		Squint
	Myxedematous cretinism	Growth-stunting
		Severe mental deficiency
	Psychomotor deficiency	
Neonate	Goiter	
	Hypothyroidism	
Child & adolescent	Goiter	
	Hypothyroidism	
	Mental deficiency	
	Low physical development	
Adult	Goiter	Mechanical compression of adjacent organs in the neck
		Endocrine disorders: hyperthyroidism; hypothyroidism; Neoplasia: benign tumors; cancer
	Mental deficiency	

The physical finding of goiter suggests the presence of iodine deficiency. Proof is provided by the concentration of iodine in urine. Concentrations $<20\mu gL^{-1}$ are indicative of severe IDD, 20–$49\mu gL^{-1}$ indicates moderate IDD, and 50–$99\mu gL^{-1}$ indicates mild IDD. Concentrations between 100 and $200\mu gL^{-1}$ are satisfactory. Thyroid-stimulating hormone (TSH) concentration in blood plasma reflects iodine nutriture. This test is used to screen newborns for hypothyroidism. In IDD endemic regions such as northern India and Zaire up to 10% of neonates were found to have increased plasma TSH, which implied severe iodine deficiency, hypothyroidism, and a high likelihood of brain damage.

Selenium is required for activity of iodothyronine deiodinase enzymes I–III that deiodinate T_4, T_3, and reverse T_3 and thus regulate the concentration of T_3. Additionally, the selenoproteins glutathione peroxidase and thioredoxin reductase, in concert with glutathione reductase, are believed to protect the thyroid gland from peroxides

produced during the synthesis of T_4 and T_3. When soils are low in selenium as well as iodine, simultaneous deficiencies can occur. This phenomenon apparently occurs in central Africa and western China. It has been speculated that selenium deficiency increases the severity of IDD and contributes to the occurrence of myxedematous cretinism.

Limited data suggest zinc nutriture affects human thyroid function. Zinc deficiency inhibits liver type I 5′-deiodinase, and lowers plasma concentrations of T_3. These effects are expressed physiologically by impaired temperature regulation. Theoretically zinc nutriture also affects the resistance of thyroid tissue to peroxidation through glutathione reductase and thioredoxin reductase. Both are flavoenzymes. Zinc is essential for flavokinase and synthesis of flavin adenine dinucleotide (FAD). Some zinc-deficient humans were found to have low plasma concentrations of T_3 and increased plasma concentrations of reverse T_3. It is possible that some zinc-deficient humans are hypothyroid. Because zinc deficiency is common among the world's poor, many of whom are at risk of iodine deficiency, interactions between low zinc status and iodine deficiency are of more than theoretical interest.

The minimal iodine requirement of adults under usual circumstances is 50–75μg/day. To meet this need and provide a margin of safety the Food and Nutrition Board (FNB), Institute of Medicine, National Academy of Medicine, United States, recommends 150μg/day for both sexes, and more during pregnancy (220μg/day) and lactation (290μg/day).

Diets containing foods of marine origin are rich in iodine. For example, marine fin and shellfish contain 300–3,000μg iodine g^{-1} as contrasted to the 20–40μg iodine g^{-1} in freshwater fish. Fertilizers that contain marine products can increase the iodine content of plants 10- to 100-fold, and iodine-enriched rations can increase the iodine in eggs and milk 100- to 1,000-fold.

Iodine in soil and water determines the iodine content of foods (see also Chap. 17, this volume). For example, in England the average daily intake of iodine is about 220μg. Similarly in the northeast United States average intakes are about 240μg. In contrast, intakes in the southwest United States are about 740μg. In Japan daily intakes are about 300μg, when little seaweed is eaten. Seaweed consumption can increase iodine intakes to 10mg daily. Approximate iodine contents of American foods are listed in Table 9.11. Of note is the wide variability. Foods raised on iodine-depleted soil are poor sources of iodine.

Prevention of primary iodine dietary deficiency requires the administration of iodine. Iodine enrichment of foods commonly consumed by the population at risk is the preferred approach. Salt is the most common vehicle. Salt iodization with KI or KIO4 is used in some areas and where used it is economical and efficacious. However, salt iodization may not be effective when customs or economic reasons cause the population to obtain salt from traditional small producers and distributors, and not from large producers equipped for iodization. In addition, iodine enrichment of salt has no effect on populations that use little or no salt. A second approach is the addition of iodine to bread. Wide variation in bread consumption and the preparation of bread in the home and small bakeries can detract from this approach. The addition of iodine to water supplies has been used in some settings. The small amount of iodized water consumed is a major limitation. When these approaches are not practical, oral administration of iodine-containing tablets, oil, or confections have been used, or iodine-containing oil has been injected. An obvious limitation of these approaches is the need for continuous cooperation by the recipients and the people responsible for delivery of the iodine. Even so, every 3years intramuscular injection of iodized oil given to women of child-bearing age has eliminated cretinism in some regions. In addition, the Chinese have used irrigation of crops with iodine-enriched water.

Table 9.11 Approximate iodine content of foods (U.S.)

	Iodine (μg/g w.w.)	
Food class	Mean ± SEM	Number of samples
Fruit	40 ± 20	18
Bread/cereals	100 ± 20	18
Dairy	130 ± 10	18
Eggs	260 ± 80	11
Meat	260 ± 70	12
Vegetables	320 ± 100	13
Marine	660 ± 180	7

9.5.9 Iron

Iron is an essential element in human nutrition but it can be toxic. Estimates of the minimum daily requirement of iron for humans depend on age, sex, physiological status, and iron bioavailability. The range is 10–18mg/day, 30mg/day if pregnant (U. S. recommended daily allowance; RDA), and 14mg/day (EU RDA).

Iron toxicity can occur at high levels of intake. The average lethal dose of iron is 200–250mgkg^{-1} of body weight. However, even an oral intake as low as 40mgkg^{-1} of body weight has been lethal. Chronic iron overload results primarily from a genetic disorder (hemochromatosis) characterized by increased iron absorption and from diseases that require frequent transfusions. Intake of 0.4–1mgkg^{-1} of body weight per day is unlikely to cause adverse effects in healthy people.

Iron deficiency, probably the most common nutritional deficiency, is believed by some authorities to affect 80% of the world's population, or about five billion people, two billion of whom are anemic. Especially at risk are young children and premenopausal women, both pregnant and non-pregnant. Low socioeconomic status (SES), high dietary phytate, and other inhibitors of iron bioavailability and low consumption of flesh foods and chronic blood loss increase the risk of deficiency. Knowledge of iron deficiency and its effects has a long and rich history. Examples of advances in understanding iron in human nutrition are listed in Table 9.12.

Iron deficiency exists when body iron stores are completely depleted and iron-dependent processes malfunction. Iron deficiency may be associated with many abnormalities (Table 9.13). However, because iron deficiency related to diet seldom occurs alone and is usually part of a syndrome of micronutrient deficiencies, a cause-and-effect relationship between iron status and some clinical phenomena is obscure or nonexistent. Controlled experiments in other species have clarified these issues (Table 9.14).

Iron deficiency related to diet and/or chronic blood loss evolves slowly. Initially iron stores are depleted and iron is less available for iron-dependent biochemical functions. This is reflected by decreased synthesis of heme, and the activity of cytochrome enzymes, aconitase, and other iron-sulfur enzymes is decreased. Iron-depleted aconitase (three irons) in intestinal mucosa, whose concentration is effected by iron in plasma, mediates mechanisms that increase intestinal iron absorption. These cellular effects occur when iron stores are depleted and before hemoglobin concentration is significantly decreased. More than half of iron-deficient people have this type of iron deficiency. Most affected individuals are unaware of its morbidity. Research data indicate that abnormalities in cold intolerance, muscle endurance, immunity, and neuro-psychological function can be detected. Other associated micronutrient deficiencies depend on the type of diet consumed. For example, diets that exclude red meat increase the likelihood of both iron and zinc deficiencies in young women who menstruate regularly.

Iron deficiency anemia occurs after all chemical indicators of deficiency are abnormal. Anemia is relatively mild in most individuals and is poorly related to symptoms or awareness of morbidity. When anemia is more severe other manifestations of the deficiency syndrome are more evident. For example, the anemia may become a "mixed anemia" as deficiencies of other micronutrients such as copper, pyridoxine, retinol, zinc, and folate become manifest. When this occurs treatment with iron alone will not fully restore erythropoiesis to normal. Other examples of conditions that may in part be caused by iron deficiency, but are also caused by associated deficiencies, include stunted growth and delayed sexual development, low immunity, impaired neuropsychological function, and abnormal pregnancy outcomes. In addition, research in animal models suggests that some clinical signs involving epithelia (skin, hair, nails, mouth, tongue, esophagus, stomach, and duodenum) that may be associated with iron deficiency are probably caused by associated micronutrient deficiencies, e.g., zinc, pyridoxine, and riboflavin.

Iron deficiency occurs when the diet provides insufficient bioavailable iron. Bleeding is a conditioning factor. Body iron content is regulated by feedback through aconitase-mediated control of iron absorption. With the exception of surface and menstrual losses, iron is reutilized.

Basal iron losses from the exterior and interior surfaces of men are about $14\mu g kg^{-1} day^{-1}$, an amount of iron readily available from usual "western" diets. Under usual circumstances premenopausal women are at risk of iron deficiency because their non-menstrual iron losses are similar to those of men, and they lose about 6–179mL of blood monthly (intrauterine devices can cause substantially greater losses). About 10% of normal women lose >80mL of blood monthly. Thus the iron needs of some women for maintenance are 2.84mg iron per day (95th percentile). Therefore women, who absorb 15% of iron when iron stores are low, must consume about 18.9mg of food iron daily to be assured of adequate iron status. This is difficult to do. Consequently, a national survey in the United States found that the 25th percentile for serum ferritin of premenopausal women was $14\mu g L^{-1}$, which suggested one out of four American premenopausal women is iron deficient by the criteria of absent iron stores.

Newborns generally have adequate iron stores for the first 4–6months of life. They then become an increased risk of iron deficiency up to 2years of age because of the low bioavailability of iron from common cereal-based weaning foods and high growth rate. Their iron requirements are about $100\mu g kg^{-1} day^{-1}$, or four times the average needs of menstruating women. Later, after age 2, growth of children slows and requirements decrease. Adolescents who are growing rapidly have iron requirements of about 20% (boys) and 30% (girls) more, respectively, than the needs of menstruating women. The requirements of girls are related to the combined effects of growth and menses.

Pregnancy and lactation substantially increase maternal iron requirements (Table 9.15). Multiple pregnancies without repletion of iron nutriture can result in severe iron deficiency. This phenomenon is especially common in societies where diets are based on unrefined cereals and legumes and flesh foods are infrequently eaten.

Bioavailability is a critical factor affecting iron adequacy. Most food iron (non-heme iron) is bound to proteins and

Table 9.12 Advances in understanding of iron in human nutrition

Date	Name	Observation
1554	Lange	Chlorosis: greenish palor and languor in young women
1661	Dr. Sydenham's practice of physick	Iron treatment for chlorosis
1713	Lemery	Iron present in blood
1832	Foedisch	Low iron content in blood of chlorotics
1832	Blaud	Ferrous sulfate efficacious for chlorosis
1846	Magendi	Dietary iron increased blood iron of experimental animals
1895	Stockman	Dietary iron of men and women, and relation to chlorosis
1897	Cloetta	Dietary iron prevented "milk anemia" in dogs
1910	Dock	Hookworm associated with growth stunting
1919	International Health Board	Hookworm associated with low intellectual performance in children and young adults
1919	Waite	Hookworm associated with retarded mental development
1925	Whipple	Iron and unidentified nutrients essential for erythropoesis
1925	Keilin	Cytochrome enzymes discovered
1927	Hart	Copper essential for iron intestinal absorption
1928	Mackay	Iron treatment of infants decreased morbidity from respiratory and gastrointestinal infections
1933	Strauss	Infants born to iron deficient women at risk of deficiency
1936	Bernhardt	Low maze learning in post-weaning rats
1943	McCance	Dietary phytate impairs iron absorption
1943	Hahn	Mucosal block theory of iron absorption
1963	Prasad	Iron and zinc deficiencies occur together
1966	Oski	Ceruloplasmin oxidizes Fe^{+2} to Fe^{+3} for binding to transferrin
1967	Sandstead	Zinc deficiency, not iron deficiency causes growth stunting
1968	Lee	Copper essential for utilization of tissue iron
1973	Webb	Low academic achievement in adolescents
1974	Cantwell	Residual neurological abnormalities in children after iron deficiency as an infant
1978	Oski	Mental development of iron deficient infants improved by iron

Table 9.13 Examples of chemical effects of iron deficiency on cells

Type	Effect
Heme proteins	Hemoglobin and myoglobin: low
	Cytochromes – cytochrome-C: low in skeletal muscle, liver, intestine, kidney; but little or no change in brain and heart
	Catalase: low in erythrocytes
Metalloflavoproteins	Monoamine oxidase: low in liver and platelets
	Aldehyde oxidase: low in brain mitochondria
	Alpha-glycerophosphate dehydrogenase: low in skeletal muscle
	Succinate dehydrogenase: low in heart and kidney
	Xanthene oxidase: low in heart and kidney
Cofactors	Aconitase: low in kidney and liver; activity in intestinal mucosa directly correlated with iron stores and inversely with iron absorption
	Protocollagen-proline hydroxylase: low
Nucleic acid metabolism	Ribonucleotide reductase: low

other food constituents and is usually about 1–5% bioavailable. In iron-deficient individuals feedback stimulation may increase absorption of non-heme iron as much as 20% if substances that bind iron and inhibit absorption are absent and facilitators are present. Flesh foods provide iron-protoporphyrin (heme iron) that is 15–35% bioavailable. Red meat is the best source in that nearly 50% of its iron is heme iron. Non-red-flesh foods contain less heme iron. The iron content of foods can be retrieved at Web sites of national agencies.

Dietary non-heme iron is in the ferric state (Fe^{+3}). Organic acids reduce ferric iron to the ferrous state (Fe^{+2}), the form in which most iron is absorbed. Absorption of ferrous iron occurs primarily in the duodenum. Nondigestible substances including phytate (hexaphosphate-inositol), certain dietary fibers, lignins, phenolic polymers, oxalate, products of

Table 9.14 Clinical abnormalities described as associated with iron deficiency in humans

Skin	Poor growth of hair
	Dry and fissured
	Infections
	Spoon shaped deformity of nails (koilonychia)
Nose	Atrophy of mucosa (ozena)
Mouth	Bilateral angular stomatitis
	Atrophy of mucosa
Tongue	Atrophy of mucosa with loss of papillae
Esophagus	Dysphagia from hyperplasia of basal cells and perakeratosis of surface cells causing a web-shaped constriction of the mucosa just below the cricoid cartilage, known as Plummer Vinson Syndrome
Gastric mucosa	Atrophy, achlorohydria, and low intrinsic factor secretion
Intestinal mucosa	Atrophy of duodenal mucosa
	Iron absorption regulated by iron depleted aconitase: increased
Liver	Reticuloendothelial cell iron: absent
Spleen	Reticuloendothelial cell iron: absent
Immune system	Morbidity from infections: increased or decreased (depends on agent)
	Circulating immunoglobulins: no change or increased
	Circulating T cells: decreased
	Lymphocyte subsets: decreased
	Lymphocyte proliferation *in vitro*: decreased, increased or unchanged
	Skin hypersensitivity to antigens: decreased
	Phagocyte function: decreased
Bone marrow	Iron: decreased to absent
	Hemoglobin synthesis: decreased
	Erythropoesis: increased
	Intermedullary hemolysis: increased
Erythrocytes	Normochromic normocytic anemia (early in deficiency)
	Hypochromic microcytic anemia
	Erythrocyte survival: decreased
	Zinc protoporphyin concentration: increased
	Transferrin receptors: increased
Blood plasma	Iron in ferritin: decreased
	Ferritin protein: decreased when acute-phase stimuli are absent
	Serum iron: decreased
	Transferrin concentration: increased
	Iron bound to transferrin: decreased
	T_4 concentrations: decreased
	Catacholamine concentrations: increased
Muscle	Exercise endurance: decreased
	Physical work productivity: decreased
	Myoglobin synthesis: decreased
Brain	Neuropsychological function: decreased
	Pica (perverse appetite: craving for and compulsive ingestion of food and non-food items)
	Affective behaviors: inconsistent symptoms, e.g., fatigue, weakness, headache, irritability, depression
Growth	Stunted
Maturation	Retarded
Fetus	Perinatal death increased; poorly reversible retarded development
Maternal status	Maternal death increased

Maillard browning, and alkaline clays bind ferrous iron and inhibit its absorption. Calcium inhibits absorption of both non-heme and heme iron. Facilitators of non-heme-iron absorption include cysteine from meat, ascorbic acid, and other organic acids.

Diets low in heme iron and facilitators of non-heme-iron absorption greatly increase the risk of iron deficiency. Such diets are common in nonindustrialized countries where poverty and/or customs limit consumption of flesh foods and cereals and legumes are the principal sources of protein. In

Table 9.15 Iron requirements of pregnancy and lactation

	Iron (mg)
Fetus	200–300
Placenta and cord	30–170
Blood lost at delivery	90–310
Maternal milk (6months)	100–180
Total	420–1,030
Daily need: gestation and 6months lactation	1–2.5

Table 9.16 Iron available to infants from human milk and cow milk preparations

Source	mgL^{-1}	% Bioavailable	Absorbed (mgL^{-1})
Breast milk	0.5	~50	0.25
Whole cow's milk	0.5	~10	0.05
Non-fortified formula	1.5–4.8[a]	~10	0.15–0.48
Iron-fortified formula[b]	10–12.8[a]	~4	0.40–0.51

[a]Common infant formulas in the USA
[b]Iron-fortified formula contains 1mg iron/100kcal formula; most iron-fortified formulas contain $680kcalL^{-1}$, equivalent to 6.8mg iron L^{-1}

developed countries food choice is an important determinant of the occurrence of dietary iron deficiency. Clay eating, a culturally determined "dietary" practice is the cause of iron deficiency in some populations.

Breast milk is adequate in iron until about 6months postpartum. Non-human milk, e.g., cows' and goats' milk, contains very little bioavailable iron. Thus infants weaned early to non-human milk are at risk for iron deficiency. Weaning foods based on cereals that are rich in phytate increases the risk of deficiency, as do formulas prepared from phytate-rich soy products. Modern iron-fortified, processed cows' milk formulas that provide 1mg iron/100kcal substantially decrease the risk of iron deficiency. A comparison between breast milk and cows' milk preparations is seen in Table 9.16.

Empirical practice suggests that iron-fortified cereals are the first foods fed infants at weaning. Recent research suggests infants also accept and tolerate weaning foods prepared from animal products rich in heme and sulfur amino acids. This innovation needs further evaluation.

Chronic blood loss is the second major cause of iron deficiency, and hookworm is the primary agent responsible. Hookworm afflicts at least two billion people. *Ancylostoma duodenale* affects the Middle East, North Africa, and southern Europe, whereas *Necator americanus* predominates in the Americas and Australia, and both also occur in central and southern Africa. The disease is endemic where the climate is warm and moist and sanitation is rudimentary. Adult worms suck blood from the mucosa of the small intestine and disperse their eggs in human feces. If dropped on appropriate soil larva mature to the filariform stage which can penetrate the skin (feet and legs are the usual points of entry) and migrate to the small intestine where they transform into adults.

9.5.10 Lead

Lead occurs in ore and soil as both inorganic and organic compounds with different types of toxicity to humans. Lead has been used in paint, in ceramics, and in home utensils. The main exposure route is via food and drinking water. It can cause neurotoxic effects in children. An early effect in adults is the interference with the hemoglobin syntheses with an increase of eryth-rocyte zinc protoporphyrin (ZPP) in whole blood. ZPP is sometimes monitored in whole blood as an indicator of lead exposure. This can also be seen in iron-deficiency anemia. Humans suffering from intermittent protoporphyria are considered a vulnerable group for lead toxicity.

In acute and chronic poisoning from environmental exposure to inorganic Pb the reference range in blood for adults is usually $<100\mu gPbL^{-1}$. For children, the levels should not exceed $40\mu gPbL^{-1}$ because of its neurotoxic effects.

The metabolism and distribution of organic Pb differs markedly from the inorganic form, and urine is used as the medium for biological monitoring. Monitoring of inorganic Pb in urine is only recommended during chelation therapy. After the removal of tetraethyl lead from gas, the concentration of lead in blood in children has markedly decreased.

9.5.11 Manganese

Manganese is an essential trace element with an estimated daily nutritional requirement of $30–50\mu gkg^{-1}$ of body weight. Manganese is required for several enzymes involved in carbohydrate metabolism. Absorption rate is influenced by actual intake, chemical form, and the presence of other metals, such as iron and copper, in the diet. In infants and young animals very high absorption rates of manganese have been observed.

The biological half-time of manganese in humans has been determined to be about 12–35days, depending on the nutritional status. The major route of excretion of manganese is through the biliary system, and a low tissue level is maintained by this mechanism. Because of its chemical similarity to iron, manganese also binds to iron-binding proteins such as transferrin. Interactions between manganese, iron, and lead have been shown in rats (Malhotra et al. 1984). Although manganese is not very toxic, its increased tissue accumulation can cause toxic effects in the brain and lung.

Excessive absorption of manganese from lungs can increase its accumulation in the brain. Brain damage has been reported in miners exposed to manganese dioxide dust, and it can progress to irreversible brain injury similar

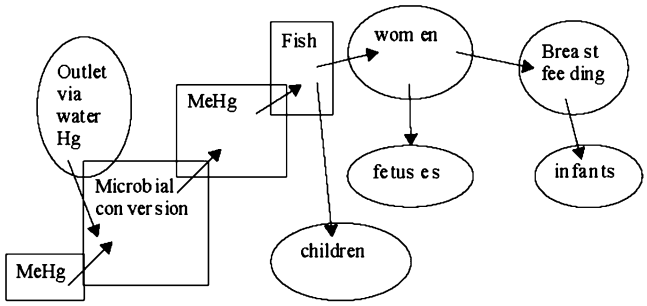


Fig. 9.3 Mercury exposure via the nutritional chain with vulnerable groups indicated in bold

to Parkinson's disease. The most important biochemical effects of manganese are on the metabolism of various neurotransmitter substances such as dopamine. The selective injury of catecholamine neurons by manganese is similar to 6-hydroxydopamine-induced neuronal injury, and it may be related to a generation of toxic free radicals. Certain manganese compounds such as manganese carbonyl compounds have been used as anti-knocking agents to replace lead in several countries. The long-term effects of these compounds on people, especially the elderly, should be monitored because of its known effects on catecholamines in brain and its relation to Parkinson's disease. The interactions of other metals on the neurotoxicity of manganese should also be studied to understand the mechanisms involved in metal-induced neurotoxicity.

9.5.12 Mercury

Mercury is found mainly in food, especially fish as methylmercury, and it is the major source of mercury exposure in various parts of the world. Exposure to mercury gives rise to adverse health effects such as neurobehavioral disorders and is an issue of public health concern. Cognitive impairment in children, following exposure of methylmercury during pregnancy, has been reported from several epidemiological studies from New Zealand (Kjellström et al. 1986), from the Seychelles (Clarkson 1997), and from the Faroe islands (Grandjean et al. 1997). The message of public health concern is that women of child-bearing age should not consume species containing methylmercury (Fig. 9.3).

The major source of exposure to mercury vapor in the general population is from dental amalgam fillings. Dental amalgam consists of about 50% mercury along with other metals such as silver and copper. They have been used widely for the last 150years because they are long lasting, inexpensive, and easier to use than other types of dental fillings. There are reports of mercury poisoning in dentists and dental technicians who handle amalgam as well as reports of side effects in patients. However, the use of amalgam in dental practice has been controversial because of the assumed toxic effects of mercury. It is known that amalgams can release mercury vapor into the oral cavity and if inhaled up to 80% is absorbed in the lungs. What is absorbed is transported to the bloodstream and part of it may enter the CNS through the blood–brain barrier, but a great part of it may be excreted in the urine. Normally the absorbed mercury vapor is excreted in the urine. About 80% of inhaled mercury vapor is taken up. The normal mercury level in urine is less than $0.05 mgL^{-1}$, and the maximum allowable level is $0.15 mgL^{-1}$. The threshold limit value (TLV) for mercury vapor is $0.05 mgm^{3-1}$, and is calculated for a working schedule of 8h/day and 40h/week.

9.5.12.1 Mercury—Blood

The blood level of mercury is a useful indicator for both acute and chronic exposure to organic mercury compounds; but it may be of limited use for acute exposure to vapor or inorganic salts. Reference range is $<4 \mu gL^{-1}$.

9.5.12.2 Mercury—Urine

Urinary levels of mercury are useful to measure exposure to mercury vapor and to inorganic mercury salts. Analysis of mercury in urine has been recommended for people concerned about its release from dental fillings. Reference range: excretion usually $<10 \mu g/g$ creatinine/$<10 \mu g/24h$.

According to the IPCS/WHO Health Criteria Document on inorganic mercury (1991), when mercury vapor exposure is above $80 \mu gm^{3-1}$, $25–80 \mu gm^{-3}$ corresponding to a urinary mercury level of $100 \mu g/g$ creatinine, the probability of developing the classical neurological signs of mercury intoxication such as tremor, erethism, and proteinuria is high. Exposure to mercury vapor in the range of $25–80 \mu gm^{3-1}$, corresponding to urinary level of 30–100μg mercury/g of creatinine can lead to increases in the incidence of certain less severe toxic effects.

9.5.13 Molybdenum

Molybdenum is considered to be an essential trace element in both animals and humans. Safe and adequate intake levels suggested for the population are 0.015–0.04mg/day for infants, for children of age 1–10 the level is $0.025–0.15 mgday^{-1}$, and $0.075–0.25 mgday^{-1}$ for humans over 10years of age.

A disease of Swedish moose is explained by interference of molybdenum with copper; see Chap. 21.

9.5.14 Selenium

Selenium is an essential trace element. Recent studies have indicated that selenium exerts a beneficial effect on coronary disease mortality, and that selenium plus garlic produces significant anticancer activity. In the Scandinavian countries the intake of selenium is low due to the fact that the soils are poor in

selenium. Also, due to acid rain selenium has less bioavailability. Depending on the soil content selenium concentration varies in grain. Countries like New Zealand, Finland, and Sweden, with a low concentration of selenium in soils, report intakes below 70μg/day and from China low intakes are reported as below 20μg/day. High concentrations are reported from Scotland, Venezuela, and from certain parts of China of up to 200–700μg/day (see also Chap. 16, this volume).

An estimated daily intake in Sweden is about 24–35μg. A recommended daily intake in the United States is 55μg for ages 19 and older. For pregnant women the recommended daily intake is 60μg and during lactation the recommended rate is 70μg. In some areas this dose is exceeded. Deficiency occurs when the daily intake is 10μg. Toxicity occurs at intakes of 500μg/day. Health effects related to deficient intake, which occurs at daily intake below 20–30μg occur in muscles and the heart. At high intakes above 1,000μg gastrointestinal irritation, hair loss, and nerve damage occur. Acute exposure may give rise to liver damage. Reported increased incidence of asthma in populations in New Zealand is associated with low selenium intakes.

Selenium deficiency gives rise to heart diseases like Keshan disease, which is most frequently seen in children. Another disease associated with selenium deficiency is Kashin-Beck. Cancer is connected to selenium and influences diseases of the muscles and joints and rheumatics and senility. High intakes of selenium influence the occurrence of caries, the garlic smell of breath, and the blue staining of nails. Selenium intoxication of cattle has been treated with arsenic supplement.

A selenium level in serum between 60 and 120μgL^{-1} reflects sufficient intake of selenium via food. Selenium concentration in hair exceeding 5μgg^{-1} reflects selenium poisoning. Selenium in plasma/serum is measured upon indications of deficiency and/or toxicity and can measure recent (months) changes in intake of or exposure to selenium. In humans the activity of the Se-dependent enzyme glutathione peroxidase is of interest. Reported reference values for neonates to 16years of age are 30–115μgL^{-1} and are age dependent and for adults ranging from 70 to 130μgL^{-1}.

9.5.15 Thallium

Thallium (Tl), is a soft metal and the oxidation states are I and III. The salts are highly toxic. Environmental samples contain μgkg^{-1} or less. However, the determination level for minerals is 20μgkg^{-1} and for aqueous solutions it is about 0.1μgL^{-1}. Thallium can be found in phyllosilicates and in sulfide deposits.

Bioavailability of thallium increases when pH decreases in soil. Thallium can also be leached to ground and surface water. It has a strong tendency to accumulate in aquatic life and plants can easily take up the element by the roots.

Oral intake of (Tl1) 20–60mg thallium kg^{-1} body weight is lethal within 1week. Thallium(III) oxide, which is water soluble, shows a somewhat lower acute toxicity compared to thallium(I) salts. The U. S. Environmental Protection Agency (EPA) suggests that drinking water exceeding action levels can lead to gastrointestinal irritation and nerve damage in the short term, and to changes in blood chemistry, damage to liver, kidney, intestinal, and testicular tissues and hair loss in the long term.

The major symptoms of acute intoxication are anorexia, vomiting, depression, and hair loss. Respiratory failure is lethal. The same symptoms are reported for chronic intoxication. Loss of hair is a typical sign of intoxication caused by thallium. The reference range for acute and chronic poisoning for blood is $<1\mu gL^{-1}$ and for urine $<1\mu gL^{-1}$.

9.5.16 Zinc

Zinc (Zn) is an essential trace element found in all food and potable water as salt or an organic complex. The principal source of zinc is normally the diet. Zinc in surface and groundwater usually does not exceed 0.01 and 0.05mgL^{-1}, respectively, and concentrations in tap water can be much higher as a result of dissolution of zinc from pipes. A daily dietary requirement of zinc of 0.3$mgkg^{-1}$ of body weight was proposed by the Joint FAO/WHO Expert Committee on Food Additives (JECFA) in 1982. At the same time JECFA also sug-gested that the provisional maximum tolerable daily intake for zinc is 1$mgkg^{-1}$ of body weight. The dietary reference values for adults range from 6 to 15mg/day. Drinking-water containing zinc at levels above 3mgL^{-1} may not be acceptable to consumers.

Zinc deficiency occurs in human beings when intake is low. This depends on the intake of diets that are low in readily bioavailable zinc. Unrefined cereals are rich in phytate and dietary fibers which all bind zinc and prohibit the bioavailability of the metal. This means that even if food is high in zinc content the intake can give rise to deficiency. This has been described as a public health problem in certain countries such as Egypt.

Zinc in serum is measured upon indications of deficiency. Another way of finding out zinc status is simply by asking about number of meals with red meat per week. This can provide a crude estimate as to whether zinc deficiency should be suspected.

9.5.16.1 Zinc Deficiency

History

Examples of advances in understanding of the nutritional role of zinc are given in Table 9.17. Reference values for zinc status are that <0.5mgL^{-1} may indicate zinc deficiency and 0.5–0.7mgL^{-1} might be of no clinical significance. The "normal" range for all ages is regarded to be 0.7–1.6mgL^{-1}

Table 9.17 Advances in understanding of the nutritional role of zinc

Date	Name	Observation
1869	Raulin	Essential for *Aspergillus niger*
1887	Lechartier	Presence in living tissues
1905	Mendel	Constituent of respiratory pigment of the snail *Sycotypus*
1910	Mazé	Growth factor for maize
1919	Birckner	Proposed "nutritive function" from its "constant occurrence" in human and cow milk
1926	Somner	Essential for higher green plants (sunflower, barley)
1927	Hubbell	Improves growth of mouse
1934	Todd	Deficiency stops growth and causes alopecia (rat)
1941	Follis	Deficiency causes dermatitis with acanthosis, parakeratosis and inflammation; and parakeratosis and basal cell hyperplasia of esophagus (rat)
1952	Mawson	Deficiency causes atrophy of testicular germinal epithelium, epididymes and prostate (rat)
1955	Tucker	Deficiency causes stunting, dermatitis with parakeratosis, diarrhea and death (swine)
1956	Vallee	Possible zinc deficiency in humans with alcoholic cirrhosis
1958	O'Dell	Deficiency in *Aves* causes dermatitis, and retards feathering, osteogenesis and lymphoid tissue
1959	Winder	Facilitates DNA, RNA and protein synthesis of *M smegmatis*
1960	Blamberg	Deficiency is teratogenic (chicks)
1961	Prasad	Deficiency thought to cause stunting and hypogonadism in teenage Iranian farmers
1962	Lieberman	Facilitates DNA synthesis in cultured mammalian cells
1962	Miller	Deficiency in calf has effects similar to those in swine
1963	Prasad	Stunted teenage farmers in Egypt have zinc metabolism consistent with zinc deficiency
1963	Sandstead	Zinc treatment improved growth and development of stunted teenage farmers in Egypt
1964	Ott	Deficiency in lambs has effects similar to those in swine
1966	Hurley	Deficiency teratogenic for rat

whereas $>1.6 mgL^{-1}$ might reflect the use of dietary supplements.

Zinc in urine is measured upon indications and metabolic studies or for chelation therapy. It is, however, of little value in assessing deficiency. This analysis may be of value in monitoring the effect on zinc body burden of long-term chelation therapy of other trace elements. The reference value for excretion is usually 0.3–0.6mg/24h.

9.6 Summary

To understand the full context of the relationship of trace elements to the geo-environment, it is important to encourage interdisciplinary research and to coordinate the knowledge obtained from various scientific fields. This approach may provide a better understanding of the mechanisms involved in both nutritional deficiency and toxicity of trace elements and metals. The nutritional deficiencies can arise from lack of the essential elements in the drinking water or food. This can be due to lack of these elements in the soil where the food is grown or can be due to the eating habits of the people. Certain genetic defects can also affect the absorption or transport of the elements to required tissues. Toxicity might occur when high concentrations of metals in soil and drinking water lead to high exposure to metals. The potential risk of developing toxicity depends on the bioavailability of the specific trace element. These influencing factors are shown in Table 9.18.

Table 9.18 Factors influencing relationship for trace elements and health effects

Eco – and biological system related to exposure, dose and effect;
Environmental exposure in the ecosystem
Environmental and biological monitoring
Acidification and species of metals and trace elements
Bioavailability, species and exposure conditions and toxicity of metals
Metals, geo-environment and environmental biochemistry
Metabolism and kinetics of trace elements
Importance of various proteins such as metallothionein for binding of elements

See also the Following Chapters. Chapter 5 (Uptake of Elements from a Chemical Point of View) • Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 7 (Biological Functions of the Elements) • Chapter 8 (Geological Impacts on Nutrition) • Chapter 22 (The Impact of Micronutrient Deficiencies in Agricultural Soils and Crops on the Nutritional Health of Humans)

Further Reading

Bremner I (1987) Nutritional and physiological significance of metallothionein. In: Kägi JHR, Kojima Y (eds) Metallothionein II. Birkhauser Verlag, Basel, pp 81–107

Cherian MG (1995) Metallothionein and its interaction with metals. In: Goyer RA, Cherian MG (eds) Toxicology of metals. Springer, Berlin, pp 121–132

Cherian MG (1997) Metallothionein and intracellular sequestration of metals. In: Sipes IG, McQueen CA, Gandolfi AJ (eds) Comprehensive toxicology, vol 3. Pergamon, Oxford, pp 489–500
Cherian MG, Nordberg M (1983) Cellular adaptation in metal Toxicology and metallothionen Toxicology 28:1–15
Clarkson TW (1997) The toxicology of mercury. Crit Rev Crit Lab Sci 34(4):369–403
Clarkson TW, Friberg L, Nordberg GF, Sager PR (eds) (1988) Biological monitoring of toxic metals. Plenum Press, New York
DelRazo LM, Quintanilla-Vega B, Brambila-Colombres E, Calderon-Aranda ES, Manno M, Albores A (2001) Stress proteins induced by arsenic. Toxicol Appl Pharmacol 177:132–148
Flanagan PR, McLellan JS, Haist J, Cherian MG, Chamberlain MJ, Valberg LS (1978) Increased dietary cadmium absorption in mice and human subjects with iron deficiency. Gastroenterology 74:841–846
Goering PL, Fisher BR (1995) Metals and stress proteins. In: Goyer RA, Cherian MG (eds) Toxicology of metals. Springer, Berlin, pp 229–266
Goyer RA, Cherian MG (1978) Metallothioneins and their role in the metabolism and toxicity of metals. Life Sci 23:1–10
Grandjean P, Weihe P, White RF (1997) Cognitive deficit in 7-year-old children with prenatal exposureto methylmercury. Neurotoxicol Teratol 19(6):417–428
IARC (1980) Some metals and metallic compounds, vol 23. IARC, Lyon
IARC (1990) Chromium, nickel and welding, vol 49. IARC, Lyon
IARC (1994) Beryllium, cadmium, mercury and exposures in the glass manufacturing industry monograph, vol 58. IARC, Lyon
Kägi JHR, Nordberg M (eds) (1979) Metallothionein. Birkhauser Verlag, Basel, pp 1–378
Kjellstrom T, Kennedy P, Wallis S, Mantell C (1986) Physical and mental development of children with prenatal exposure to mercury from fish. National Swedish Environmental Protection Board Report 3080, Stockholm, Sweden
Malhotra KM, Murthy RC, Srivastrava RS, Chandra SV (1984) Concurrent exposure of lead and manganese to iron-deficient rats: effect of lipid peroxidation and contents of some metals in the brain. J Appl Toxicol 4:22–30
Nordberg M (1998) Metallothioneins: historical review and state of knowledge. Talanta 46:243–254
Nordberg M, Nordberg GF (2000) Toxicological aspects of metallothionein. Cell Mol Biol 46(2):451–463
Nordberg M, Nordberg GF (2002) Cadmium. In: Sarkar B (ed) Handbook of heavy metals in the environment. Marcel Dekker, Inc., New York, pp 231–269
Piscator M (1964) On cadmium in normal human kidneys together with a report on the isolation of metallothionein from livers of cadmium-exposed rabbits. Nord Hyg Tidskv 48:76
Radtke F, Heuchel R, Georgiev O, Hergersberg M, Garglio M, Dembic Z, Schaffner W (1993) Cloned transcription factor MTF-1 activates the mouse metallothionein 1 promoter. EMBO J 12:1355–1362
Robertson A, Morrison JN, Woods AM, Bremner I (1989) Effects of iron deficiency on metallothionein-1 concentrations in blood and tissues of rats. J Nutr 119:439–445
Seguin C (1991) A nuclear factor requires Zn^{2+} to bind a regulatory MRE element of the mouse gene encoding metallothionein-1. Gene 97:295–300

WHO/IPCS Environmental Health Criteria Documents for Various Metals

WHO/IPCS (1991) Environmental health criteria document 118 inorganic mercury. WHO, Geneva
WHO/IPCS (1992a) Environmental health criteria document 134 cadmium. WHO, Geneva
WHO/IPCS (1992b) Environmental health criteria document 182 thallium. WHO, Geneva
WHO/IPCS (2001a) Environmental health criteria document 224 arsenic, 2nd edn. WHO, Geneva
WHO/IPCS (2001b) Environmental health criteria document 221 zinc. WHO, Geneva
WHO/IPCS (2002) Environmental health criteria document 228 principles and methods for the assessment of risk from essential trace elements. WHO, Geneva

Part II

Pathways and Exposure

Ron Fuge

Institute of Geography and Earth Sciences,
Aberystwyth University, Aberystwyth, UK
e-mail: rrf@aber.ac.uk

Introduction

It has long been said that "we are what we eat"; however, in terms of medical geology we are in fact what we eat, drink, and breathe. The major pathways into the human body of all elements and compounds that are needed for well-being, together with those that are harmful, derive from the food and drink we consume and the air we breathe. Although direct absorption through the skin can also serve as a pathway, this is of relatively minor importance. The composition of the air, water, and food is directly influenced by interactions with geological media. It should, however, be stated that air, soil, and water are subject to chemical changes due to anthropogenic activities, as discussed in Chap. 4. Although pollutants from this source can have a dramatic effect on the composition of air, water, and food, which can in turn seriously impact human and animal health, this book is essentially concerned with the influence of natural sources of the elements on human health. Therefore, anthropogenic sources are largely ignored in this section. The chapters in the section discuss the importance of the various pathways, together with examples of the influence of various elements from different sources, on human and animal health.

The atmosphere is of particular importance for life on Earth because it is the source of essential oxygen. However, in addition to the oxygen, the air inhaled can contain substances that are detrimental to health. As outlined in Chap. 10, volcanic eruptions can directly result in large loss of life due to phenomena such as explosive vulcanicity, pyroclastic flows, and lahars. In addition, volcanic eruptions can result in the emission of large quantities of fine ash and dust that can be inhaled and cause serious health problems. Similarly, volcanic gases, such as the sulfur gases and hydrogen fluoride, can be extremely harmful to exposed populations, both from inhalation and from skin contact. The radioactive gas radon, derived from the natural decay of uranium and thorium, can be emitted during volcanic eruptions and can also leak out of crustal rocks that contain uranium and thorium minerals (see Chap. 11). Inhalation of radon can increase the risk of lung cancer.

The trace element chemistry of terrestrial water is essentially controlled by its interaction with rocks and soils. The composition of the water deriving from precipitation depends essentially on geography. Because rain and snow are derived from evaporation of seawater, they contain some remnant salts of marine origin, the content being controlled essentially by the proximity of the coast. Thus, near the coast precipitation is enriched in sea salts, whereas well inland the influence is much weaker. However, once precipitated, the water interacts with soils and surface rocks, whereas some penetrates into the lithosphere to become groundwater, where it comes into more intimate contact with the rocks through which it flows and in which

it becomes stored. Thus, the chemistry of the water that animals and humans consume can be greatly modified by interaction with the lithosphere, particularly in the case of groundwater, which can be many millions of years old.

Two elements that can have a marked effect on health are arsenic (Chap. 12) and fluorine (Chap. 13). Chronic exposure to high concentrations of these elements is most often through drinking water. Consumption of arsenic-rich drinking water is having serious and detrimental health effects on large populations in several countries. Fluorine, occurring in groundwater as the fluoride ion, is essential in small quantities and harmful in larger quantities. Again, the harmful effects of drinking fluoride-rich water are a serious problem in many countries.

Water hardness is essentially dependent on the concentrations of calcium and magnesium in the water. It has been suggested that the hardness of drinking water can influence human health (Chap. 14), with harder waters appearing to offer some protection against cardiovascular disease. It appears that magnesium is the important component in this.

Plants take up nutrients from soil through their roots and from the atmosphere through their aerial parts. However, the major pathway for most trace elements is from soil. To that end, it is important to understand soil formation and the chemistry of soils, together with bioavailability of elements from this source (Chap. 15). Soil chemistry is very strongly influenced by the chemistry of the parent materials. Thus, the origin of most trace elements in plants is the lithosphere, from which soils are derived through weathering. Consumption of plants then becomes a major pathway into animals and humans.

Chapters 16 and 17 deal with the geochemistries of selenium and iodine, respectively, with particular reference to their behavior in soils and their bioavailability. These two elements have long been known to be essential in trace quantities. Although excesses of both, particularly selenium, are detrimental to health, it is their deficiencies in many areas of the world that has resulted in large geographically defined areas where endemic diseases related to their low concentrations and bioavailability have been and are prevalent.

In addition, soil can have a more direct pathway into humans through its ingestion intentionally, geophagy, or by involuntary consumption (Chap. 18). This can also result in the ingestion of soil microbes, some of which can be pathogenic (Chap. 20). Wind-ablated soil and dust particles can also represent a pathway for pathogen ingestion and inhalation. Such dusts, particularly in drylands, can also cause serious respiratory problems and can represent a pathway into the body of many elements, some of which may be harmful (Chap. 19).

Chapter 21 deals with medical geology in relation to animals. This is a particularly important topic because animals, whose movements and sources of food are more limited than humans, are more frequently affected by deficiency and excess diseases related directly to geology. They are also part of the food chain.

Humans are generally dependent on crops as a food source and the final chapter (Chap. 22), deals with trace element deficiencies in agricultural soils and the resultant effects on human health.

Volcanic Emissions and Health

10

Philip Weinstein, Claire J. Horwell, and Angus Cook

Contents

10.1 An Introduction to Volcanic Toxicology

Volcanoes provide a conduit by which magma—the molten rock, gases, and water within the earth—may interact with human biological systems (Fig. 10.1). Because of the range of materials that are ejected during eruptions, the consequent effects on human health are diverse. Contact may occur dramatically and immediately for people living close to the vent, such as from pyroclastic density currents or the emission of large projectiles. Alternatively, effects on health may occur slowly or at great distances from the volcano as a result of dispersal of volcanic material such as ash and aerosols.

The vast majority of volcanogenic fatalities in the past few centuries have resulted from 'proximal' events—such as pyroclastic density currents, lahars, and suffocation or building/roof collapse from ash or debris—and 'distal' events, such as tsunamis, which may travel for hundreds of miles from the active site, and indirect consequences of eruptions, such as famine or infectious disease outbreaks. Witham (2005) placed the death toll over the last 400 years at almost 300,000, with seven eruptions dominating the historical record (including Tambora in 1815, Krakatau in 1883, Mount Pelée in 1902, and Nevado del Ruiz in 1985). Apart from the obvious thermal and physical injuries resulting from an eruption, ejecta may also contain toxic elements and compounds that disrupt biological systems. These compounds may be released in the form of volcanic gases or carried with volcanic matter falling from eruptive columns or plumes. Some of the material ejected has the potential to induce disease by undergoing radioactive decay: among the best documented of such products is radon. Although these toxic agents are often not the major causes of mortality in volcanic eruptions, they may persist and have the potential to cause long-term morbidity.

Considering the pathological consequences of exposure to volcanic toxic compounds is a complex task. There is no simple, predictable path between the 'emergence' of a toxic

P. Weinstein (✉)
Barbara Hardy Institute, University of South Australia, Adelaide, Australia
e-mail: philip.weinstein@unisa.edu.au

C.J. Horwell
Institute of Hazard, Risk and Resilience, Department of Earth Sciences, Durham University, Durham, UK

A. Cook
School of Population Health, University of Western Australia, Crawley, Perth, Australia

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_10,

Fig. 10.1 Vent of an active volcano: White Island, New Zealand (Photo: Michael Durand)

Fig. 10.2 Krafla fissure eruption, Iceland (Photo: Olle Selinus)

agent from the magma to the eventual health consequences it imparts to a particular individual. This chapter reviews the varied mechanisms by which the emissions associated with volcanism may compromise health and generate disease. In general, these mechanisms may be influenced by one or more of the following factors:

1. Eruptive variables
2. Toxicant-specific properties
3. Patterns of toxicant dispersal and persistence
4. Biological variables

Eruptive variables—The nature of the eruption (or other volcanic event) influences the duration of emissions, the chemical composition of the toxic compounds expelled, and the range of dispersal. For example, eruptions may be broadly grouped as *explosive* (releasing large quantities of gas and fragmented magma, as ash and blocks, as with Mount St. Helens, 1980), *effusive* (associated with lava flows and gas but less ash, as with the basaltic volcanoes of Hawaii) (Fig. 10.2), or *mixed* (switching between the two styles). The magnitude and style of eruption may be measured using the Volcanic Explosivity Index, which incorporates variables such as the volume of ejecta (total erupted product) and the plume height (Newhall and Self 1982).

Toxicant-specific properties—These primarily pertain to the chemical and physical properties of toxic compounds. Volcanic products vary in terms of particle size, composition, surface area, concentration, pH, and water solubility. All of these factors can influence the bioavailability of toxic agents, and thus their patho-physiological effects. A particular cause for concern is the generation of large amounts of respirable crystalline silica during volcanic eruptions.

Patterns of toxicant dispersal and persistence—In terms of evaluating possible effects on human populations, physical proximity to the vent or eruptive site is an important component of risk assessment. Populations who fall within the 'near-vent' range may be exposed to the full array of ejected materials, often at high concentrations (Fig. 10.3). By contrast, areas that are distal from the volcanic site (or are less vulnerable to volcanic products for some other reason, such as the presence of a natural barrier) tend to be exposed to a smaller range of toxic compounds, and at concentrations less likely to result in injury. There are exceptions, however, with toxic agents such as fluoride adhering to fine, wind-dispersed ash particles and thereby occurring in highest concentrations some distance from the vent. In this discussion, near-vent is used loosely to refer to an area extending tens of kilometers from the eruptive site.

The mode of toxicant dispersal should also be considered. Eruptive products may travel along many routes, and in a variety of chemical forms, before finally entering human biological systems. Carriage in the atmosphere and hydrosphere are the most common modes of dispersal, but poisoning may also occur as a result of volcanic products entering the soil and food chain.

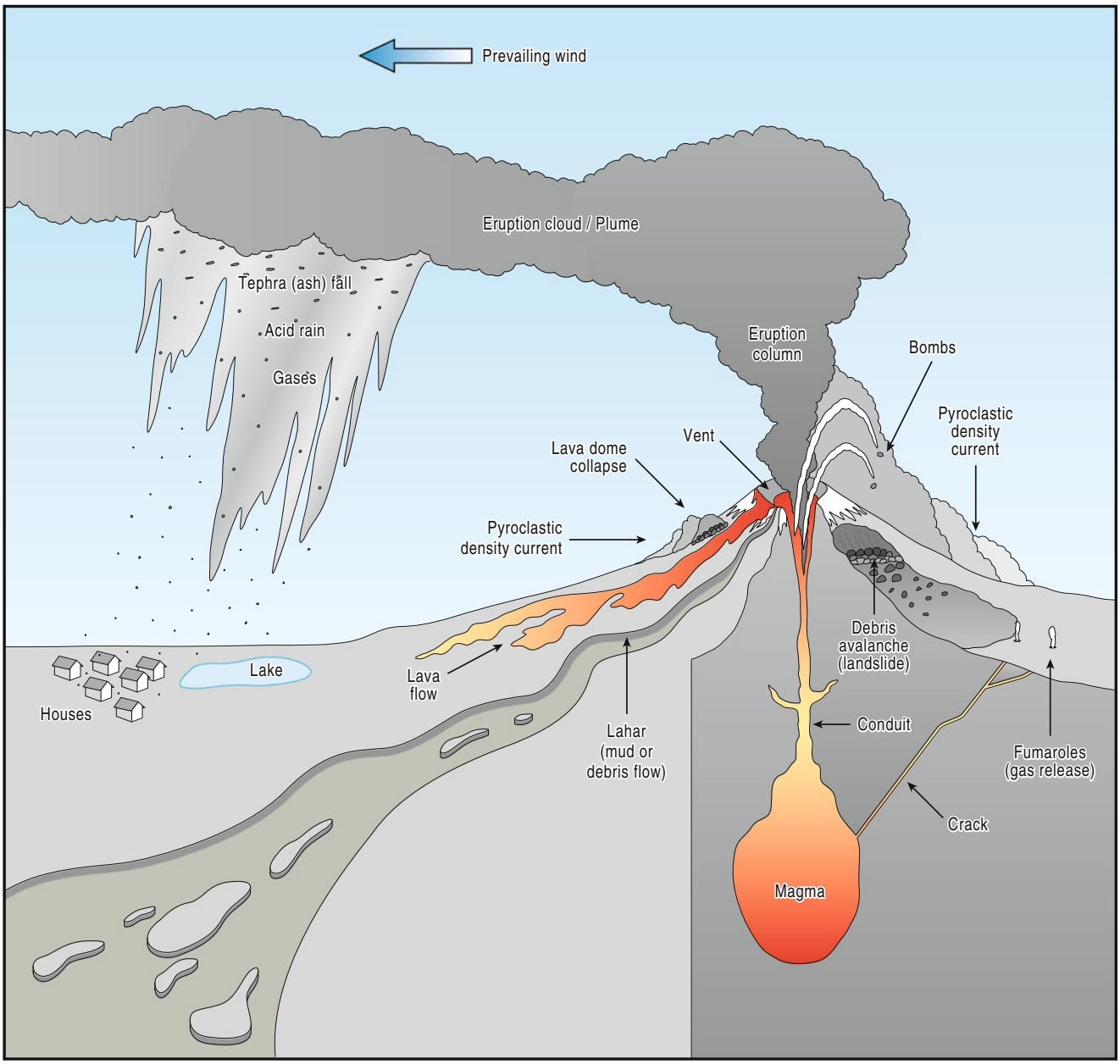


Fig. 10.3 Schematic of potential near-vent volcanic hazards for all eruption styles (Adapted from USGS poster 'Geological hazards at volcanoes' (Myers and Driedger 2008))

The duration of exposure plays one of the most crucial roles in determining health outcomes. For example, some exposures may be short-lived and reversible, as with conjunctival irritation from ash particles, or may be chronic, as with inhalation of ash particles over a long-lived dome-forming eruption. Some toxic compounds, such as radon and gases, may persist in volcanic products, or continue to be released from the volcano (and continue to cause injury) long after the eruptive event ceases. These patterns of injury, and the body systems that are predominantly affected, are discussed in detail in Sects. 10.3, 10.4, and 10.5.

The properties of the environment also affect the pattern of dispersal and settlement of volcanic toxicants, including features of the physical environment (geography, air pressure, and climate) or man-made surroundings (for example, the opportunities for asphyxiant gases to enter low, enclosed spaces such as cellars).

Biological variables—The mechanism of damage must be considered at a histopathological level. Volcanic products have the capability to cause injury in human tissues and cells, either individually or in combination, in the following ways: (1) direct physical interaction (e.g., skin contact with

acidic gases); (2) acute interactions of particles and gases with the upper airways, triggering asthma and bronchitis in susceptible individuals; (3) initiation of a chronic process of damage and repair (e.g., fibrosis from the deposition of respirable silica particles in lung tissue); (4) metabolic disruption (e.g., with carbon monoxide toxicity); or (5) genotoxicity and genetic alteration (e.g., from exposure to carcinogenic agents, such as radon).

Finally, the characteristics of the affected individuals play an important role in determining health outcomes. Important parameters include age, the presence of coexisting cardiac or respiratory diseases, and access to appropriate diagnostic services. There is also an inherent response spectrum, both phenotypic and genotypic, in any human population and some individuals are more susceptible than others. Such variation in susceptibility is discussed in Sect. 10.3.3.

10.2 The Health Effects of Tephra Dispersal

10.2.1 Atmospheric Dispersal of Toxic Compounds

Tephra dispersal is a major cause of morbidity following eruptions. Tephra (fragmented rock) thrown into the atmosphere may cause disease through the fallout of particles from eruption columns or plumes on human populations, or through the movement of individuals into eruptive clouds (such as aircraft passengers and crew). The emission of large fragments of rock, such as 'blocks' and 'bombs' may cause severe physical injury including lacerations and fractures. Heavy fallout (especially of pumice or ash) can lead to burial and asphyxiation, either directly or through roof collapse. Smaller particles of pumice, scoria, and ash may be distributed over a wide area around the eruptive site and, in some cases, plumes may affect settlements situated hundreds or thousands of kilometers away. This chapter considers the health effects of tephra by primarily examining the effects of ash (the constituents of tephra less than 2 mm in diameter). The pathophysiological effects of the compounds discussed are summarized in Table 10.1.

The eyes are particularly vulnerable to the emission of fine tephra particles. Common ocular injuries include abrasions of the cornea and conjunctivitis from accumulation of ash in the conjunctival sac (Blong 1984). Ocular irritation has been reported in people using contact lenses, because of the interposition of matter between the contact lens and the eye. Swelling of the eyelids and other facial tissues around the eyes has been reported less frequently. Symptomatically, ash has been associated with ocular redness, itchiness, throbbing pain, and discharge. Superficial tissues such as the skin, lips, mouth, and other mucous membranes may also be exposed. Nasal and throat irritation occurs, and high rates of nasopharyngeal irritation and nasal stuffiness were observed in a group of loggers following the Mount St. Helens, 1980 eruption (Baxter et al. 1986). A less well-documented effect of tephra is skin irritation, including in the axillary area, following deposition of volcanic ash particles on the skin—the so-called "ash-rash." Irritation of superficial mucous membranes, eyes, and exposed skin commonly resolves shortly after the exposure to the ash ceases, and longer term toxic injury to such body structures following ash-fall is unusual.

The lungs may be exposed to any particulate matter able to penetrate into the respiratory passages. This 'inhalable' portion of tephra refers to particles less than 100 μm in diameter, with particles >10 μm lodging in the upper respiratory tract. Particles <10 μm diameter (the 'thoracic' fraction) may enter the bronchioles, and those under 4 μm, the 'respirable' fraction, may penetrate further into the terminal bronchi and alveoli. It has been postulated that acute disease may tend to be initiated by the thoracic fraction and chronic disease by the respirable fraction in the alveoli (Horwell and Baxter 2006). The proportion of respirable ash varies greatly across eruptions with the greatest amounts generated by explosive eruptions which interact with water, known as phreato-magmatic eruptions, or from collapse of silicic lava domes (Horwell 2007). High levels of total suspended particles (TSP) caused by ash-fall may exacerbate some pre-existing respiratory complaints including asthma and bronchitis (Baxter et al. 1981; Baxter 1983; Yano et al. 1990). An increase in the frequency of bronchitis was also detected after the 1996 Mount Ruapehu eruption in New Zealand, despite the small amount of ash and the relatively low respirable portion (Hickling et al. 1999). The probable mechanism by which ash produces such respiratory symptoms is by provoking hypersecretion of mucus and bronchoconstriction (narrowing of the air passages): both are reversible, however, and diminish once exposure ceases (Buist 1988).

Dispersal of tephra may also produce health effects over a longer duration, and respiratory diseases are among the most common subacute, recurring, and chronic problems in this context. Baxter et al. (1983) reported higher rates of asthma and bronchitis exacerbations for over 3 months after the 1980 Mount St. Helens eruption. There appeared to be a clear correlation with respirable ash, which persisted in the ambient air of surrounding settlements long after the eruption. This effect was less prominent following the 1991 Mount Pinatubo eruption, despite the high volume of tephra. This lower degree of respiratory morbidity may have occurred because subsequent rainfall acted to settle volcanic dust and limit opportunities for respiration of particles. Alternatively, the chemical nature of the respirable portion of ash may have differed significantly between the two eruptions: since 2008, such differences have been routinely

Table 10.1 A review of major toxic compounds of volcanic origin and their potential pathophysiological effects

Toxic agent (active form)	Mode of dispersal	Mechanism of injury	Acute systemic effects	Chronic or recurrent systemic effects
Sulfur compounds				
Sulfur dioxide, SO_2; sulfur trioxide, SO_3; sulfuric acid, H_2SO_4	Gas emissions during eruptions; lava flows; degassing episodes	Acidic irritant	**RESP**: upper airway irritation; pneumonitis; pulmonary edema; acute ARDS[a]	**RESP**: recurrent or prolonged exacerbation of respiratory disease[a]
			HEENT: nose and throat irritation; conjunctivitis	
			DERM: skin irritation	
Hydrogen sulfide, H_2S	Gas emissions during and after eruptions	Irritant; asphyxiant; inhibition of metabolic enzymes	**GENERAL**: headache; nausea; vomiting; confusion; collapse; paralysis of respiratory centres[a]	
			GIT: diarrhea	
			GU/REPRO: pain on urination	
			RESP: cough; shortness of breath; pulmonary edema	
			HEENT: eye and throat irritation	
Fluoride compounds				
Fluoride compounds (including related acidic gases, aerosols and liquids)	Gas emissions during eruptions; ash leaching	Acidic irritant on inhalation or contact with skin, conjunctiva, or mucous membranes	**GENERAL**: hypocalcaemia (low serum calcium); low serum magnesium; collapse and shock[a]	**RESP**: permanent lung injury from toxic inhalation **HEENT**: mottling and /or pitting of teeth
			RESP: (e.g., HF inhalation) coughing; laryngeal spasm; bronchitis; pneumonitis; pulmonary edema; acute ARDS[a]	
			GIT: nausea, vomiting, gastrointestinal hemorrhage[a]	**MSS**: osteoporosis; osteosclerosis; calcification of ligaments and tendons; kyphosis of spine with bony exostoses
			GU/REPRO: nephritis	
			HEENT: eye and throat irritation	
			DERM: severe, slow healing burns; may be absorbed through skin causing internal effects	
Chlorine compounds				
Hydrochloric acid (HCl)	Gas emissions during eruptions; lava (e.g., with seawater contact)	Acidic irritant	**GENERAL**: collapse[a]	**RESP**: permanent lung injury from toxic inhalation
			RESP: coughing; laryngeal spasm; pneumonitis; pulmonary oedema; acute ARDS[a]	
			HEENT: eye and throat irritation	
Carbon compounds				
CO	Gas emissions during eruptions	Noxious asphyxiant; binds to hemoglobin	**GENERAL**: collapse; coma; "cherry red" skin[a]	**NEURO**: permanent neurological impairment from brain injury
			NEURO: headache; impaired dexterity	
CO_2	Gas emissions during eruptions or lake overturn	Inert asphyxiant	**GENERAL**: asphyxia; collapse[a]	
Minerals				
Crystalline silica	Ash plumes	Mineral dusts may initiate inflammatory response and fibrosis (scarring)	**RESP**: acute exacerbation of respiratory disease (e.g., asthma)	**RESP**: chronic silicosis
Metals				
Mercury vapour, Hg	Gas emissions during eruptions	Oxidant irritant	**RESP**: bronchitis; pneumonitis; pulmonary edema	**NEURO**: neurotoxicity
			NEURO: neurotoxicity (May lead to acute or chronic mercury intoxication[a])	

Key to major body systems: *CVS* cardiovascular system; *DERM* dermatological system (skin and adnexa); *GENERAL* includes multi-organ, metabolic, and endocrine effects; *GIT* gastrointestinal system; *GU/REPRO* genitourinary and reproductive systems; *HEENT* head, eyes, ears, nose and throat, including the scalp, face, eyes, and adnexa, ears, nasal cavities, sinuses, pharynx, oral cavity, or dentition; *MSS* musculoskeletal system; *NEURO* neurological system; *RESP* respiratory system

[a]Potential cause of mortality

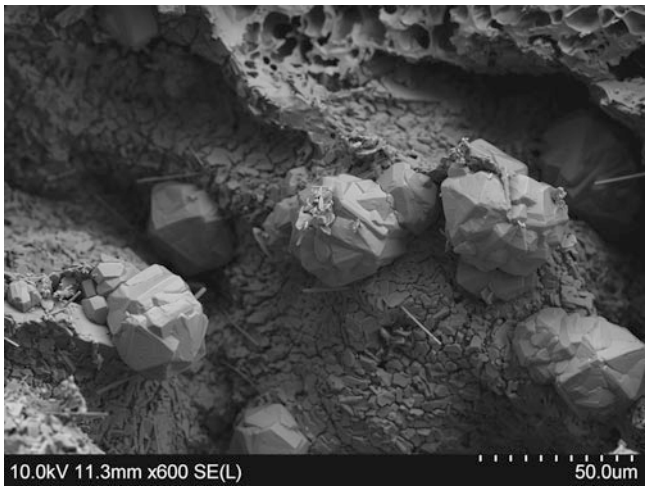


Fig. 10.4 Vapor-phase cristobalite crystals sitting on a 'crust' of vapor-phase feldspar within a pore space of Soufrière Hills, Montserrat dome rock (Image: David Damby)

measured and are important with respect to health protection (see Sect. 10.7).

In terms of chronic lung pathogenesis, one of the most troublesome compounds produced by volcanic activity is crystalline silica. Certain forms of silica, such as quartz, cristobalite and tridymite, occur in felsic eruptions (where bulk silica is >63 wt.% SiO_2) and may be formed in several ways: (1) quartz may grow as a crystal within the ascending magma; (2) cristobalite and tridymite may grow as vapor-phase or devitrification products within lava domes (Fig. 10.4); (3) Hydrothermal alteration of volcanic edifices may result in devitrification of glassy groundmass, forming cristobalite and/or quartz (Horwell et al. 2013a). Vapor-phase crystallization within lava domes is by far the most efficient mechanism of cristobalite growth at volcanoes, with up to 20 wt.% of bulk dome-collapse ash at Soufrière Hills volcano, Montserrat being cristobalite (Horwell et al. 2013b). Inhalation of fine particles of crystalline silica, including quartz, is a well-established cause of both acute and chronic inflammatory reactions in lung tissue (see also Chap. 19, this volume). If large quantities of crystalline silica are inhaled over a short period, exposure may result in the appearance of inflammatory cells in the interstitium and filling of alveoli with proteinaceous material. If exposure to silica is prolonged, the particles may evoke a chronic inflammatory reaction which recurs until the lung tissue displays extensive signs of fibrosis (tissue repair with collagen fibers), known as silicosis. In the chronic phase of silicosis, the classical pathological features of the disease are lung nodules, which are masses of concentric collagen fibers with a rim of silica dust and macrophages (phagocytes) containing trapped silica particles. These whorled masses slowly enlarge and may impinge on airways and the vasculature of the lung. It is has also been suggested that respirable dusts from volcanic eruptions may have carcinogenic effects as crystalline silica is classified as a human carcinogen by the International Agency for Research on Cancer (IARC 1997).

The mechanisms by which crystalline silica evokes these reactions are probably related to the chemical reactivity of the particle surface. SiOH groups form bonds with, and ultimately damage, proteins and phospholipids embedded in cellular membranes. Cell death may result. Furthermore, silica interacts with macrophages, causing the activation of these cells and the release of inflammatory mediators, such as interleukin-1. This process triggers the migration of other inflammatory cells to the alveoli, which amplifies the body's response and encourages intense deposition of collagen fibers.

During the Mount St. Helens eruption, the crystalline silica content (sometimes referred to as 'free silica') was 3–7% of the sub-10 μm fraction of ash. Due to this relatively low content, it was estimated that the chances of silicosis developing even in high-exposure groups (such as cleanup crews or loggers) was negligible (Baxter et al. 1986). The risk of silicosis is thought to be greater for those occupationally exposed (e.g. gardeners; 2–4% risk of silicosis over 20 years continuous exposure, Hincks et al. 2006) on the island of Montserrat, which has experienced many years of dome-forming eruptions of the Soufrière Hills volcano starting from 1995. At this location, the crystalline silica content has been much higher (10–24% by weight of the sub-10 μm fraction), and the exposure much more prolonged (Baxter et al. 1999). Nevertheless, to date, no cases of silicosis have been diagnosed as a result of inhaling volcanic ash. This may be due to difficulties in conducting clinical investigations where long-term follow up is required. However, toxicological and mineralogical investigations have produced evidence that volcanic crystalline silica may not be as toxic as some other types of crystalline silica, due to inherent characteristics of the silica and external factors such as the close association of other minerals (Cullen et al. 2002; Horwell et al. 2012).

Within the broader classification of dust-related lung diseases known as pneumoconioses (which encompass inflammatory lung reactions that occur from contact with respirable mineral dusts, but that are distinct from asthma or bronchitis), volcanic ash also has the propensity to cause mixed-dust pneumoconiosis, where dust macules or fibrotic nodules occur following exposure to mixed, non-fibrous silicate dust. Following the 1980 Mount St. Helens eruption, workers exposed to tephra over a long duration, such as loggers and road gangs around the eruptive site, were examined for the advent of pneumoconiosis. One deceased victim appeared to have been suffering from distinctive granulomatous lesions of the lungs, a chronic inflammatory

reaction similar to that observed in rats exposed to respirable tephra (Green et al. 1982). However, a longitudinal study of survivors found that respiratory symptoms largely disappeared 3 years after exposure, and no permanent changes in chest X-rays were observed relating to ash exposure (Buist et al. 1986). Animal studies indicated that volcanic ash from various eruptions (including Mount St. Helens ash) displayed less cytotoxicity than crystalline silica, and was more readily phagocytosed by macrophages in the alveoli (Vallyathan et al. 1984).

Mafic, basaltic rock (<52 wt.% bulk SiO_2) may also have the potential to be toxic despite not containing any crystalline silica. Basaltic ash can generate substantial amounts of the potentially inflammogenic and carcinogenic hydroxyl radical, by an iron-catalysed reaction with endogenous hydrogen peroxide (Horwell et al. 2007).

Other minerals, formed by metamorphic alteration of rock of volcanic origin, including fibrous silicates and zeolites such as talc, erionite and asbestos may also cause prolonged fibrotic diseases such as asbestosis. These asbestiform minerals have been linked to endemic pleural disease and a high rate of mesotheliomas (a highly lethal malignancy which typically infiltrates the pleura or peritoneum) (Rohl et al. 1982), but are not considered further here.

10.2.2 Hydrospheric Dispersal of Toxic Compounds

During and following volcanic activity, tephra particles may affect water supplies in two principal ways. First, they may be deposited onto bodies of water, which include irrigation or filtration plants, thus rendering the water highly turbid and unusable. Secondly, tephra may carry a variety of adsorbed chemicals. Plumes of ash, which may cover hundreds of kilometers, may effectively disperse such 'stowaway' toxic agents. After the ash settles to the ground, these agents may leach into the environment, often by rain water. The dominant chemicals which may adsorb to tephra, and thus act as leachates, include chlorine (Cl), sulfur (S) compounds, sodium (Na), calcium (Ca), potassium (K), magnesium (Mg), and fluorine (F). Other elements that are present in smaller quantities are manganese (Mn), zinc (Zn), copper (Cu), barium (Ba), selenium (Se), bromine (Br), boron (B), aluminum (Al), silicon (Si), and iron (Fe) (Witham et al. 2005). Furthermore, even light tephra-falls may have significant effects on water pH. During the 1953 eruption of Mount Spurr (Alaska), a 3–6 mm tephra fall on Anchorage caused the pH of the town water supply to fall to 4.5, returning to 7.9 after a few hours (Blong 1984). Such excursions in pH and turbidity can alter the chlorine demand at treatment plants, adding a significant, albeit indirect, microbial water hazard.

Some of these elements and compounds have safety levels established in drinking water and could potentially cause harm if ingested in quantities exceeding these concentrations. There is a paucity of conclusive studies relating to the health effects of ash leachates, although some effects may be inferred from accidental or occupational exposures. An example is provided by fluoride, a relatively common volcanic product, for which numerous cases of toxicity have been reported in various settings.

Fluoride is rapidly and effectively absorbed following ingestion (up to 90% absorption from the gastrointestinal tract), and most of the ionic fluoride remaining in the body has a high affinity for calcified tissues, particularly bones and developing teeth. Despite its usefulness in reducing the development of dental caries, fluoride may produce toxicity if ingested in high concentrations. Acute fluoride poisoning can result in death, such as following the accidental contamination of public water systems (Gessner 1994). Acute symptoms of excessive fluoride ingestion include nausea, vomiting, excessive sweating, and abdominal discomfort (probably from the formation of the corrosive hydrofluoric acid in the stomach) (Grandjean 1982). Clinical and laboratory evidence indicates that kidney damage (nephritis) may also result, thus disturbing water and sodium balances in affected patients. Higher concentrations of fluoride may produce cardiovascular convulsions, collapse, and shock partly as a result of fluid loss from vomiting and gastrointestinal bleeding. Fluoride may also lower levels of calcium in the blood, possibly by a process of precipitation in which insoluble CaF_2 is formed. Painful and involuntary muscle contractions may result from the depletion of available calcium. Although not recorded in humans following volcanic eruptions, these signs and symptoms are well-established in grazing livestock in Iceland (see below) (see also Chap. 13, this volume).

Chronic ingestion of fluoride may cause a number of low-grade symptoms, including loss of appetite, headache, vertigo, and joint pain. The most characteristic effects, however, involve the dentition and skeletal system. Although not directly a result of recent ash leaching, elevated fluoride levels in potable water adjacent to the Furnas volcano in the Azores have resulted in dental fluorosis, a condition characterized by mottling of the teeth (Baxter and Coutinho 1999). Mottling—the appearance of chalky white patches distributed irregularly over the teeth—develops if chronic fluorosis occurs during the development phase of permanent dentition. The enamel is structurally weak, and may eventually become pitted, in which case the affected teeth acquire a corroded appearance. Such dental changes are often the only signs of chronic fluorosis.

In areas where waters are highly fluoridated, however, more serious problems may emerge. For example, teeth and gums may degenerate. Skeletal symptoms start to appear,

which include back pain and limitation of spinal movement. Studies of fluoride exposure in various contexts have indicated that, as a result of the fixation of bone calcium by fluorine, bones undergo a number of changes: increased hardening and rigidity (sclerosis) of bones; the development of bony 'projections' (osteophytes and exostoses) on the ribs, pelvis, and vertebrae; increased thickness of long bones; and calcification of the ligaments (Grandjean 1982). The results of long-term fluoride exposure may thus be crippling.

Other evidence relating to emission of ash rich in soluble fluorine have been obtained from studies of effects on other mammals. It is known that the ingestion of high concentrations of fluoride is highly toxic to livestock (Baxter 1983). Following the 1947 eruption of the Icelandic volcano Hekla, a mere 1-mm deposition of fluoride-rich ash was sufficient to kill thousands of sheep. Icelanders in the eighteenth century also suffered from the long-term consequences of fluorosis following the eruption of Laki (1783). During this eruption, in which over 140 vents formed along a 27-km fissure, it is estimated that eight megatons of fluoride compounds were discharged. (Fluoride may also be expelled from volcanic vents as unadsorbed gas, including fluorine, F_2 and HF; volcanic gases are discussed subsequently.) The deposition of high concentrations of fluorine on pastures and waterways proved fatal for numerous animals. Death occurred from at least two causes. First, the volcanic discharge resulted in extensive crop damage, and thus stock starved from loss of feed. In addition, ingestion of fluoride from pastures and waterways caused direct toxicity (see also Chap. 21, this volume).

Records of the period indicate that fluoride may also have directly poisoned the Icelandic population. A contemporary account describes a curious disorder affecting farmers living around the eruption site. These symptoms included severe swelling in many joints that was associated with painful cramps. The most striking effects, however, appeared to affect the oral cavity: sores appeared on the palates and tongues of the victims, and eventually teeth and blackened portions of gum fell from their mouths. Although it is difficult to exclude the effects of scurvy, Blong (1984) suggested that the illness may be the result of toxic fluorosis. In modern times, studies of meltwater around ash particles from Mount Hekla, Iceland, show fluoride levels up to 2,000 ppm. Although rapidly diluted to about 200 ppm in standing water and 20 ppm in running surface water, these fluoride concentrations could readily produce symptomatic fluorosis if ingested by humans, and there is circumstantial, historical evidence that chronic ingestion in Viking-age Iceland may have resulted in disease (Weinstein 2005).

Determining areas at risk from fluoride leachates remains difficult, and the relationship between concentrations and distance from the volcanic vents is highly variable. For example, during the Mount Ruapehu eruption in New Zealand in 1995, it was concluded that the levels of leachable fluoride in adjacent water supplies did not pose a threat to human health (Weinstein and Patel 1997). Following the Popocatepetl eruption on May 12, 1997, however, levels of fluoride were maximal at a distance of 13.4 km from the eruptive site. It was predicted that had such ash entered the water supplies, fluoride levels would have exceeded the national standard of 1.5 mg/L (Armienta et al. 2001). Such results suggest that the possibility of significant water contamination by fluoride should be considered during and following the eruptive phase.

10.2.3 Toxic Compounds in the Soil and Food Chain

Volcanic material may also enter the human body by direct ingestion of contaminated foods or soils. As with the hydrospheric effects, toxic chemicals may be carried into the soil or food supplies by adsorption to tephra particles and subsequent leaching. (This is discussed in Sect. 10.3. Emissions of volcanic gases also contribute to chemical deposition.) Usually ingestion of these compounds is unintentional, but humans may consume volcanic products such as mineral-rich muds for therapeutic or nutritional reasons.

Selenium is an example of a volcanic compound that may potentially produce health effects through food borne exposure. This element may be emitted in substantial concentrations (e.g., 630 kg of aerosolized selenium per day during a 1976 eruption of Mount Etna (Faivre-Pierret and Le Guern 1983)), thus providing unusually large Se loads for soils and edible plants. In some districts in China, cases of chronic toxicity from the consumption of foods with high levels of selenium have been reported, possibly occurring in conjunction with fluorosis (Yang et al. 1988). Signs of chronic exposure include loss of hair and nails, skin lesions, and abnormalities of the nervous system. Elevated selenium levels may increase the risk of dental caries. The exact mechanism of selenium toxicity remains uncertain, although it has been suggested that the element may cause disruption of sulfur metabolism, inhibition of protein synthesis, or oxidation of sulfhydryl groups. The role of selenium and cancer is debated. Some studies indicate that there is a protective effect against cancer, and others have argued that selenium displays pro-oxidant—and thus possibly carcinogenic—effects (Spallholz 1994) (see also Chap. 16, this volume).

The effects of tephra on the food chain and the hydrosphere intersect through the damage caused to aquatic life. During the 1979 eruption of Karkar, Papua New Guinea, heavy ash-falls (and possibly small lahars) tainted the headwaters of surrounding rivers, which resulted in the

death or contamination of aquatic life. However, edible fish, eels, and prawns continued to be consumed by local farmers, despite complaints of the sulfurous taste (Blong 1984). The health effects of this form of food contamination are unknown.

10.3 The Health Effects of Volcanic Gas Emissions

10.3.1 Description of Volcanic Gases

The gaseous substances produced by volcanic activity are varied, and may be classified as follows:

1. Gases and vapors: The gaseous state of an element which normally exists in a liquid or solid form and can be readily reverted to this form by decreases in temperature or increases in pressure (such as mercury)
2. Aerosols: Droplets or particles suspended in a gaseous medium
3. Fumes: Aerosols of solid particles, usually less than 0.1 μm in size, usually formed by escape of volatiles from molten materials
4. Smoke: Volatile gases and particles, usually less than 0.5 μm in size, produced by combustion (Kizer 1984).

The pathophysiological effects of the compounds discussed are summarized in Table 10.1.

Steam, from both magmatic and superficial sources (such as overlying lakes or groundwater), is the most common volcanic gas. Other, often very toxic, gases are also emitted during eruptive events and may impact on human health (Fig. 10.5). There are numerous accounts of volcanic gases causing death. In Japan, for example, around 35 people were killed during the twentieth century by becoming overwhelmed by volcanic gas emissions (Hayakawa 1999).

Among the most notorious gases are those which are heavier than air, such as CO_2 and H_2S. These may pool at ground level and result in asphyxia. HF, sulfur dioxide (SO_2) hydrochloric acid (HCl), hydrogen (H), helium (He), carbon monoxide (CO), and radon may also be produced in considerable quantities. Vaporized metals, such as mercury (Hg), are also found in plumes. Gases may be released even in the absence of obvious volcanic activity (passive degassing).

In assessing the impact of volcanic gases, it is not always possible to separate the toxicological effects of each gas released during a particular eruption. Numerous gases may be emitted simultaneously, or the gaseous components or concentrations may vary over time as the eruptive process evolves. For example, the sequence of activity around Rabaul, Papua New Guinea, over the last 70 years included: relatively quiet solfataric emissions of CO_2 and H_2S (prior to 1937); a violent steam explosion resulting in over 500 fatalities (between May and June 2, 1937); the appearance

Fig. 10.5 Volcanic gas emissions (Photo: Michael Durand)

of vigorous vents producing large clouds of SO_2, with periodic emissions of HCl, gradually increasing until another major eruption commenced (1937–1941); a release of CO_2 in a pit crater at the Tavurvur vent, causing the death of six people (1990); an eruption from the two major vents with an accompanying SO_2 plume with an estimated size of 45,000 km^2, and a maximum SO_2 mass of 80 kt (1994), producing large quantities of ejecta, widespread building collapse, the evacuation of over 50,000 residents from Rabaul; and the emission of a large bubble of flammable gas, presumed to be methane, which ignited and initiated bush fires (1997).

Although not as dramatic as full-scale eruptions, other types of geothermal activity (including hot springs and fumaroles) also generate a variety of toxic gases (Sparks et al. 1997). Because such activity may often continue unabated over prolonged periods, and because the benefits of hot springs may encourage the presence of humans, there is a risk of toxic exposure. For example, fumaroles are very numerous in some areas (such as Yellowstone Park) and may emit high levels of numerous gases including CO_2, SO_2, HCl, and H_2S. A full discussion of geothermal toxicants is beyond the scope of this chapter.

In terms of adverse impacts on human health, volcanic gases may be classified into the following groups: gases with *irritant effects* on the respiratory system; those which act as *inert asphyxiants*; and those which combine both properties and act as *noxious asphyxiants*.

10.3.2 Inert Asphyxiants

Carbon dioxide, CO_2, illustrates the effect of an inert asphyxiant gas: it replaces oxygen, but does not have a directly toxic effect on biological tissue. Concentrations of CO_2 are particularly high near emission vents (Faivre-Pierret and Le Guern 1983). The degassing of volcanic soil may

result in the collection of carbon dioxide in cellars, huts, and in low-lying areas (Baubron and Toutain 1990; Baxter and Coutinho 1999). Low concentrations (e.g., under 5%) produce accelerated breathing, and often feelings of discomfort, by direct activation of the respiratory centers in the brain. Headache and vertigo are early symptoms. If sufficient concentrations are reached (for example, concentrations of 7–10% for a few minutes), fainting occurs. Elevated levels of CO_2 in the bloodstream (hypercapnia) eventually result in circulatory failure and death from acidosis.

In Java's Dieng Volcanic Complex, it is believed that emissions of CO_2 from a fissure resulted in the deaths of 149 people (Le Guern and Faivre-Pierret 1982). The dramatic effect of CO_2 emissions is also illustrated by two disasters in Cameroon at Lake Monoun in 1984 and at Lake Nyos in 1986. It is thought that large volumes of CO_2 accumulated at the lake bed and were then released as the lakes' thermal stratification became disrupted, causing overturning of the water as CO_2 came out of solution (known as a limnic eruption). The CO_2 over-topped the craters and, carried downward by gravity, engulfed whole villages in their path. As a result, 37 people were asphyxiated near Lake Monoun, and the death toll reached 1,700 at Lake Nyos. The survivors from the Lake Nyos disaster reported falling in a deep state of unconsciousness for up to 36 h. It appears that no long-term respiratory effects occurred in the survivors, although some sustained burns by falling into cooking fires during a period of CO_2-induced coma (Wagner et al. 1988; Afane et al. 1996). In areas with the highest CO_2 concentrations, a variety of animal life ranging from insects to livestock also perished.

10.3.3 Irritant Gases

Volcanic gases which have primarily *irritative* (directly injurious) *effects* include the hydrogen halides, HF and HCl. At low doses, eye and throat irritation may occur. At higher levels, both may cause ulceration of the respiratory tract upon inhalation and corrosive burns upon contact with the skin or mucous membranes. Cutaneous burns from contact with HF are particularly severe and notoriously slow to heal. Fatalities may occur from pulmonary edema (the accumulation of unwanted fluid in the respiratory organs, for which there are numerous other etiologies) and laryngeal spasm (discussed below).

Although there are few clinical accounts of direct toxicity from these gases in a volcanic context, an indication of likely health effects is provided by fumes from Hawaii's fumaroles and basaltic lava, which contain high concentrations of both HCl, and, to a lesser degree, HF (Murata 1966; Kullmann et al. 1994). HCl and HF concentrations are often highest in dense plumes arising close to the ocean, from which sea winds carry mist clouds to the adjacent village of the Hawaii Volcanoes National Park. The HCl-/HF-acidic aerosol may exacerbate pre-existing lung disease (Ostro et al. 1991), even at great distances from volcanic vents (Mannino et al. 1996). On the island of Ambrym, Vanuatu, cutaneous burns resulted from acid rain following a 1979 emission of eruptive gases high in HCl and sulfur compounds. Gastric upsets were also reported from the ingestion of acidic cistern water (Scientific Event Alert Network 1989).

At a pathophysiological level, irritant gases affect the respiratory tract in a number of ways. The relatively more soluble gases, which include HF and HCl, tend to be removed by mucus linings before they reach the alveoli. Therefore, such gases predominantly cause inflammation in the upper airways, which results in symptoms such as cough and reactive bronchoconstriction. These effects, although often relatively short-lived, may be severe, particularly in people with hyperreactive airways (such as asthma sufferers). Less soluble volcanic gases, including H_2S and mercury vapor (both discussed below), are less likely to be 'cleared' by the mucus secretions of the upper airways. Thus, there is greater risk of penetration to the level of the smaller airways and alveoli.

As the concentration of respired irritant gases rises, there is a greater chance that alveolar damage will result. The most vulnerable targets are the epithelial cells lining the airways, the mucosal tissues underlying these cells, and endothelial cells lining the pulmonary blood vessels (including the capillaries of the alveoli). Damage to these cells results in an increase in vessel permeability. Ultimately, the alveoli fill with protein-rich fluid from plasma, which interrupts the process of effective gas exchange. In addition, there may be denudation and sloughing of alveolar epithelium and the mucous membranes of bronchiolar walls.

During the acute phase, the spectrum of disease caused by irritant gases is highly variable and includes pneumonitis, which results in a dry cough, shortness of breath, and evidence of patches of inflammation on chest radiography. The condition may often resolve without long-term sequelae such as pulmonary edema. In other cases, pulmonary edema emerges within a day of exposure, and may be associated with the formation of mucus plugs and collapse of areas of the lung. Because of compromised oxygen intake, patients may experience breathlessness and acute (or adult) respiratory distress syndrome (ARDS). This syndrome is associated with severe alveolar injury, edema, and hemorrhage. It has a mortality rate of around 50%, and represents the end point for a number of disease processes apart from toxic inhalation.

If the victim survives the acute stage of gas exposure, a number of further clinical outcomes may occur. Often the symptoms may rapidly reverse once the gas is removed and full recovery occurs. If the gas has produced severe

epithelial damage, however, cellular regeneration may take weeks. A secondary effect of this alveolar disruption is that the lung tissues become more vulnerable to bacterial invasion, and thus chest infections may intervene in the post-exposure period. Occasionally, toxic gas exposure results in permanent fibrosis (scarring) of lung tissue. Another potential outcome from toxic inhalation (described in greater detail below) is prolonged airway hyperreactivity, although the specific cause of this disorder remains uncertain.

Sulfur dioxide, SO_2, is well-established as a cause of acute and chronic disease. Both the gas, and the sulfuric acid aerosols into which it forms, are highly irritant, particularly to the eyes, nasal passages, throat, and respiratory tract. High SO_2 concentrations may also act to cause asphyxia, although in volcanic contexts the effects appear primarily irritative.

Sulfur dioxide exposure may provoke exacerbations of asthma, even at low concentrations. Increased airway resistance has been noted in asthmatics exposed to SO_2 at concentrations of 0.5 ppm when exercising, and at 1 ppm during rest. In non-asthmatic controls, an increase in airway resistance has been provoked at 5 ppm (Bethel et al. 1983). Cough and eye irritation occurs at concentrations of 20 ppm or above. At higher doses, pneumonitis, pulmonary edema, ARDS, and ultimately death may result.

In the context of volcanically active sites, SO_2 concentrations of 1 ppm have been recorded far downwind (30 km) from Masaya Volcano in Nicaragua (Baxter et al. 1982a). The effects of long-term emissions have been examined at Kilauea, Hawaii, which has been erupting for 29 years and (in 2012) continues to release around 400 t of SO_2 per day. Episodes of increased SO_2 in the ambient air exceeded health standards 80 times in 15 years and appeared to correlate to ongoing eye irritations, throat pain, and respiratory problems, including asthma exacerbations (Elias and Sutton 2001). During elevated volcanic emissions in 2008, increased morbidity of medically-diagnosed acute illnesses (cough, headache, acute pharyngitis, and acute airway problems) were observed in age and ethnic subgroups of the general Hawaiian population (Longo et al. 2010).

At the histopathological level, SO_2 has a diverse range of respiratory effects. Although larger aerosols are filtered out by the nose and nasopharynx and are rapidly removed by mucociliary clearance, smaller SO_2 particles (particularly those less than 10 μm in diameter) may be deposited in deeper airways. Sulfur dioxide also acts to increase levels of mucus secretion and viscosity. Animal studies also indicate that the gas impairs the ability of macrophages to destroy bacteria in the alveoli, thus increasing the risk of respiratory infection.

Apart from affecting mucus production, SO_2 may produce airway narrowing through other mechanisms. As mentioned, people with high levels of airway hyperresponsiveness, such as those with asthma and/or atopy (tendencies to allergic responses), are particularly vulnerable. Sulfur dioxide may provoke the recruitment of inflammatory cells, including histamine-secreting mast cells and eosinophils, which persist long after exposure ceases. Such cells contribute to the risk of recurrent airway constriction, particularly if provocation with gases (or other agents and allergens) recurs. This bronchoconstrictive response may enter a chronic phase (sometimes after only one exposure of a particular irritant gas). Other irritant gases and aerosols, such as nitrogen dioxide (discussed below), may precipitate a similar response. Other changes from SO_2 exposure include chronic neutrophil infiltration and edema in the mucosa, which is more distinctive of chronic bronchitis.

In addition to acting directly, irritant gases may create health problems by their interaction with other atmospheric processes. For example, droplets of rainfall or mud may pass through the irritant gases present in toxic clouds or ash plumes, thus creating acid rain. The health effects of this process were experienced by people on Kodiak Island, who suffered from falls of both ash and sulfuric acid rain following the Katmai eruption on the Alaskan mainland in 1912. A number of Kodiak islanders, 160 km downwind from the active vent, suffered from "stinging burns" when this rain contacted their lips or skin (Blong 1984). Acid rains in other sites have also resulted in eye irritation (e.g., with Masaya in Nicaragua, Baxter et al. (1982a)) and apparently even hair loss (during the 1917 eruption of Boqueron). A secondary health risk is posed by the interaction between acid rains and heavy metals: acid rain has reacted with zinc in galvanized roofs and tainted water supplies as a result (Baxter et al. 1982b).

As with tephra fallout, chemicals dispersed with volcanic gas emissions may affect water and food supplies. During the 1783 eruption of Laki in Iceland, apart from the high levels of expelled fluorine, an estimated 150 megatons of SO_2 was discharged. This vast quantity of gas, and the aerosolized sulfuric acid it formed, had destructive consequences for vast tracts of surrounding pasture lands. The event tipped the balance in the already marginal farming environment of Iceland. The ensuing period of crop damage resulted in massive livestock losses: half of all the island's cattle and horses and four-fifths of the sheep perished. One-fifth of Iceland's population, around 10,000 people, died as a consequence of famine (called the haze famine from the persistent presence of sulfur compounds in the atmosphere) with toxic fluorosis probably also contributing to some fatalities. The sulfurous volcanic haze from Laki also stretched over Western Europe and has been associated with increased mortality in the UK in the late-summer of 1783 (Witham and Oppenheimer 2005).

10.3.4 Noxious Asphyxiants

The pungent gas hydrogen sulfide, H_2S, is a noxious asphyxiant: that is, it acts as both an asphyxiant and a powerful irritant. Its metabolic effect is to inhibit cytochrome oxidase, one of the enzymatic drivers of cellular metabolism (Jappinen et al. 1990).

In volcanic environments, it is particularly released from cooling magmatic systems i.e. areas displaying geothermal/hot spring activity at the surface, such as Yellowstone, USA or Rotorua, New Zealand. At low concentrations, H_2S may cause irritation of the conjunctivae and mucous membranes. Short exposures at concentrations of 2 ppm have not shown any statistically significant effects on respiratory function (Jappinen et al. 1990). However, an analysis of the effects of chronic low-level exposure to H_2S in the geothermal centre of Rotorua, New Zealand, revealed a positive association with diseases of the nervous system and respiratory system (Bates et al. 2002). Once the H_2S concentration in ambient air increases, the effects become more definitive, and exposure even for a few seconds may be fatal. For example, at the Kusatsu-Shirane crater in Japan, a high school teacher and two students were overcome and killed by H_2S emitted from nearby fumaroles that had concentrated in a bowl-shaped depression (Scientific Event Alert Network 1989). Early signs of poisoning include headaches, ocular and respiratory irritation, and loss of smell (anosmia). Apart from these effects, inhalation of the gas also directly damages the respiratory tract and precipitates pulmonary edema in the lungs. At 1,000 ppm, fainting occurs. Ultimately, H_2S causes cessation of breathing by direct action on the respiratory centers of the brain. Those who recover from poisoning may suffer from long-term neuropsychological effects. These sequelae are probably related to acute hypoxic effects on the brain either during poisoning or subsequently from pulmonary damage.

10.3.5 Health Risks from Other Volcanic Gases

Information relating to the effects of exposures to other volcanic gases is limited. After symptomatic (or fatal) episodes of inhalation, it may be difficult to determine the nature of the gas or combinations of gases that were present, and whether the concentrations were high enough to cause illness. Other gases produced during eruptive or degassing events that may *potentially* reach concentrations sufficient to produce symptoms in humans include: CO; nitrogen dioxide, NO_2; carbon disulfide, CS_2; methane, CH_4; and ammonia, NH_4.

Carbon monoxide—Like H_2S, CO is an example of a noxious asphyxiant. Although concentrations in plumes are usually low, the gas may reach high concentrations in certain eruptions and near fumaroles (Tazieff and Sabroux 1983; Williams and Moore 1983). CO is toxic to humans in small doses. Once inhaled and absorbed, it rapidly permeates across the membrane of red blood cells and binds to the iron component of hemoglobin. Its affinity for hemoglobin is profound, over 200 times greater than oxygen, and it rapidly reduces the oxygen-carrying capacity of blood. Symptomatically, low levels of blood saturation with CO produce headache, nausea, and impaired judgment and dexterity. As the percentage saturation exceeds 30–40%, severe nausea, vomiting, confusion, hyperventilation, and collapse ensue. Coma and death occur if treatment is not provided rapidly. The characteristic 'cherry red' skin associated with CO poisoning occurs because of the persistent saturation of both arterial and venous blood with this 'hemoglobin-loving' gas (which forms carboxyhemoglobin). Survivors of CO poisoning may suffer from permanent neurological and behavioral symptoms, which include disorientation, mood changes, and movement disorders.

Nitrogen dioxide—Significant concentrations of NO_2 were detected in plumes from the 1980 Mount St. Helens eruption (Olsen and Fruchter 1986). If high concentrations are inhaled, hydrolysis of NO_2 results in the production of nitric acid, particularly in the lower respiratory tract (including the alveoli). Nitric acid in turn produces nitrates and nitrites, both of which are capable of producing cytotoxicity and disruption of cellular membranes. The clinical outcomes of NO_2 inhalation are in some respects comparable to those of contact with SO_2. Acutely, victims may suffer from chemical pneumonitis, pulmonary edema, ARDS, or fatal lung injury. Susceptible individuals may suffer from exacerbations of asthma, increased need for bronchodilators, and decreased pulmonary function (Moseholm et al. 1993). As with SO_2, emergency room visits and hospital admissions for asthma are also increased by exposure to NO_2.

Carbon disulfide—CS_2 has been detected in volcanic plumes, although concentrations are usually low. This gas acts as a powerful toxicant and produces headaches, muscular weakness, and delirium.

Methane—Methane, together with CO and CO_2, was among the gases which hampered rescue efforts during the 1973 eruption on the island of Heimaey, Iceland (Williams and Moore 1983). It may also be concentrated around fumaroles. Primarily, it is an inert asphyxiant and acts by replacing oxygen in breathed air.

Ammonia—Exposure to this highly water-soluble gas results in severe irritation to the eyes, skin, and upper airways. Injuries are produced by thermal and chemical (alkali) burns as ammonia reacts to form hydroxyl ions on exposed mucosal surfaces.

10.4 Volcanic Dispersal of Metals and Trace Elements

The dispersal of toxic agents such as fluoride and selenium have been discussed above. In addition, there may be widespread dispersal of other elements and metals with identified risks for humans. For example, using a plume dispersal model for a 1976 eruption of Mount Etna, a variety of toxic aerosols was estimated to have exceeded permissible air concentrations at a distance of 10 km from the vent. These included lead (Pb), mercury (Hg), copper (Cu), zinc (Zn), selenium (Se), and cadmium (Cd). Arsenic (As) exceeded recommended concentrations at a distance of 5 km (Faivre-Pierret and Le Guern 1983). Daily aerosol outputs were also estimated, which included 360 kg Pb, 110 kg As, 75 kg Hg, and 28 kg Cd per day. These last four elements, which have well-established toxic effects even in small doses, are discussed briefly below as well as in greater detail in other chapters in this book.

Metal vapors, such as those produced by mercury, may act directly as an irritant gas. It has been suggested that mercury is sometimes present in concentrations sufficient to be hazardous if emissions were inhaled over a prolonged period of time (Baxter 1983). Mercury vapors act to cause pulmonary edema and bronchitis and are neurotoxic once absorbed.

Arsenic can be deposited in soil or water (including seepage into subterranean wells) following volcanic or low-level geothermal activity (Welch et al. 1988). Chronic arsenic poisoning may affect many organ systems. For example, in affected populations in Bangladesh, India and Taiwan, ingestion of drinking water high in arsenic has been associated with a variety of skin lesions, which include increased or decreased pigmentation and keratosis. Gangrene has also been reported (blackfoot disease). The carcinogenic role of arsenic has been suggested in a range of studies: exposure increases the risk of cancer of the skin, lung, liver, bladder, and kidneys (see also Chap. 12, this volume).

Lead produced by volcanic processes may potentially be inhaled or ingested in contaminated water and foodstuffs. The health effects, particularly on the neurological development of children, have been well-documented.

Like lead, cadmium enters the human body through inhalation or ingestion. Cadmium deposited in the soil crosses readily into plants and thus the food chain. Having entered the body, this toxic metal tends to be retained in the liver and kidneys and is excreted extremely slowly. The effective half-life of cadmium in humans may exceed 30 years. Cadmium thus has the potential to result in prolonged illness, and adverse effects may occur in virtually any organ. The more common manifestations of acute ingestion or inhalation include vomiting, cramps, respiratory difficulties, and ultimately loss of consciousness. Chronic effects include anemia and renal disorders. More controversial is the connection between elevated cadmium and hypertension.

It is important to note that the valency state (species) of these metals and metalloids can be a key determinant of their potential biological activity. The issue is discussed in more detail in other chapters.

10.5 Volcanism and Radiation Hazards

Radioactive decay of volcanic material may also have consequences for human health. Ash may have a high uranium content and carry adherent particles of radon, an alpha-radioactive gas that has been linked to the development of lung cancer (Baubron and Toutain 1990). Exposure may also occur from use of volcanic material, such as for building. Radon is discussed in detail in *Radon in Air and Water* (Chap. 11), this volume.

10.6 Toxic Exposure with other Eruptive Events

10.6.1 Lava Flows

One of the more visually dramatic outcomes of volcanism is the ejection of low viscosity material, such as basaltic lava (Fig. 10.6). In some locations (for example, Hawaii), eruptions may be associated with fountaining of molten material, in which globules of plastic lava are sprayed over a kilometer high. These may feed into lava lakes and lava flows or tubes that course away from the volcano. The direct threats to health posed by lava flows are primarily thermal

Fig. 10.6 Lava from the Krafla fissure, Iceland (Photo: Olle Selinus)

injuries. Often fatalities occur because of unexpectedly rapid flows, because escape routes have been cut off, or from steam explosions created when the lava strikes a water source.

Lava flows may result in illness less directly by exposing humans to toxic chemicals. For example, the basaltic lava flows in Hawaii are often associated with the release of SO_2 and aerosolized droplets of sulfuric acid. As discussed, HCl and (a lesser degree) HF may also be formed, particularly when molten lava strikes the ocean, thus creating falls of acid rain from the steam plume (Mannino et al. 1996). Lava may also act to taint subterranean wells by the process of leaching. The toxicity and health effects of compounds released from lava flows do not differ from those already discussed, although the exposure may at times be intense, such as in the area around an erupting crater. Emergency crews often work in close proximity to lava flows. On Heimaey in 1973, for example, crews spent many days on or near lava flows, applying cooling waters in a successful attempt to solidify and direct flows away from the main town.

Fig. 10.7 Crater lake formation adjacent to a volcanic vent (Photo: Michael Durand)

10.6.2 Pyroclastic Density Currents

Pyroclastic density currents (also known as pyroclastic flows) are intensely hot flows (up to 1,000°C) of blocks, ash and gas which may travel at speeds of up to 700 km/h down the volcanic flanks. They occur from collapse of buoyant eruption columns during explosive events or during the collapse of viscous lava domes which over-grow the crater. The exact composition and temperature varies greatly but, during fragmentation of the magma/lava, respirable particulate will be generated. The finer ash is lofted from the flow, leaving a fines-depleted deposit and fines-enriched airborne plume which will deposit on the surrounding environment. The hazards of such ash were discussed in Sect. 10.2. Such flows present an immediate risk to humans close to the vent. With their considerable kinetic energy, the fatality rate of those caught in such flows is usually extremely high, and the common causes of death include asphyxiation (often from burial), trauma, and severe burns (especially of the respiratory system). For example, during the 1902 eruption of Mount Pelée on the island of Martinique, such a superheated cloud rapidly enveloped the city of St. Pierre, resulting in around 30,000 fatalities.

Pyroclastic density currents result in varying degrees of thermal injury to the skin ranging from superficial erythema, to deep penetration into the subcutaneous tissues, to the extreme of complete incineration. Victims are commonly described as appearing dried and "mummified" rather than charred (the outcome usually observed with fire injuries), with a 'pugilistic' attitude caused by contraction of muscles. Respiratory effects appear to occur as a result of intense heat, oxygen deficiency, ash inhalation, and toxicity of the gas. Asphyxia from plugs of ash in the upper airways was described as the cause of death in those caught in the flow of the Mount St. Helens eruption. Survivors of the devastating flows from Lamington, Papua New Guinea, in 1953, have suffered from symptoms suggestive of pharyngeal burns, including throat pain, shortness of breath, and inability to swallow (Taylor 1958). Health effects subsequent to the acute injury include pneumonia, tracheobronchitis, and ARDS, presumably from irritation and secondary infection of injured respiratory tissues (Eisele et al. 1981).

Flows, surges, and debris avalanches also impact human health because of their capacity to disperse toxic compounds. For example, during the Mount Pinatubo eruption in 1991, pyroclastic density currents and lahars (mud flows) contributed to the volcanic material that covered a wide area and filled surrounding valleys. In later years, erosion, often triggered from monsoon rains, then acted to mobilize volcanic chemicals.

10.6.3 Volcanic Activity and Aquatic Environments

Apart from the interaction between tephra and the hydrosphere described above, volcanic and aquatic processes may intersect in other ways (Fig. 10.7). Crater lakes, for example, can act as a reservoir of toxic compounds which, in some circumstances, may affect human populations. Some lakes, such as the Poas volcano, Costa Rica, sit atop a degassing system. At this site, steady emissions of SO_2 pass through the shallow lake, which is often hot and intensely acidic (pH <1). These emissions, together with particles of rock dissolved in the acid lake water, are

periodically dispersed out of the crater and have been linked to respiratory problems in downwind communities (Baxter 1997). At Poas and Kawah Ijen in Java, crater lakes have contaminated water supplies with fluoride and other elements (Baxter 2001).

A fast-moving and potentially lethal consequence of volcanic eruptions is the lahar. These torrential flows of mud, water, and debris wash down the sides of the volcano, and may occur in association with crater lake rupture, the melting of snow and ice during or after eruptive events, or from heavy rainfall. For communities situated in the path of lahars, the opportunities for timely warnings may be limited—sometimes with lethal consequences. Lahars from some volcanic lakes, such as Kelut in Indonesia, may be hot and often acidic. Five thousand people died from a lahar generated from the 1919 eruption of Kelut. Those caught in the flow often suffer from drowning, suffocation while entrapped, or severe trauma from penetrating wounds and fractures (Baxter 1990). Following the Nevado del Ruiz lahar in Colombia in 1985, when around 25,000 people were buried in the town of Armero, burns were also noted and may have been acidic in origin (Lowe 1986).

For submarine eruptions, rafts of pumice and areas of discolored water on the surface may be the only evidence that volcanic activity has occurred. At the other end of the hazard continuum, such events may produce tsunamis, which pose a major threat to coastal populations even at considerable distances from the vent. Submarine activity may manifest itself in other ways, as illustrated by the destruction of a Japanese research vessel *No. 5 Kaiyo-Maru* in 1952. The vessel had been traveling to monitor a submarine eruption 420 km south of Tokyo when radio contact with the vessel was lost. Examination of the debris (which included boat fragments containing embedded rock particles) suggested that an explosion beneath the ship had caused the disaster, in which all 31 crew members were killed.

Apart from such infrequent events, little is known about the threat to health posed by submarine emissions of gases and particles. It is probable that most ejected material is thoroughly diluted in seawater and thus poses little risk for humans. However, in some regions, activity may occur near coastal populations. In the Antilles, for example, recently active submarine sites include Kick 'Em Jenny', situated 8 km north of Grenada, and other vents near Martinique, where periodic episodes of "boiling water" have been reported (Roobol and Smith 1989). Along the shallow Reyjkanes oceanic ridge near Iceland, bubbles rich in methane and CO_2 may rise to the sea surface (German et al. 1994). These eruptions and emissions, although under water are, in effect, near vent phenomena. The potential for toxicity remains, albeit tempered by the aquatic environment.

10.7 Volcanic Monitoring and Health Protection

HEKLA, perpetuis damnata estib. et ni: uib. horrendo boatu lapides evomit (Hekla, perpetually condemned to storms and snow, vomits stones under a terrible sound)

—an early surveillance report of the volcano Hekla (from 1585 map of Iceland)

10.7.1 Introduction

From the perspective of health protection, the purposes of volcanic monitoring are fourfold: (1) to provide an early warning system of potential health hazards, thus providing a preparation period for resident populations, health services, transport services, electricity utilities, etc., that are located in the hazard zone; (2) to minimize illness from contaminated water supplies, tainted or fouled foodstuffs, and air pollutants, particularly for susceptible groups; (3) to provide accurate records of adverse health events for use in epidemiological and clinical research and; (4) to provide detailed mineralogical and toxicological information early in the crisis to inform hazard mitigation and both the need for medical studies and the purpose of those studies if they are commissioned.

The phases of volcanic monitoring may be usefully classified as follows:

- Pre-eruptive phase: This period extends from the elevation of alert levels above the baseline state, hence indicating increased likelihood of volcanic activity, to the actual start of the eruption.
- Eruptive phase: The period surrounding the eruptive event (or degassing episode), including event imminence, the primary volcanic event, and periods of ongoing volcanic activity.
- Post-eruptive phase: The period over which volcanic activity wanes, allowing recovery and rehabilitation measures to take priority.

The current discussion will be based on the assumption that these phases provide the loci of intervention designed to protect the public health.

10.7.2 The Pre-eruptive Phase

This phase is primarily directed at preparing at-risk populations for an impending volcanic event. Early volcanic activity is usually monitored by geologists and involves assessments of seismic activity, ground deformation, gas emissions, geophysical variables and hydrology (Fig. 10.8). Using hazard maps, the primary aim during the pre-eruptive

Fig. 10.8 Monitoring volcanic emissions (Photo: Michael Durand)

stage is to predict the temporal and spatial pattern of an eruption, together with the scale and path of any flows. Disaster scenarios will usually include the need for, and possible extent of, evacuation from the danger zone. Secure locations and safe travel routes for displaced populations should also be considered.

An important strategy during the pre-eruptive phase is the dissemination of information regarding potential health effects and strategies for minimizing exposure. Ideally, this process of training and educating should be continuous (even in periods of quiescence), and only escalated in the event of increased alert levels. (In Japan, for example, evacuation drills are conducted routinely in areas at risk.) Baseline tests of water and air quality should be conducted. The International Volcanic Health Hazard Network (IVHHN) has produced pamphlets for the public on volcanic health hazards and how to prepare for ashfalls which are translated into local languages, adapted for cultural factors, locally printed and mass-distributed prior to, and during eruptions (www.ivhhn.org).

From an epidemiological perspective, this period may be an opportunity for selection of groups for prospective studies (for example, pre-eruptive baseline screening of people with asthma who will be followed through and subsequent to the eruption and ash-fall).

10.7.3 The Eruptive Phase

Early warning systems should be established for affected communities in order to ensure that the public, and organizations designed to preserve the societal infrastructure, are prepared for the arrival of the eruptive phase. For example, hospital staff and general practitioners should be informed of possible consequences of tephra fallout, including an increased frequency of motor vehicle accidents, ocular problems, and exacerbations of asthma.

Where risks are minimal (such as from light ash-fall), it may not always be necessary to recommend major changes in daily routines. Some groups of susceptible individuals, such as those with pre-existing respiratory disease, may benefit from consultation with health professionals. Instruction of disease management in the event of ash-fall or increased levels of gaseous emissions should be initiated if appropriate. For some patients, such as those with asthma who may react to raised levels of sulfuric acid aerosol, medical advice may be as simple as recommending that bronchodilator medications (such as inhalers) are always readily at hand. In those with chronic illness, strategies for ongoing medical management are particularly important should evacuation become necessary. More general advice may be required each day, including avoidance of areas of likely ash-fall. Ongoing status reports of hazards and strategies to avoid risk should be provided using newspapers, radio and television broadcasts, internet and online social media and public notices. In at-risk areas, it is imperative that such communication plans *pre-exist* the volcanic event to avoid the need to generate the information on an ad hoc basis.

Table 10.2 summarizes health impacts of eruptive events and relevant health preservation strategies. The hazards are separated according to proximity to the eruptive site using the general categories near vent and distal from vent.

Given the range of health consequences from exposure to volcanic ash, monitoring of ash fall is an important component of eruption-phase management. Ideally, exposure will be systematically measured using PM_{10} real-time aerosol monitoring instrumentation but often these are not available. In recent years, the IVHHN has developed a protocol for rapid mineralogical and toxicological analyses of ash samples, that has been applied during the eruptions of Rabaul, Papua New Guinea (2007–2008), Chaitén, Chile (2008), Kilauea, Hawaii (2008), Eyjafjallajökull, Iceland (2010), Merapi, Indonesia (2010), Grimsvötn, Iceland (2011) and Puyehue Cordón Caulle, Chile (2011) (Horwell et al. 2010; Le Blond et al. 2010; Damby et al. 2013). As a first step, grain-size is measured to determine the amount of health-pertinent material in the ash. Data are then disseminated directly to hazard managers to allow critical decisions on hazard mitigation such as distribution of dust masks. Following this initial stage, the crystalline silica content (for silicic eruptions), bulk composition, particle morphology, surface area and leachable elements are measured. For those samples displaying characteristics which could indicate a health hazard, further surface reactivity and *in vitro* toxicological analyses are undertaken. The size and composition of ash particles vary among volcanoes and from eruption to eruption; this makes it difficult to

Table 10.2 Comparison of proximal (near vent) and distal health impacts and corresponding health preservation strategies

Eruptive event	Consequence	Health impact	Health prevention strategies
Near vent			
Explosion/dome collapse	Blast, 'bombs', shock waves	Trauma, skin burns, lacerations	Evacuation, movement to secure shelters
	Lightning	Electrocution	Evacuation
	Forest and bush fires; combustion of buildings and vehicles	Burns; smoke inhalation	Evacuation
	Pyroclastic density currents	Skin and lung burns; asphyxiation	Wear high efficiency masks; minimise exposure; protect homes; offices from ash infiltration
	Ash-fall	Asthma, exacerbation of pre-existing lung disease; silicosis if crystalline silica content high and exposure prolonged	
Sector collapse	Landslides/debris avalanche	Burial	Evacuation
	Tsunami	Drowning	
Drainage of crater lakes; melting ice, snow, or rain accompanying eruption	Lahars, floods	Engulfing, drowning	Evacuation, diversion barriers
Lava flow	Building/forest/bush fires	Engulfing and burns (rare)	Evacuation, diversion barriers
Gas emissions H_2O, SO_2, CO, CO_2, H_2S, HF	Pooling in low lying areas	Asphyxiation, airways constriction (exacerbation of asthma, COPD)	Evacuation, avoidance
			Respiratory protective equipment
Radon	Radiation exposure	Lung cancer	Evacuation
Distal from vent			
Ash-fall	Ash <10 μm in diameter	Asthma, exacerbation of pre-existing lung disease; silicosis if crystalline silica content high and exposure prolonged; conjunctivitis and corneal abrasions	Wear high efficiency masks; minimise exposure; protect homes; offices from ash infiltration; goggles/eye glasses and avoid contact lenses
	Water contamination from fluoride, possibly also heavy metals (e.g. cobalt, arsenic)	Gastrointestinal upset and electrolyte disturbance	Avoid water that has not yet been approved, avoid surface water, use water from wells
	Food contamination	As above	
Gas emissions	Acid rain	Eye and skin irritation; possible toxic contamination	Protection during rainfall, avoid collection of rain water for drinking, especially from metal roofs, etc.
	Dispersal of irritant and/or asphyxiant gases	Exacerbation of pre-existing lung disease (especially asthma); mucosal and conjunctival irritation	Prophylactic use of medication; avoidance/ evacuation

Adapted from Baxter et al. (1986)

predict the chemical composition of any given volcanic product, although the growing database accumulated by IVHHN is allowing generalizations based on eruption type (e.g. silicic dome-collapse eruptions tend to generate very fine-grained, crystalline silica-rich ash). Although the health effects of tephra fallout are often not immediately life-threatening for populations distant from the eruption site, acute anxiety is often felt and populations need rapid assurance that ash in the environment is unlikely to result in severe illness, and this is possible by carrying out such rapid studies. Prior warnings of ash-fall, together with regular announcements on air quality, will enable susceptible individuals to minimize their exposure or to seek medical advice where appropriate.

As described above, tephra may also have disruptive effects on water supplies. Following some eruptions there is a direct risk of toxicity, which includes acute fluorosis. Water quality may also be rapidly degraded by increased turbidity and fluctuations in pH levels. Apart from direct testing and pH monitoring, it may often be possible to anticipate potential effects on water quality by estimating volumes of ash falling in water catchment areas and by using results from rapidly performed leachate studies. Such analyses are of importance not only to populations drawing

Table 10.3 Strategies for monitoring tephra dispersal: a health protection perspective

Monitored component of tephra fallout	Location of monitoring	Optimal time for monitoring	Health-related objectives of monitoring
Mapping of ash-fall (isopach)	Throughout ash-fall area	As soon as possible after ash-fall in order to determine load of ash in water catchment areas, pastures etc	(a) To maintain water quality, including monitoring of turbidity and pH
			(b) To minimise the risk of ingesting toxic substances in drinking water (e.g., fluoride)
Airborne PM_{10} (measure of particulate matter <10 μm) accounting for relative proportions from volcanogenic and other sources (e.g., man-made pollutants)	(a) Priority to plume-vulnerable areas with greatest population density	(a) Continuously and preferably at locations were humans are likely to receive maximum exposure	(a) To monitor daily levels and provide information to individuals with respiratory disease (e.g., to stay indoors on certain days); also for research purposes in linking the respirable portion of ash to health effects.
	(b) Personal monitors on individuals exposed to much ash	(b) Whenever exposed to ash	(b) To monitor personal exposure and any associated health effects. This is especially important if the ash has a high silica content.
Proportion bulk ash which is 'thoracic' (<10 μm)	Priority to plume-vulnerable areas with greatest population density	Immediately after ash-fall	To quantify the level of health-pertinent ash to assess the risk of acute and chronic illness (e.g., silicosis)
Crystalline silica content	(a) Priority to plume-vulnerable areas with greatest population density	(a) Immediately after ashfall	(a) To assess, and hopefully, alleviate public anxiety concerning silicosis
	(b) Individuals that will be exposed to high levels e.g. clean up workers	(b) More detailed studies at later date	(b) To determine the relative risk of silicosis for individuals exposed to varying concentrations and types of silica
Bulk composition	Areas with ash-fall	Immediately after ash-fall	Determines magma type (hence likelihood of crystalline silica and iron availability)
Particle morphology	Areas with ash-fall	Immediately after ash-fall	To check for the presence of fibre-like mineral particles
Surface area	Areas with ash-fall	Immediately after ash-fall	Surface area is a proxy for toxicity
Leachates	(a) In water catchment areas supplying public drinking water supplies	(a) Immediately after ash-fall	(a) To assess risk to water quality in the catchment region, and to initiate clearance of toxic agents and particles; this is often required urgently because water demand is likely to be high during the process of clean-up
	(b) In water catchment areas of private water supplies (e.g., irrigation; wells, water tanks)	(b) Immediately after ash-fall if practicable	(b) To assess potential effects on water quality
			(c) To determine leachable elements in the RESP or GIT (see Table 10.1)

Table adapted from IVHHN protocol supplied by Le Blond et al. (2010)

water from large catchment areas, but also to those using non-public water supplies (such as for crop irrigation) that are not routinely monitored for quality.

Other health hazards of ash include the mechanical effects (particularly of wet ash), which cause roofs to collapse resulting in trauma. Heavy ash-falls may also result in traffic accidents, a product of both poor visibility and ash coating on roads (Dent et al. 1995). Cleanup crews are particularly at risk, and falls from roof cleaners often cause greater morbidity than any direct effect of the eruption.

Table 10.3 summarizes the strategies for tephra monitoring, which indicates how such information may assist in health protection for vulnerable populations.

As discussed in the previous section, volcanic gas emissions—and the aerosols that are derived from them—also pose a major threat to health. Therefore, as with tephra, gas production at the vent should be monitored in order to alert downwind communities of the potential hazards on a daily basis Table 10.4. Information of gas (and tephra) dispersal production will usually be combined with weather

Table 10.4 Strategies for monitoring gas (and aerosol) emissions: a health protection perspective

Monitored component of gas emission	Location of monitoring	Optimal time for monitoring	Health-related objectives of monitoring
CO_2	(a) At vent/remote sensing	(a) At regular intervals preferably daily	(a) To minimise risk of asphyxiation from down-hill flow of CO_2 into populated areas
	(b) In areas where CO_2 (heavier than air) can pool.	(b) If the areas need to be accessed	(b) To minimise risk of asphyxiation from CO_2 collecting in low-lying areas
	(c) Soil degassing	(c) Continuously	(c) Same as for (a)
SO_2	(a) At vent/remote sensing	(a) At regular intervals preferably daily	(a) To provide early warnings to vulnerable populations
	(b) Priority to emission-vulnerable areas with greatest population density	(b) Continuously during eruption or in response to early warning	(b) To minimise risks of respiratory disease, including exacerbations of asthma
H_2S	(a) At vent	(a) At regular intervals preferably daily	(a) To provide early warnings to vulnerable populations
	(b) Downwind in built up areas	(b) Continuously during eruption/ passive degassing or in response to early warning	(b) To minimise risks of respiratory disease, including exacerbations of asthma
	(c) In areas where H_2S (heavier than air) can pool	(c) If the areas need to be accessed	(c) To minimise risk of noxious asphyxiation
HCl and HF	(a) At vent/remote sensing	(a) At regular intervals preferably daily	(a) To provide early warnings to vulnerable populations
	(b) Priority to emission-vulnerable areas with greatest population density	(b) Continuously during eruption or in response to early warning	(b) To minimise risks of respiratory disease, including toxic lung injury
	(c) Areas where much water is collected from roofs	(c) Continuously during eruption or in response to early warning	(c) To minimise effects of acid rain dissolution of metal roof components (which may release heavy metals into drinking water)
Radon	At vent	Continuously	To minimise possible long-term sequelae of radon exposure, such as lung cancer

predictions. Combining climatic and volcanic forecasts—including predictions for acid rain or combined ash and fog contributing to poor visibility—will allow communities to implement appropriate precautions on a day-by-day (or hour-by-hour) basis.

The monitoring of gases and pH often requires specialized equipment that may not be available for permanent use in all settlements at risk of exposure. A possible solution may be the provision of a mobile air quality team that responds to predictions based on geologic and climatic data.

10.7.4 The Post-eruptive Phase

As the post-eruptive phase commences, and recovery becomes possible, it is critical to continue monitoring for as long as toxic compounds remain present in the environment. Recovery from volcanic events is often a drawn-out process. Ongoing assistance is often required for long-term physical needs, and adverse impacts on psycho-social well-being can last for months or years (Cook et al. 2008). For example, as noted above, volcanic hazards may destroy croplands, poison livestock and contaminate available water supplies (Lohr et al. 2005; Hansell et al. 2006). The conclusion of the 'disaster phase' does not indicate the cessation of monitoring requirements, especially as volcanic degassing and lahars may continue for many years after a volcano becomes dormant. Collection of geological and health-related data should be ongoing for epidemiological reviews in order to anticipate which measurements are important for any subsequent eruptions. It may be necessary to specifically review exposure patterns in high-risk groups (such as cleanup workers) and in areas where dust will be disturbed (e.g., roads, schoolyards, city centers). In areas with persistent ash re-suspension, advice and hazard communication similar to that provided during the eruptive phase is important.

10.7.5 Utilizing and Integrating Data from Eruptions

In a coordinated response to volcanic hazards, ideally a number of information sources should be utilized. Geological monitoring may often help predict when and where an eruption will take place, as well as the type and scale of activity expected. After appropriate assessment of hazards,

the information should be disseminated to all relevant organizations, including territorial authorities or regional councils responsible for identifying risk areas and likely effects upon their constituent populations. Regional authorities are usually also responsible for ensuring that the public are, and remain, informed and educated throughout each of the eruption phases. Finally, local and international public health representatives and medical providers should participate in the assessment of the possible eruptive effects on human welfare.

Problems may arise, however, with the interpretation and integration of geological and medical information to ensure beneficial health outcomes. A number of failures to obtain and integrate health-related data have emerged during recent volcanic events, including the 1995–1996 Ruapehu eruptions in New Zealand (Davies 1998). For example, during the height of Ruapehu's activity, numerous measurements of ash-fall and gas emissions were undertaken, but these data were used almost exclusively for geological purposes. There were few avenues available to usefully relay such information or to discuss its implications with public health officials or medical practitioners. Furthermore, air monitoring of gases following eruptive activity was difficult, because many areas had no existing facilities in place to test pollutant levels during non-eruptive periods. Data regarding water contamination from falling ash was often unavailable because testing was completed at supply level, thus excluding many catchment areas that were vulnerable to ash-fall (such as small streams serving remote communities).

During disaster responses, there may be a degree of mismatching between the acquisition of eruption data and its health-related utilization. Given the diverse range of effects of volcanism on humans and ecosystems, full and accurate risk assessment requires integration of data collected by numerous ministries, departments, and regional authorities. Protocols for sharing information and achieving a consensus should be established *prior to* a volcanic event. For an effective response, joint consultation and action should involve organizations responsible for agriculture, fisheries, forestry, water provision, power generation, environmental management, and health. Furthermore, consensus needs to be achieved regarding the process of information transfer to the public, which includes the nature and frequency of health-related information, the organization that should impart the message, and the most appropriate media to use.

Geological data which may be generated before, during, and subsequent to eruptions therefore represent only the first step in the process of decreasing volcanogenic morbidity. The ideal end result would involve information collected from monitoring (e.g., gas production at the vent) being disseminated to all relevant regional organizations and health providers, and then to vulnerable communities. Recommendations communicated to the public should take into account practical realities (including economic or geographical constraints) and should be simple, coherent, non-contradictory, and delivered in a prompt fashion. From a medical perspective, geologic monitoring has a pivotal, but often underutilized, role in helping those living in the shadows of volcanoes. In keeping with the integrative aims of this book, the authors hope that this chapter will help to encourage the cross-disciplinary use of geological and medical data from locations where populations are exposed to volcanic eruptions. Such an approach offers an opportunity not only to advance scientific understanding, but ultimately to better protect public health.

See Also the Following Chapters. Chapter 2 (Natural Distribution and Abundance of Elements) • Chapter 11 (Radon in Air and Water) • Chapter 13 (Fluoride in Natural Waters) • Chapter 16 (Selenium Deficiency and Toxicity in the Environment) • Chapter 21 (Animals and Medical Geology) • Chapter 25 (Environmental Pathology)

References

Afane Z, Atchou G, Carteret P, Huchon GJ (1996) Respiratory symptoms and peak expiratory flow in survivors of the Nyos disaster. Chest 110(5):1278–1281

Armienta M, de la Cruz-Reyna S, Morton O (2001) Compositional variations of ash deposits as a function of distance at popocatepetl volcano. Cities on volcanoes In: 2nd conference proceedings, Auckland, New Zealand

Bates MN, Garrett N, Shoemack P (2002) Investigation of health effects of hydrogen sulfide from a geothermal source. Arch Environ Health 57(5):405–411

Baubron J-C, Toutain JP (1990) Diffuse volcanic emissions of carbon dioxide from volcano island, Italy. Nature 344:51–53

Baxter PJ (1983) Health hazards of volcanic eruptions. J R Coll Phys Lond 17(3):180–182

Baxter PJ (1990) Medical effects of volcanic eruptions. Bull Volcanol 52:532–544

Baxter PJ (1997) Volcanoes. In: Noji EK (ed) The public health consequences of disasters. Oxford University Press, Oxford

Baxter PJ (2001) Human health and volcanoes: recent developments. In: Stewart C (ed) Proceedings of the cities on volcanoes 2 conference, Auckland, New Zealand, 12–14 Feb 2001, Institute of Geological and Nuclear Sciences Information Series, vol 49. Institute of Geological and Nuclear Sciences Limited, Lower Hutt, New Zealand

Baxter PJ, Coutinho R (1999) Health hazards and disaster potential of ground gas emissions at Furnas volcano, Sao Miguel, Azores. J Volcanol Geotherm Res 92:95–106

Baxter PJ, Ing R, Falk H et al (1981) Mount St. Helens eruptions, May 18–Jun 12, 1980: an overview of the acute health impact. JAMA 246:2585–2589

Baxter PJ, Falk RS, Falk H et al (1982a) Medical aspects of volcanic disasters: an outline of hazards and emergency response measures. Disasters 6:268–276

Baxter PJ, Stroiber RE, Williams SN (1982b) Volcanic gases and health: Masaya volcano, Nicaragua. Lancet 2:150–151

Baxter PJ, Ing R, Falk H et al (1983) Mount St. Helens eruptions: the acute respiratory effects of volcanic ash in a north American community. Arch Environ Health 38:138–143

Baxter PJ, Bolyard ML, Buist AS et al (1986) Health effects of volcanoes. An approach to evaluating the health effects of an environmental hazard. Am J Public Health 76:1–90

Baxter PJ, Dupree R, Hards VL et al (1999) Cristobalite volcanic ash of the Soufriere hills volcano, Montserrat, British West Indies. Science 283:1142–1145

Bethel RA, Sheppard D, Nadel JA, Boushey HA (1983) Sulfur dioxide-induced bronchoconstriction in freely breathing, exercising, asthmatic subjects. Am Rev Respir Dis 128:987–990

Blong RJ (1984) Volcanic hazards: a sourcebook on the effects of eruptions. Academic, Sydney

Buist AS (1988) Evaluation of the short and long term effects of exposure to inhaled volcanic ash from Mt. St. Helens. In: Kagoshima international conference on volcanoes proceedings, Kagoshima, Japan, pp 709–712

Buist AS, Vollmer WM, Johnson LR et al (1986) A four-year prospective study of the respiratory effects of volcanic ash from Mount St. Helens. Am Rev Respir Dis 133:526–534

Cook A, Watson J, van Buynder P, Robsrtson A, Weinstein P (2008) Natural disasters and their long-term impacts on the health of communities. J Environ Monit 10:167–175

Cullen RT, Jones AD, Miller BG et al (2002) Toxicity of volcanic ash from Montserrat. IOM TM/02/01. Institute of Occupational Medicine, Edinburgh

Damby DE, Horwell CJ, Baxter PJ, Delmelle P, Donaldson K, Dunster C, Fubini B, Murphy F, Nattrass C, Sweeney S (2013) The respiratory health hazard of tephra from the 2010 Centennial eruption of Merapi with implications for occupational mining of deposits. Journal of Volcanology and Geothermal Research, doi: org/10.1016/j.jvolgeores.2012.09.001

Davies H (1998) Review of the priority 2 identification programme 1997/98. Institute of Environmental Science and Research Limited, Wellington

Dent AW, Barret P, de Saint Ours PJA (1995) The 1994 eruption of the Rabaul volcano, Papua New Guinea: injuries sustained and medical response. Med J Aust 163:635–639

Eisele JW, O'Halloran RL, Reay DT et al (1981) Deaths during the May 18, 1980, eruption of Mount St. Helens. N Engl J Med 305:931–936

Elias T, Sutton AJ (2001) Volcanic air pollution creates health concerns on the Island of Hawai'i, cities on volcanoes In: 2nd conference proceedings, Auckland, New Zealand

Faivre-Pierret R, Le Guern F (1983) Health risks linked with inhalation of volcanic gases and aerosols. In: Tazieff H, Sabroux J-C (eds) Forecasting volcanic events. Elsevier Press, Amsterdam

German CR, Briem J, Chin C et al (1994) Hydrothermal activity on the Reykjanes ridge: the Steinholl vent-field at 63°06′N. Earth Planet Sci Lett 121:647–654

Gessner B (1994) Acute fluoride poisoning from a public water system. N Engl J Med 330(2):95–99

Grandjean P (1982) Occupational fluorosis through 50 years: clinical and epidemiological experiences. Am J Ind Med 3:227–236

Green FHY et al (1982) Health implications of the Mount St. Helens eruption, laboratory investigation. Ann Occup Hyg 26(1–4): 921–933

Hansell AL, Horwell CJ, Oppenheimer C (2006) The health hazards of volcanoes and geothermal areas. Occup Environ Med 63(2):149–156

Hayakawa Y (1999) Catalog of volcanic eruptions during the past 2000 years in Japan. J Geogr 108(4):472–488

Hickling J, Weinstein P, Woodward A (1999) Acute health effects of the Mount Ruapehu (New Zealand) volcanic eruption of June 1996. Int J Environ Health Res 9:97–107

Hincks TK, Aspinall WP, Baxter PJ et al (2006) Long term exposure to respirable volcanic ash on Montserrat: a time series simulation. Bull Volcanol 68:266–284

Horwell CJ (2007) Grain size analysis of volcanic ash for the rapid assessment of respiratory health hazard. J Environ Monitor 9: 1107–1115

Horwell CJ, Baxter PJ (2006) The respiratory health hazards of volcanic ash: a review for volcanic risk mitigation. Bull Volcanol 69(1): 1–24

Horwell CJ, Fenoglio I, Fubini B (2007) Iron-induced hydroxyl radical generation from basaltic volcanic ash. Earth Planet Sci Lett 261(3–4):662–669

Horwell CJ, Le Blond JS, Michnowicz SAK, Cressey G (2010) Cristobalite in a rhyolitic lava dome: evolution of ash hazard. Bull Volcanol 72:249–253

Horwell CJ, Williamson BJ, Llewellin EW et al (2013a) Nature and formation of cristobalite at the Soufrière Hills volcano, Montserrat: implications for the petrology and stability of silicic volcanic domes. Bull Volcanol. In revision

Horwell CJ, Cole PD, Loughlin SC et al (2013 b) Cristobalite content of ash generated by 15 years of activity of the Soufrière Hills volcano, Montserrat. J Volcanol Geotherm Res. In revision

Horwell CJ, Williamson BJ, Le Blond JS, Donaldson K, Damby DE, Bowen L (2012) The structure of volcanic cristobalite in relation to its toxicity; relevance for the variable crystalline silica hazard. Particle and Fibre Toxicology 2012, 9:44

International Agency for Research on Cancer (1997) Silica, some silicates, coal dust and para-aramid fibrils. International Agency for Research on Cancer, Lyon, 506 pp

Jappinen P, Marttila O, Haahtela T (1990) Exposure to hydrogen sulphide and respiratory function. Br J Ind Med 47:824–828

Kizer KW (1984) Toxic inhalations. Emerg Med Clin N Am 2: 649–666

Kullmann CJ, Jones WG, Cornwell RJ et al (1994) Characterization of air contaminants formed by the interaction of lava and sea water. Environ Health Perspect 102(5):478–482

Le Blond JS, Horwell CJ, Baxter PJ et al (2010) Mineralogical analyses and in vitro screening tests for the rapid evaluation of the health hazard of volcanic ash at Rabaul volcano, Papua New Guinea. Bull Volcanol 72:1077–1092

Le Guern PL, Faivre-Pierret R (1982) An example of health hazard: people killed by gas during a phreatic eruption: Dieng Plateau (Java Indonesia), February 20th 1979. Bull Volcanol 45:153–156

Lohr AJ, Bogaard TA, Heikens A et al (2005) Natural pollution caused by the extremely acidic crater Lake Kawah Ijen, East Java, Indonesia. Environ Sci Pollut Res 12(2):89–95

Longo BM, Yang W, Green JB et al (2010) Acute health effects associated with exposure to volcanic air pollution (vog) from increased activity at Kilauea Volcano in 2008. J Toxicol Environ Health-Part A-Curr Issue 73(20):1370–1381

Lowe D (1986) Lahars Initiated by the 13 November 1985 Eruption of Nevado del Ruiz, Colombia. Nature 324:51–53

Mannino DM, Holschuh FC, Holscuh TC et al (1996) Emergency department visits and hospitalizations for respiratory disease on the island of Hawaii, 1981 to 1991. Hawaii Med J 55:48–54

Moseholm L, Taudorf E, Frosig A (1993) Pulmonary function changes in asthmatics associated with low-level SO_2 and NO_2 air pollution, weather, and medicine intake. An 8-month prospective study analyzed by neural networks. Allergy 48:334–344

Murata KJ (1966) The 1959–60 Eruption of Kilauea volcano, Hawaii; an acidic Fumarolic gas from Kilauea Iki. U. S. Geological survey professional paper, 537-C, 1–6

Myers B Driedger C (2008) Geologic hazards at volcanoes. U.S. Geological survey general information product 64, 1 sheet. http://pubs.usgs.gov/gip/64/. Accessed 3 March 2012

Newhall CG, Self S (1982) The volcanic explosivity index (VEI). J Geophys Res 87:1231–1238
Olsen KB, Fruchter JS (1986) Identification of hazards associated with volcanic emissions. Am J Publ Health 76(Suppl):45–52
Ostro BD, Wiener MB, Selner JC (1991) Asthmatic responses to airborne acid aerosols. Am J Publ Health 81(6):694–702
Rohl AM, Langer AM, Moncure G et al (1982) Endemic pleural disease associated with exposure to mixed fibrous dust in Turkey. Science 216:518–520
Roobol MJ, Smith AL (1989) Volcanic and associated hazards in the lesser antilles. In: Latter JH (ed) Volcanic hazards: assessment and monitoring. Springer, Berlin
Scientific Event Alert Network (SEAN) (1989) Global volcanism 1975–1985: the first decade of reports from the Smithsonian Institution's Scientific Event Alert Network. Prentice Hall, Upper Saddle River
Spallholz JE (1994) On the nature of selenium toxicity and carcinostatic activity. Free Radic Biol Med 17:45–64
Sparks RSJ, Bursik MI, Carey SN et al (1997) Volcanic plumes. Wiley, Chichester
Taylor GAM (1958) The 1951 eruption of Mt. Lamington. Papua. Bureau of Mineral Resources of Australia, Bulletin 38
Tazieff H, Sabroux J-C (1983) Forecasting volcanic events. Elsevier Press, Amsterdam
Vallyathan V, Robinson V, Reasor M et al (1984) Comparative *in vitro* cytotoxicity of volcanic ashes from Mount St. Helens, El Chichon, and Galunggung. J Toxicol Environ Health 14: 641–654
Wagner GN, Clark MA, Koenigsberg EJ et al (1988) Medical evaluation of the victims of the 1986 Lake Nyos disaster. J Forensic Sci 33:899–909
Weinstein P (2005) Palaeopathology by Proxy: the case of Egil's Bones. J Archaeol Sci 32:1077–1082
Weinstein P, Patel A (1997) The Mount Ruapehu eruption, 1996: a review of potential health effects. Aust N Z J Publ Health 21(7):773–778
Welch AH, Lico MS, Hughes JL (1988) Arsenic in ground water of the western United States. Ground Water 26:333–347
Williams SN, Moore JG (1983) Man against the volcano: the eruption on Heimaey, Vestmannaeyjar, Iceland. U. S. Geological Survey, Washington, DC
Witham CS (2005) Volcanic disasters and incidents: a new database. J Volcanol Geotherm Res 148(3–4):191–233
Witham CS, Oppenheimer C (2005) Mortality in England during the 1783–4 Laki Craters eruption. Bull Volcanol 67(1):15–26
Witham CS, Oppenheimer C, Horwell CJ (2005) Volcanic ash leachates: a review and recommendations for sampling methods. J Volcanol Geotherm Res 141:299–326
Yano E, Yokohama Y, Higashi H et al (1990) Health effects of volcanic ash: a repeat study. Arch Environ Health 45(6):367–373
Yang G, Ge K, Chen J, Chen X (1988) Selenium-related endemic diseases and the daily nutritional requirements of humans. World Rev Nutr Diet 55:98–152

Radon in Air and Water

11

J. Donald Appleton

Contents

11.1 Introduction

Radon is a natural radioactive gas that you cannot see, smell, or taste and that can only be detected with special equipment. It is produced by the radioactive decay of radium, which in turn is derived from the radioactive decay of uranium. Uranium is found in small quantities in all soils and rocks, although the amount varies from place to place. Radon decays to form radioactive particles that can enter the body by inhalation. Inhalation of the short-lived decay products of radon has been linked to an increase in the risk of developing cancers of the respiratory tract, especially of the lungs. Breathing radon in the indoor air of homes contributes to about 15,000 lung cancer deaths each year in the United States and 1,100 in the UK (HPA 2009). Only smoking causes more lung cancer deaths.

Geology is the most important factor controlling the source and distribution of radon (Appleton and Miles 2010). Relatively high levels of radon emissions are associated with particular types of bedrock and unconsolidated deposits, for example some, but not all, granites, phosphatic rocks, and shales rich in organic materials. The release of radon from rocks and soils is controlled largely by the types of minerals in which uranium and radium occur. Once radon gas is released from minerals, its migration to the surface is controlled by the transmission characteristics of the bedrock and soil; the nature of the carrier fluids, including carbon dioxide gas and groundwater; meteorological factors such as barometric pressure, wind, relative humidity, and rainfall; and soil permeability, drainage, and moisture content. The most important factors controlling the migration and accumulation of radon in buildings include (i) characteristics of the bedrock and soils that affect fluid transport, including porosity and permeability; (ii) the construction of the building and its use which includes the level of ventilation and heating; and (iii) environmental factors such as temperature (increased heating in buildings during the colder months causes a chimney effect which

J.D. Appleton (✉)
British Geological Survey, Nottingham, UK
e-mail: jda@bgs.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_11, British Geological Survey © NERC 2013

draws soil gasses including radon into the property), plus wind speed and direction which can increase the chimney effect. Radon-containing soil air can be drawn into buildings through gaps and cracks in solid floors, walls and service pipes below construction level; through the voids in suspended floors and crawl spaces, and via small cracks or pores in hollow-block walls or wall cavities (see also Chap. 10, this volume).

Radon levels in outdoor air, indoor air, soil air, and groundwater can be very different. Radon released from rocks and soils is quickly diluted in the atmosphere. Concentrations in the open air are normally very low and probably do not present a hazard. Radon that enters poorly ventilated buildings, caves, mines, and tunnels can reach high concentrations in some circumstances. The construction method and the degree of ventilation can influence radon levels in buildings. A person's exposure to radon will also vary according to how particular buildings and spaces are used.

The radon potential of the ground may be assessed from a geologically based interpretation of indoor radon measurements in conjunction with permeability, uranium, soil gas radon, and ground and airborne gamma spectrometric data. The categorization of a group of rocks or unconsolidated deposits as having known or suspected high levels of radon emissions does not imply that there is any problem. That would depend on whether pathways, locations for accumulation, and protracted exposure occur. Whereas geological radon potential maps do not give a direct guide to the level of radon in specific buildings or cavities, there is, in general, a higher likelihood that problems may occur at specific sites within areas of potentially high radon emissions.

Radon potential maps have important applications, particularly in the control of radon through environmental health and building control legislation. They can be used (1) to assess whether radon protective measures may be required in new buildings, (2) for the cost-effective targeting of radon monitoring in existing dwellings and workplaces, (3) to provide a radon potential assessment for homebuyers and sellers, and (4) for exposure data for epidemiological studies of the links between radon and cancer. Whereas a geological radon potential map can indicate the relative radon hazard, it cannot predict the radon risk for an individual building. This can only be established by having the building tested.

Radon dissolved in groundwater migrates over long distances along fractures and caverns depending on the velocity of fluid flow. Radon is soluble in water and may thus be transported for distances of up to 5 km in streams flowing underground in limestone. Radon remains in solution in the water until a gas phase is introduced (e.g., by turbulence or by pressure release). If emitted directly into the gas phase, as may happen above the water table, the presence of a carrier gas, such as carbon dioxide, would tend to induce migration of the radon. This appears to be the case in certain limestone formations, where underground caves and fissures enable the rapid transfer of the gas phase. Radon in water supplies can result in radiation exposure of people in two ways: by ingestion of the water or by release of the radon into the air during showering or bathing, allowing radon and its decay products to be inhaled. Radon in soil under homes is the biggest source of radon in indoor air, and it presents a greater risk of cancer than radon in drinking water.

This chapter explains how radon forms, the associated health risks, the kinds of rocks and soils it comes from, and how it moves through the ground and into buildings. It also explains how the radon potential of an area can be estimated. Greater detail on radon health risks, radon epidemiology, radon measurement methods, radon preventive measures in buildings, cost-effectiveness of radon control, radon risk communication and national radon programmes can be found in WHO (2009) and HPA (2009).

11.2 Nature and Measurement

11.2.1 Radioactivity and Radiation

All matter, including the materials that constitute the Earth's crust, consists of atoms, which are usually combined in various chemical compounds. Each atom comprises a nucleus, made up of protons, neutrons, and electrons, which orbit around the nucleus. Nuclei identified by the name of the element and the number of protons and neutrons are referred to as nuclides. All nuclei of the same chemical element have the same number of protons, but they can have different numbers of neutrons, and these are then called isotopes of that element. Many atoms are unstable and will change quite naturally into atoms of another element accompanied by the emission of ionizing radiation. This process is called radioactivity and the change is called radioactive decay. Unstable atoms that change through radioactive decay to form other nuclides are said to be radioactive and are referred to as radionuclides or radioisotopes. The rate of change or decay of an unstable radionuclide is indicated by its half-life, which is the period of time during which half the original number of atoms will have decayed.

The radiations most commonly emitted by radionuclides are alpha particles, beta particles, and gamma rays. The principal geological sources of radiation are *gamma radiation* from the ground and buildings and *radon* gas, which is derived mainly from uranium minerals in the ground. *Terrestrial gamma rays* originate chiefly from the radioactive decay of the natural potassium, uranium, and thorium, which

Table 11.1 The uranium-238 decay series

Nuclide	Principle mode of decay	Half-life
^{238}U	α	4.5×10^9 years
^{234}Th	ß	24.1 days
^{234}Pa	ß	1.2 min
^{234}U	α	2.5×10^5 years
^{230}Th	α	7.5×10^4 years
^{226}Ra	α	1,602 years
^{222}Rn	α	3.8 days
^{218}Po	α	3.1 min
^{214}Pb	ß	26.8 min
^{218}At	α	1.5 s
^{214}Bi	α	19.9 min
^{214}Po	α	1.6×10^{-4} s
^{210}Tl	ß	1.3 min
^{210}Pb	ß	22.6 years
^{210}Bi	ß	5.0 days
^{210}Po	α	138.4 days
^{206}Tl	ß	4.2 min
^{206}Pb	Stable	Stable

Table 11.2 The thorium-232 decay series

Nuclide	Principle mode of decay	Half-life
^{232}Th	α	1.4×10^{10} years
^{228}Ra	ß	5.8 years
^{228}Ac	ß	6.1 h
^{228}Th	α	1.9 years
^{224}Ra	α	3.7 days
^{220}Rn	α	55.6 s
^{226}Po	α	0.15 s
^{212}Pb	ß	10.6 h
^{212}Bi	α 36%	60.5 min
	ß 64%	
^{212}Po	α	3.0×10^{-7} s
^{208}Tl	ß	3.1 min
^{207}Pb	Stable	Stable

are widely distributed in terrestrial materials including rocks, soils, and building materials extracted from the Earth.

There are three naturally occurring radon (Rn) isotopes: ^{219}Rn (actinon), ^{220}Rn (thoron), and ^{222}Rn, which is commonly called radon. ^{219}Rn (actinon) has a very short half-life of about 3 s and this, together with its occurrence in the decay chain of ^{235}U (which is only present as 0.7% of natural uranium), restricts its abundance in gases from most geological sources. Actinon does not escape to air in significant quantities. ^{222}Rn (radon) is the main radon isotope of concern to man. It occurs in the uranium-238 decay series (Table 11.1), has a half-life of 3.82 days and provides about 50% of the total radiation dose to the average person. ^{222}Rn is produced by the radioactive decay of solid radium (^{226}Ra). ^{220}Rn (thoron) is produced in the thorium-232 decay series (Table 11.2). ^{220}Rn has been recorded in houses, and about 4% of the average total radiation dose for a member of the UK population is from this source.

Most of the radon that is inhaled is exhaled again before it has time to decay and irradiate tissues in the respiratory tract. Radon (^{222}Rn), however, decays to form very small solid radioactive particles, including polonium-218, that become attached to natural aerosol and dust particles. These may remain suspended in the air or settle onto surfaces. When these particles are inhaled, they irradiate the lining of the bronchi in the lung with alpha particles and this may increase the risk of developing lung cancer.

11.2.2 Measurement of Radioactivity

There are a number of different ways to measure radioactivity. These include (1) the *radioactivity* of a radioactive material, such as radon gas; (2) the *dose* to living tissue, e.g., to the lungs from solid decay products of radon gas; and (3) the *exposure* caused by the presence of radioactivity. There are also environmental or safety thresholds of radioactivity such as *dose limit, action level,* and *reference level,* which are used in legislation and advice. The units of radioactivity and dose are summarized in Table 11.3. In the United States, radioactivity is commonly measured in pico curies (pCi), named after the French physicist Marie Curie, who was a pioneer in the research on radioactive elements and their decay. In most other countries, and throughout this chapter, radioactivity is measured using the SI unit becquerel (Bq). One becquerel represents one atomic disintegration per second. The level of radioactivity in the air due to radon is measured in becquerels per cubic meter (Bq m^{-3}) of air. The average radon concentration in houses in Great Britain is 20 Bq m^{-3}, that is, 20 radon atoms disintegrate every second in every cubic meter of air. The average in the United States is 46 Bq m^{-1}. A 1,000-square foot house with 46 Bq m^{-3} of radon has nearly two million radon atoms decaying in it every minute.

Radon levels in outdoor air, indoor air, soil air, and groundwater can be very different. Radon in outdoor air is generally low (4–8 Bq m^{-3}) but may be as high as 100 Bq m^{-3} in some valleys when measured in the early morning. Radon in indoor air ranges from less than 20 Bq m^{-3} to about 110,000 Bq m^{-3} with a population-weighted world average of 40 Bq m^{-3} and country averages ranging from nine in Egypt, 20 in the UK, 22 in New Zealand, 44 in China, 46 in the United States,108 in Sweden, 120 in Finland to 140 Bq m^{-3} in the Czech Republic (UNSCEAR 2009; HPA 2009). Radon concentrations in outdoor air in the UK are generally low, on average 4 Bq m^{-3} whilst radon in indoor air in UK dwellings ranges from less than 10 Bq m^{-3} to over 17,000 Bq m^{-3} (Rees et al. 2011). Radon in soil air (the air that occupies the pores in soil) ranges from less than 1 to

Table 11.3 Units of measurement of radioactivity and dose

Quantity	Unit	Purpose	Comments
Activity	Becquerel (Bq)	Measure activity of a radioactive material (solid or gas); The International System of Units (SI) definition of Activity	1 Bq = 1 Atomic disintegration per second
	Curie	In the USA, the activity (rate of decay) of ^{222}Rn is expressed in units called curies	The curie is based on the rate of decay of one gram of ^{226}Ra or 3.7×10^{10} disintegrations per second
	Pico curies (pCi)		1 pCi = one trillionth of a Curie; 0.037 disintegrations per second, or 2.22 disintegrations per minute
Radioactivity in air or water	Becquerels/m^{-3} (Bq m^{-3})	Measure average concentration of radon gas in building or in soil air. Bq/L used to measure radon in water	Average level of radon in houses in Great Britain is 20 Bq m^{-3}; in Sweden 108 Bq m^{-3}
	Picuries/L^{-1} (pCi L^{-1})	Unit used in USA	Average level of radon in houses in USA is 1.24 pCi L^{-1} equivalent to 46 Bq m^{-3}
Absorbed dose	Gray (Gy)	Measure energy per unit mass absorbed by tissue	1 J of energy absorbed by 1 kg of tissue
	Rad	Old unit of absorbed dose	1 rad = 0.01 Gy
Equivalent dose	Sievert (Sv)	Measure of absorbed doses caused by different types of radiation	Absorbed dose weighted for harmfulness of different radiations
	Roentgen man equivalent (rem)	Old measure of absorbed dose	The rem is being replaced by the Sievert, which is equal to 100 rem

more than 2,500 Bq L^{-1}; most soils in the United States contain between 5 and 55 Bq L^{-1} radon in soil air. The amount of radon dissolved in groundwater ranges from about 3 to nearly 80,000 Bq L^{-1}.

The *absorbed dose* is the energy absorbed by a unit mass of tissue whereas the *dose equivalent* takes account of the relative potential for damage to living tissue of the different types of radiation. The dose equivalent is the absorbed dose multiplied by a "quality factor," which is 1 for beta and gamma rays and 20 for alpha particles. This is because alpha particles deposit their energy more densely. In addition, alpha particles transfer all their energy in short distances so that a relatively small volume of tissue receives a high dose of radiation. The commonly used unit for dose equivalent is the *sievert* (1Sv = 1,000 millsieverts; mSv).

The dose equivalent indicates the potential risk of harm to particular tissues by different radiations, irrespective of their type or energy. *Risk weighting factors* are an approximate measure of the risk to particular parts of the body for a given dose equivalent. Some parts are more susceptible to radiation damage (e.g., lungs, bone marrow, or gonads). These have higher risk-weighting factors than other parts of the body. An overall *effective dose equivalent* for the whole body can be calculated from the organ dose equivalents and risk-weighting factors. The annual effective dose equivalent for the average member of the UK population arising from all sources is 2.5 mSv, to which exposure of the lungs by radon and its daughters contributes about half. Exposure in the home to a radon gas concentration of 48 Bq m^{-3} would lead, in the course of a year, to an effective dose equivalent of 1 mSv (ICRP 1993). In the United States, the average person is exposed to an effective dose equivalent of approximately 3.6 mSv (whole-body exposure) per year from all sources.

Governments set occupational *dose limits* in order to ensure that individuals are not exposed to an unacceptable degree of risk from artificial radiation. Occupational levels are conventionally expressed in working level (WL) units. A WL is any combination of short-lived radon daughters (decay products or progeny ^{218}Po, ^{214}Pb, ^{214}Bi, and ^{214}Po) in one liter of air that will result in the emission of 1.3 X 10^5 MeV of potential alpha energy. Exposures are measured in working level months (WLM). A WLM is the cumulative exposure equivalent to 1 WL for a working month (170 h). In SI units, a WLM is defined as 3.54 mJhm^{-3} (ICRP 1993). One WL is approximately equal to a radon exposure of 7,500 Bq m^{-3} and 1 WLM to an average radon exposure of about 144 Bq m^{-3} y (on the assumption that people spend most of their time indoors) (NRPB 2000).

The International Commission for Radiological Protection (ICRP) has recommended that all radiation exposures should be kept as low as reasonably achievable taking into account economic and social factors. In the UK, statutory regulations apply to any work carried out in an atmosphere containing ^{222}Rn gas at a concentration in air, averaged over any 24-h period, exceeding 400 Bq m^{-3} except where the concentration of the short-lived daughters of ^{222}Rn in air averaged over any 8-h working period does not exceed 6.24 X 10^{-7}Jm^{-3}. The limit on effective dose for any employee of 18 years of age or above is 20 mSv in any calendar year. This dose limit may be compared with the

dose to the average person in the UK of 2.5 mSv, the dose of 7.5 mSv to the average person living in the high radon area of Cornwall, UK, and 4.5 mSv to the average nuclear worker in the UK.

In the United States, exposure limits vary by regulating agency and type of worker. The Miners Safety and Health Act (MSHA) covers underground miners, whereas the Occupational Safety and Health Act (OSHA) regulates exposure to ^{222}Rn gas and ^{222}Rn progeny for workers other than miners. The MSHA sets limits so that no employee can be exposed to air containing ^{222}Rn progeny in excess of 1 WL (100 pCi L^{-1}) in active work areas. The MSHA also limits annual exposure to ^{222}Rn progeny to less than 4 WLM per year. OSHA limits exposure to either 30 pCi L^{-1} or 0.33 WL based on continuous workplace exposure for 40 h per week, 52 weeks per year.

A number of occupations have the potential for high exposure to ^{222}Rn progeny: mine workers, including uranium, hard rock, and vanadium; workers remediating radioactive contaminated sites, including uranium mill sites and mill tailings; workers at underground nuclear waste repositories; radon mitigation contractors and testers; employees and recreational visitors of natural caves (Gillmore et al. 2000, 2002); phosphate fertilizer plant workers; oil refinery workers; utility tunnel workers; subway tunnel workers; construction excavators; power plant workers, including geothermal power and coal; employees of radon health mines; employees of radon balneotherapy spas (waterborne ^{222}Rn source); workers and visitors to hot spring hotels (Song et al. 2011); water plant operators (waterborne ^{222}Rn source); fish hatchery attendants (waterborne ^{222}Rn source); employees who come in contact with technologically enhanced sources of naturally occurring radioactive materials; and incidental exposure in almost any occupation from local geologic ^{222}Rn sources. Recreational and other visitors to abandoned metalliferous mines may also be exposed to high radon concentrations (Gillmore et al. 2001).

11.3 Health Effects of Radiation and Radon

Radiation can interact with the electrons in surrounding molecules in the cells and induce changes such as ionization. Ionization of water molecules in organic tissues can alter important molecules in that tissue. Radiation can also ionize and produce chemical changes in DNA molecules, the basic material that controls the structure and function of the cells that make up the human body. This can lead to biological effects, including abnormal cell development, some of which may not be seen for some time after radiation exposure. Alpha particles are considered to be the most dangerous type of radiation. Although they do not penetrate very far, the mass and charge of the particles is so high that it can cause intense ionization. Whereas alpha radiation cannot penetrate the surface layer of the skin (stratum corneum), the interior of the lungs lacks a protective epidermis so that alpha decay particles emitted by radon progeny can damage important molecules in the cells. Gamma rays are very penetrative and can cause ionization and tissue damage comparable in effect to x-radiation, but are usually much more energetic. Provided the radioactive sources remain outside the body, gamma radiation is the greatest problem because it is so penetrating, whereas alpha particles are stopped by clothing and the outer layers of the skin. Beta particles are intermediate in penetrating power.

When radioactive sources are taken into the body however, the situation changes markedly. The major pathways by which alpha activity enters the human body are the ingestion of radioactive elements and inhalation of radon, and more importantly its daughter products, some of which are alpha particle emitters (^{218}Po and ^{214}Po). Alpha particles give up their energy to a very small volume of tissue and can thus cause intensive damage, which has been shown to result in cancers. Much of the inhaled radon is exhaled and relatively few alpha particles are emitted by it within the body. However, the four immediate decay products of ^{222}Rn have short half-lives and are all radioactive isotopes of solid elements (Table 11.4). The decay products, which remain in suspension attached to the surface of aerosols, dust, smoke, or moisture particles, or are unattached, may remain in the respiratory system where they may become trapped in the lungs and irradiate the cells of mucous membranes, bronchi, and other pulmonary tissues. Overall doses are due largely to irradiation of the bronchial epithelium and secretory cells by alpha particles from the short-lived decay products of ^{222}Rn. It is believed that the ionizing radiation energy affecting the bronchial epithelial cells initiates carcinogenesis. As a consequence the main danger is an increased risk of developing cancers of the respiratory tract, especially the lungs. Whereas radon-related lung cancers occur primarily in the upper airways, radon increases the incidence of all histological types of lung cancer, including small cell carcinoma, adenocarcinoma, and squamous cell carcinoma. The contribution to both lung dose equivalent and effective dose equivalent by the beta and gamma radiations may be ignored, as they are small compared to those from alpha radiation.

It is interesting to note that radon remained a chemical curiosity for decades, even promoted at times as a "health-giving" gas at various spas. Initially radon was regarded as a fairly innocuous or even benign component of geological gases, and its importance as the major contributor to the radioactive dose received by the general population has been recognized relatively recently. In contrast to the early dramatic effects of high radiation doses on humans, which

Table 11.4 Principal decay properties of radon (^{222}Rn) and short-lived decay products (After Green et al. 1992)

			Main radiation energies and intensities					
			α		ß		γ	
Radionuclide		Half-life	MeV	%	MeV	%	MeV	%
^{222}Rn	Gas	3.824 day	5.49	100	–	–	–	–
^{218}Po	Solid	3.11 min	6.00	100	–	–	–	–
^{214}Pb	Solid	26.8 min	–	–	1.02	6	0.35	37
					0.70	42	0.30	19
					0.65	48	0.24	8
^{214}Bi	Solid	19.7 min	–	–	3.27	18	0.61	46
					1.54	18	1.77	16
					1.51	18	1.12	15
^{214}Po	Solid	1.64×10^{-4} s	7.69	100	–	–	–	–

can cause death in a few days or weeks, or obvious skin damage when a limited area of the body is exposed to a high radiation dose, the effects of the relatively low doses of natural radiation (e.g., cancer) usually occur a long time after exposure.

The overall hazard to human health from gamma radiation, either indoors or outdoors, is negligible compared with the hazard associated with radon.

11.4 Radon Epidemiology

Evidence linking the exposure to high levels of radon and an increase in the risk of lung cancer is becoming overwhelming. Indeed more is known about the health risks of radon exposure than about most other human carcinogens. A large body of epidemiological data has accumulated over several decades relating to studies of the incidence of lung cancer in miners and risk estimates have been derived from these data (NAS 1998; NRPB 2000). Supporting evidence comes from experimental studies of animals and from radiobiology.

11.4.1 Cohort Studies of Miners

High death rates from lung cancer recorded in the Middle Ages among miners in Germany and the Czech Republic are now recognized as radon-induced. Studies of thousands of miners, some with follow-up periods of more than 30 years, have been conducted in uranium, iron, tin, and fluorspar mines in Australia, Canada, China, Europe, and the United States. These studies consistently demonstrated an increase in lung cancer incidence with exposure to radon decay products, despite several differences in study populations and methodologies. The miner studies demonstrated that (1) at equal cumulative exposures, low exposures in the range of the U. S. Environmental Protection Agency's (EPA) 4 pCi L^{-1} (148 Bq m^{-3}) action level over longer periods produced greater lung cancer risk than high exposures over short periods; (2) increased lung cancer risk with radon exposure was observed even after controlling for, or in the absence of, other potentially confounding mine exposures such as asbestos, silica, diesel fumes, arsenic, chromium, nickel, and ore dust; (3) increased lung cancer risk has been observed in miners at relatively low cumulative exposures in the range of the U. S. EPA's 4 pCi L^{-1} (148 Bq m^{-3}) action level; and (4) nonsmoking miners exposed to radon have been observed to have an increased risk of lung cancer.

A major reassessment of health information mainly on uranium miners from Czechoslovakia, Colorado, Sweden, and Canada by the IRCP (1993) demonstrated a pronounced excess of lung cancers. More recently the Sixth Committee on the Biological Effects of Ionizing Radiation (BEIR VI) of the American National Academy of Sciences (NAS 1998) re-analyzed the data for miner cohorts and used the most recent information available to estimate the risks posed by exposure to radon in homes.

Results from 12 major epidemiological studies involving a total of more than 60,000 miners clearly indicate a correlation between excess mortality from lung cancer and radon exposure (Fig. 11.1). A combined analysis of studies on underground miners revealed an increase in relative risk from about 2% at a mean exposure of 250 WLM to 10% at 2,500 WLM (Lubin et al. 1994).

Differences between mine and home environments could limit the validity of extrapolating risk estimates based on miner data to the home environment. These include generally higher ^{222}Rn gas concentrations, more airborne dust, and larger dust particle diameter in mines compared with homes as well as different activity size distributions of radon progeny and rates of attachment. Other toxic pollutants present in mine air may act as confounders. In addition, there are age and sex differences between miners and the general population; higher levels of physical activity among miners, which affects respiration rates; and more oral (as opposed to nasal) breathing in the miners, which leads to increased deposition of larger

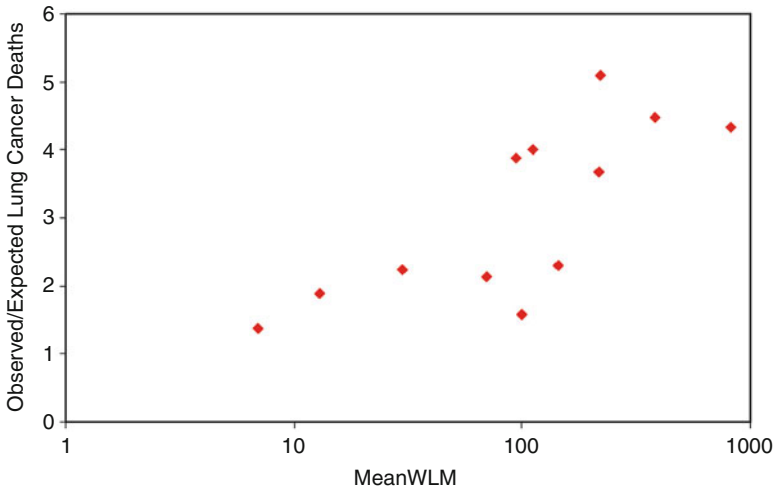


Fig. 11.1 Excess mortality from lung cancer among miners exposed to high levels of radon. *WLM* working level month, a unit of exposure to radon (Data from NRPB 2000)

particles into the lung. Miners have shorter term high exposure compared with the lifelong lower concentration exposure for the general population. Finally, most miners were smokers compared with a minority of the general population. Even allowing for these factors, the evidence for a causal association between lung cancer and occupational radon exposure in underground miners is overwhelming (see also Chap. 13, this volume).

11.4.2 Case Control Studies of the Effects of Domestic Exposure

Comparison of radon exposures among people who have lung cancer with exposures among people who have not developed lung cancer is the most direct way to assess the risks posed by radon in homes. Many factors must be considered when designing a domestic case control radon epidemiology study. These factors include

1. Mobility: People move a lot over their lifetime and it is virtually impossible to test every home where an individual has lived; estimates of radon exposure have to be used to fill in gaps in exposure history.
2. Housing stock changes: Over time, older homes are often destroyed or reconstructed so that radon measurements will be not available or vary dramatically from the time of occupancy by the case; a home's radon level may increase or decrease over time if new ventilation systems are installed, the occupancy patterns may change substantially, or the home's foundation may shift and cracks appear.
3. Inaccurate histories: Often a majority of the lung cancer cases (individuals) studied are deceased or too sick to be interviewed by researchers. This requires reliance on second-hand information, which may not be accurate.

These inaccuracies primarily affect

1. Residence history: A child or other relative may not be aware of all residences occupied by the patient, particularly if the occupancy is distant in time or of relatively short duration. Even if the surrogate respondent is aware of a residence he or she may not have enough additional information to allow researchers to locate the home.
2. Smoking history: Smoking history historically has reliability problems. Individuals may underestimate the amount they smoke. Conversely, relatives or friends may overestimate smoking history.
3. Other factors: Complicating factors other than variations in smoking habits include an individual's genetics, lifestyle, exposure to other carcinogens, and home heating, venting, and air conditioning preferences.

Several such case control (or cohort) studies have been completed but they have not produced a definitive answer, principally because the risk is small at the low exposure of most domestic environments. In addition, many people involved in the studies moved a number of times so it was difficult to estimate the radon exposures that people had received over their lifetimes. The greatest problem, however, was caused by the fact that far more lung cancers are caused by smoking than are caused by radon (NAS 1998).

Residential epidemiological case–control studies examining the relationship between contemporary ^{222}Rn gas concentrations in homes and lung cancer have been performed in Canada, China, Finland, Germany, Sweden, the UK, and the United States. These studies indicate that higher lung cancer rates occur in those exposed to higher levels of radon, although in most cases this did not reach a statistical level of significance (Lubin and Boice 1997; Darby et al. 1998).

Meta-analysis is a statistical attempt to analyze the results of several different studies to assess the presence or absence

of a trend or to summarize results. Meta-analysis of the largest case control studies produced a positive risk estimate that was statistically significant and close to that derived from the miner data (Lubin and Boice 1997). The meta-analysis of eight studies using weighted linear regression found a summary excessive risk of 14% at an average indoor ^{222}Rn gas concentration of 4 pCi L^{-1} (148 Bq m^{-3}). The excess risk at 4 pCi L^{-1} (148 Bq m^{-3}) in recent studies in Germany and the UK was in close agreement with risk estimates obtained from the meta-analysis. Lubin and Boice (1997) concluded that the results of their meta-analysis are consistent with the current miner-based estimates of lung cancer risk from radon, which place the number of radon-related deaths at approximately 15,000 per year in the United States. Because meta-analysis has several inherent limitations (such as the inability to adequately explore the consistency of results within and between studies and to control for confounding factors), meta-analysis is not able to prove that residential radon causes lung cancer. But it does provide additional good suggestive evidence. It is one more link in the "chain of evidence" connecting residential radon exposure to increased lung cancer risk. Because the investigators performing a meta-analysis do not have access to the raw data on the individual study subjects, the analysis is based on the published relative risks and confidence intervals of the individual studies. Frequently, the impact of each study is weighted based on some factor considered relevant to the reliability of each study's data. In the Lubin and Boice (1997) meta-analysis, the results of each individual study were weighted so that each study contributed in relation to the precision (i.e., relative lack of random or sampling errors) of its estimate.

An exposure-rate effect is the alteration of an effect by intensity of an exposure. An inverse exposure-rate effect would be the enhancement of an effect as the intensity of the exposure decreases (i.e., low-level chronic exposures would be riskier than high-level more acute exposures). An "inverse exposure-rate effect" was observed in the miner data. This means that for miners who received the same exposure, those that received it over a longer period of time had a greater risk of lung cancer. The inverse exposure-rate effect diminished in miners exposed below 50–100 WLM. The finding that the inverse exposure-rate effect does not seem to apply in residential situations will not change the EPA's risk assessment since the EPA had not included the inverse exposure-rate effect in their latest, 1992, risk estimate.

Collaborative analysis of individual data from 13 case–control studies of residential radon and lung cancer in 9 European countries showed that the risk of lung cancer increased by 16% (95% confidence interval = 5–31%) per 100 Bq m^{-3} increase in radon after correction for random uncertainties in measuring radon concentrations (Darby et al. 2005). The dose–response relation seemed to be linear with no threshold and remained significant (P = 0.04) in analyses limited to individuals from homes with measured radon <200 Bq m^{-3}. The results suggest that radon is responsible for about 2% of all deaths from cancer in Europe. Systematic analysis of pooled data from seven North American residential case–control radon studies showed that the odds ratios (ORs) for lung cancer increased with residential radon concentration. The estimated OR after exposure to radon at a concentration of 100 Bq m^{-3} over a 25 year period was 1.11 (1.00–1.28), which is compatible with the estimate of 1.12 (1.02–1.25) predicted by downward extrapolation of the results from occupational studies of radon-exposed underground miners (Krewski et al. 2005, 2006).

11.4.3 Ecological (Geographical) Epidemiological Studies

Ecological epidemiological studies of the associations between average lung cancer rates and radon concentrations in geographical areas are considered to be much less reliable than case–control studies that consider individual radon exposure and smoking histories. A negative correlation between mean radon and lung cancer rates in counties in the United States (Cohen 1997) is not well understood and the study methodology has been criticized by epidemiologists (e.g., Lubin 1998; Smith et al. 1998). Ecological studies of the association between childhood cancer and indoor radon exposure are reviewed in HPA (2009) (see also Chap. 22, this volume).

11.4.4 Extrapolation from Mines to Homes

Whereas it has been suggested that the dose response relationship seen in miners may not extend to the much lower levels present in most homes, there appears to be sufficient evidence to suggest that a dose threshold for radiation carcinogenesis does not apply to lung cancer. Ionizing radiation is thought to induce specific gene mutations in DNA in single target cells in tissue and, as such, act principally at the initial stage of cancer. The number of cells hit by alpha particles will be broadly proportional to the dose (i.e., the radon concentration in a dwelling). The general consensus is that low dose (i.e., domestic radiation) cancer risk rises in proportion to the dose and there is not a threshold below which risk may be discounted (NRPB 2000).

More information about residential exposure to radon is needed to answer important questions about radon's effect on women and children—groups not included in the occupational studies of miners. Although children have been reported to be at greater risk than adults for developing certain types of cancer from radiation, currently there is no conclusive evidence that radon exposure places children at

any greater risk. Some miner studies and animal studies indicate that for the same total exposure, a lower exposure over a longer time is more hazardous than short, high exposures. These findings increase concerns about residential radon exposures.

11.4.5 Experimental Studies with Animals

Results from animal experiments conducted in the United States and France are generally consistent with the human epidemiological data. Health effects observed in animals exposed to radon and radon decay products include lung carcinomas, pulmonary fibrosis, emphysema, and a shortening of life span (USDOE 1988). The incidence of respiratory tract tumors increased with an increase in cumulative exposure and with a decrease in rate of exposure (NAS 1998). Increased incidence of respiratory tract tumors was observed in rats at cumulative exposures as low as 20 WLM (NAS 1998). Exposure to ore dust or diesel fumes simultaneously with radon did not increase the incidence of lung tumors above that produced by radon progeny exposures alone (USDOE 1988). Lifetime lung tumor risk coefficients observed in animals are similar to the lifetime lung cancer risk coefficients observed in human studies (USDOE 1988). In a study of rats simultaneously exposed to radon progeny and uranium ore dust, it was observed that the risk of lung cancer was elevated at exposure levels similar to those found in homes. The risk decreased in proportion to the decrease in radon-progeny exposure (Cross 1992). Confounding factors such as smoking are more readily controlled in animal experiments and qualitatively confirmed that radon can indeed induce lung cancer in the absence of smoking. Animal evidence on radon and lung cancer is reviewed in HPA (2009).

11.4.6 Other Cancers and Radon Exposure

No consistent association has been observed between radon exposure and other types of cancer. A combined analysis of the data from 11 miner cohort studies involving more than 60,000 miners did not find convincing links (Darby et al. 1995). No clear association between radon and childhood cancer (especially leukemia) emerged from a number of ecological studies, and a review of ecological, miner, and cohort studies did not find an association between radon and leukemia (Laurier et al. 2001).

Radon ingested in drinking water may lead in some circumstances to organs of the gastrointestinal tract receiving the largest dose. Ingested radon is absorbed by the blood. Most of the radon is lost quickly from the bloodstream through the lungs but some will deliver a dose to other body organs, especially those with a high fat content due to the higher solubility of radon in fat compared with water. Other body organs may be irradiated to some extent although the doses involved will be much smaller. Comparative estimated doses to various organs from exposure to radon are indicated in Table 11.5. Apart from lung cancer, there is no epidemiological proof of radon causing any other type of cancer. If radon and its decay products have an effect on organs other than the lungs, this is too weak to be readily detectable (HPA 2009).

Table 11.5 Estimated annual absorbed doses to adult tissues from ^{222}Rn and its short-lived progeny for domestic ^{222}Rn concentration of 20 Bq m^{-3} (NAS 1998)

Tissue	Annual dose (μGy y^{-1})
Lung	500
Skin[a]	50–1,000
Red bone marrow	0.5–6
Bone surface	0.4–4.4
Breast	1.2–1.5
Blood	1.1
Liver	2.5
Kidney	14.4

[a]Basal cells at 50 μm in exposed skin

11.5 Radon Health Risks

Because a valid risk estimate could not be derived only from the results of studies in homes, the BEIR VI committee chose to use data from studies of miners to estimate the risks posed by radon exposures in homes (NAS 1998). The committee statistically analyzed the data from 11 major studies of underground miners, which together involved about 68,000 men, of whom 2,700 have died from lung cancer. A range of models was used to try to explain the relationship between radon and smoking. In the multiplicative model it is assumed that a specific radon exposure will multiply the base risk rate for smokers and non-smokers by the same factor. BEIR VI models take into account total exposure, age and duration of exposure or total exposure, and age and average radon exposure with predicted risks at about 50% higher under the first of these two models. In general, the risk of developing lung cancer increases linearly as the exposure increases.

Epidemiological studies confirm that exposure to radon in homes increases the risk of lung cancer in people. It is estimated that between 3 and 14% of all lung cancers are linked to radon. Radon is the second most important cause of lung cancer after smoking and is much more likely to cause lung cancer in people who smoke, or who have smoked, than in lifelong non-smokers for whom radon is the primary cause of lung cancer (WHO 2009).

A study of lung cancer deaths from indoor radon in the UK (HPA 2009) estimated that: (i) 1,100 lung cancer deaths were caused by radon, representing 3.3% of total lung cancer deaths (34,000); (ii) the dose–response relationship appears linear, in that the greater the concentration of indoor radon, the greater the risk of developing lung cancer; (iii) there is no evidence of any threshold below which there is no risk; (iv) of the 3.3% of lung cancer deaths attributable to radon exposure, only 0.5% were due to radon acting alone: the remaining 2.8% were caused by a combination of radon and smoking, with nearly half the deaths likely to occur in people who had already given up smoking. The HPA research also highlighted that the vast majority of radon-induced lung cancer deaths occur in areas that are not currently designated as 'Radon Affected Areas' (areas in which over 1% of homes are estimated to exceed 200 Bq m^{-3}, the UK radon Action Level). The overwhelming majority of the population live outside such areas, and around 95% of radon-attributable deaths are estimated to occur with residential concentrations below 200 Bq m^{-3}, and 70% at concentrations of less than 50 Bq m^{-3}. This has major implications for the cost-effectiveness of government intervention strategies designed to manage exposure to radon in the domestic environment (HPA 2009; WHO 2009).

The number of lung cancer cases due to residential radon exposure in the United States was estimated to be 15,400 (exposure-age-duration model) or 21,800 (exposure-age-concentration model), which is 10–15% of lung cancer deaths. Radon causes 11% of lung cancer deaths among smokers (most of whom die of smoking) but 23% of never-smokers. The BEIR VI committee's uncertainty analyses using the constant relative risk model suggested that the number of lung cancer cases could range from about 3,000–32,000. The 95% upper confidence limit for the exposure-age-concentration model was approximately 38,000, but it was considered that such a high upper limit was highly unlikely given the uncertainty distributions. The major shortcomings in the existing data relate to estimating lung cancer risks near 148 Bq m^{-3} (4 pCi L^{-1}) and down to the average U. S. indoor level of 46 Bq m^{-3} (1.24 pCi L^{-1}), especially the risks to never-smokers.

Most of the radon-related deaths among smokers in the United States would not have occurred if the victims had not smoked. Whereas there is evidence for a synergistic interaction between smoking and radon, the number of cancers induced in ever-smokers by radon is greater than one would expect from the additive effects of smoking alone and radon alone. The estimated number of deaths attributable to radon in combination with cigarette smoking and radon alone in never-smokers constitutes a significant public-health problem and makes indoor radon the second leading cause of lung cancer after cigarette smoking.

Table 11.6 Fatal lifetime lung cancer risks for lifetime radon exposure at 200 Bq m^{-3} based on BEIR VI models

	Risk (%)
General population	3–5
Smokers	10–15
Non-smokers	1–3

From NRPB (2000)

If the radon risk in China and India is approximately the same as in the UK, then, based on similar average indoor radon concentrations (Cheng et al. 2002; Ramachandan and Sathish 2011), radon may cause over 200,000 deaths per year in each of these two countries. The percentage of lung cancer deaths attributed to radon is 7.8% in Canada, 5% in Germany, 8.3% in Switzerland, and between 5 and 12% in France (WHO 2009).

The BEIR VI committee suggested that about one-third of the radon-attributed cases (about 4% of the total lung cancer deaths) would be avoided if all homes had concentrations below the EPA's action guideline of 148 Bq m^{-3} (4 pCi L^{-1}). Of these deaths, about 87% would be in ever-smokers. Deaths from radon-attributable lung cancer in smokers could be reduced most effectively through reduction in smoking, because most of the radon-related deaths among smokers would not have occurred if the victims had not smoked. Whereas the relative risks for smokers and nonsmokers is still disputed, evidence from miners who never smoked demonstrates a clear relationship between cumulative exposure and relative risk. Existing biologic evidence indicates that even very low exposure to radon might pose some risk but that a threshold level of exposure, below which there is no effect of radon, cannot be excluded.

BEIR VI risk models have been used to estimate fatal lifetime lung cancer risk for lifetime exposure at 200 Bq m^{-3} (Table 11.6). More recently, the UK HPA used risk estimates from the European pooling study (Darby et al. 2005) to calculate the cumulative absolute risk of death from lung cancer to age 75 years in the UK by long-term average residential radon concentrations and smoking history (Table 11.7). Combining the results of the European, North American and Chinese pooling studies gives an estimated increased risk of lung cancer of 10% per 100 Bq m^{-3} increase in measured radon concentration (WHO 2009).

At the mean domestic radon concentration of 21 Bq m^{-3}, the risk for a continuing cigarette smoker is more than 30 times greater than for a lifelong non-smoker. The risk is about 30% higher in houses with radon concentrations at the current UK action level of 200 Bq m^{-3}. To put this in perspective, risk of death from accidents in the home is 0.7%; risk of premature death from accidents on the road is 2.5%, while there is a 5–7.5% overall risk of lung cancer in the UK.

Table 11.7 Cumulative absolute risk of death from lung cancer to age 75 years in the UK by long-term average residential radon concentrations and smoking history (Adapted from Table 4.3, HPA 2009)

	Cumulative risk of death from lung cancer to age 75 years (%)		
Long-term average radon concentration (Bq m^{-3})	Lifelong non-smoker	Ex-cigarette smoker stopped at age 50	Continuing cigarette smoker
0	0.41	5.5	14.7
21[a]	0.42	5.7	15.2
100	0.47	6.4	16.9
200	0.53	7.2	19
400	0.66	8.9	23
800	0.92	12.2	30.5

[a]Mean UK long-term average residential radon concentration

Duport (2002) questioned whether radon risk is overestimated because only the exposure to inhaled radon decay products is generally taken into account in the determination of risk of radiogenic lung cancer in uranium miners, whereas the risk actually reflects the total dose of radiation received by the lung. Radiation dose from sources other than ^{222}Rn decay products may account for 25–75% of the total effective dose, absorbed dose, or equivalent lung dose and this varies between mines. Neglecting these doses would lead to overestimation of risk both through dose underestimation and misclassification. Correction for neglected doses and dose misclassification would reduce the risk per unit of radon exposure by a factor of at least two or three and bring the overall dose-effect relationship toward the no-effect null hypothesis. This would increase the likelihood of a radon exposure threshold for lung cancer risk at current indoor exposure levels (Duport 2002).

The U. S. EPA estimates that radon in drinking water causes about 168 cancer deaths per year, 89% from lung cancer caused by breathing radon released from water, and 11% from stomach cancer caused by drinking radon-containing water. In general, radon released from tap water and inhaled will present a greater risk than radon ingested through drinking water (NRPB 2000; HPA 2009).

Estimation of the economic cost of radon-induced lung cancers is difficult. Using the Quality Adjusted Life Years approach (Gray et al. 2009), the total estimated cost of radon-induced lung cancers is £347 million per year in the UK.

11.6 Sources of Natural Radiation

11.6.1 Introduction

The average person in the UK receives an annual effective radiation dose of 2.8 mSv, of which about 85% is from natural sources: cosmic rays, terrestrial gamma rays, the decay products of ^{220}Rn and ^{222}Rn, and the natural radionuclides in the body ingested through food and drink. Of these the major proportion is from geological sources. About 60% of the total natural radiation dose is from radon isotopes (mostly due to alpha particle activity) while about 15% is thought to be due to gamma radiation from the U, Th, and K in rocks and soils and from building products produced from geological raw materials (Åkerblom et al. 2005a). X-rays and radioactive materials used to diagnose disease are the largest source of artificial exposure to people. The average dose due to anthropogenic isotopes (radioactive fallout, fuel cycle, etc.) is less than 1% of the total annual dose (Fig. 11.2). Similar average annual effective doses apply worldwide (UNSCEAR 2009). In European countries, the average annual dose from natural sources is 2 mSv in Denmark and the UK rising to 3 mSv in Finland and Sweden where indoor radon concentrations and gamma radiation are much higher (NRPA 2000).

On an individual basis, the dose would be dependent upon where one lived, one's lifestyle, and the nature and extent of any medical treatment. Most of the exposures to terrestrial gamma rays and to ^{220}Rn and ^{222}Rn decay products result from living indoors. Building materials are the main source of thoron (^{220}Rn) in room air although a minor contribution comes from soil gas. Radon contributes by far the largest variation in the average dose from natural radiation sources.

11.6.2 Radon

The average annual dose to the UK population from radon is 1.2 mSv with a range of 0.3 to more than 100 mSv. In the most radon-prone area in Great Britain, the average person receives a total annual radiation dose of 7.8 mSv of which 81% is from radon. The production of radon by the radioactive decay of uranium in rock, overburden, and soil is controlled primarily by the amount of uranium within the rock-forming minerals and their weathering products. The ^{238}U decay chain may be divided into two sections separated by ^{226}Ra (radium), which has a half-life of 1,622 years (Table 11.1). Earlier isotopes mostly have long half-lives, while the later isotopes, including radon (^{222}Rn),

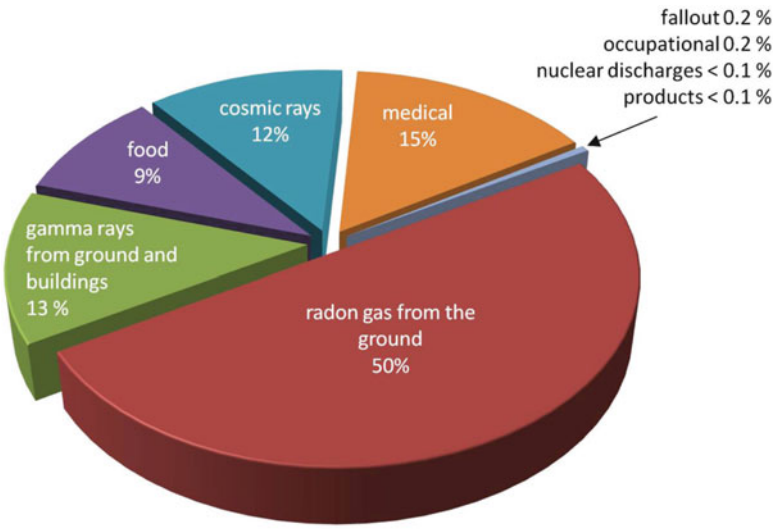


Fig. 11.2 Sources of radiation exposure contributing to average effective dose in the UK. Other = occupation 0.3%, fallout 0.2%, nuclear discharges <0.1%, and products <0.1% (Data from Watson et al. 2005)

have relatively short half-lives. Outdoors, radon normally disperses in the air whereas in confined spaces such as buildings, mines, and caves (Gillmore et al. 2000, 2001, 2002) it may accumulate. Higher radon (average 480 Bq m^{-3}) was found in houses located ±150 m from the surface projection of a closed uranium mine tunnel in Hungary compared with houses located further away (average 291 Bq m^{-3} (Somlai et al. 2006). Radon is thought to migrate through fissures that intersect the mine tunnel and run up to the surface. Radon in indoor air comes principally from soil gas derived from soils and rocks beneath a building with smaller amounts from the degassing of domestic water into the indoor air and from building materials. Soil gas represents the predominant source of indoor ^{222}Rn gas. Outdoor radon concentrations may occasionally reach potentially hazardous levels. For example, air escaping from an open uranium mine gallery in the town of Schneeberg, Germany, contained up to 10,000 Bq m^{-3} radon and ventilation facilities had to be installed to prevent ingress of this air into an adjacent factory. Radon is potentially emitted from some anthropogenic sources, such as near-surface radioactive waste disposal sites (Appleton et al. 2011a). Radon exhalation and releases from uranium mining and milling has the potential to increase long-term radon releases into the adjacent environment relative to pre-mining baseline concentrations (Mudd 2008). Precipitation and soil moisture can influence radon flux densities from uranium mining waste rock dumps, ore stockpiles and where effluent enriched in ^{226}Ra has been spray irrigated over land in wet-dry tropical regions (Lawrence et al. 2009).

Water in rivers and reservoirs usually contains very little radon so homes that use surface water do not have a radon problem from their water. Water processing in large municipal systems aerates the water, which allows radon to escape and also delays the use of water until most of the remaining radon has decayed.

Radon dissolved in groundwater can migrate over long distances along fractures and caverns depending on the velocity of fluid flow. Mains water supplies pass through treatments which tend to remove radon gas but small public water works and private domestic wells often have closed systems and short transit times that do not remove radon from the water or permit it to decay. In such situations, radon from the domestic water released during showering and other household activities could add radon to the indoor air. Areas most likely to have problems with radon from domestic water supplies include those with high levels of uranium in the underlying rocks. This association has been observed in the USA, the UK and Sweden (Åkerblom et al. 2005a, b; Skeppstrom and Olofsson 2006).

Radon from degassing of domestic water accounts for about 5% of the total indoor radon for homes that use groundwater sources in the United States. In some cases, radon from this source may account for a higher proportion of indoor radon. In Maine (U. S.), radon concentrations in domestic water wells sometimes exceed 37,000 Bq L^{-1} and more than 10% of private well water supplies exceed the action level (740 Bq L^{-1}). Radon from the domestic water supply is inhaled when it is released from the water during showering (Song et al. 2011), washing clothes, and washing dishes. It is estimated that 370 Bq L^{-1} of radon in the domestic water supply contributes about 37 Bq m^{-3} to the indoor air of a home. Areas most likely to have problems with radon from domestic water supplies include those with high levels of uranium in the underlying rocks, such as uraniferous granites. This association has been observed both in the United States and the UK. In Maine, for example, the average for well water in granite areas is over 500 Bq L^{-1}. Approximately 10% of the drilled wells in population-based random study of 472 private wells in Finland exceeded a radon concentration of 1,000 Bq L^{-1}

(Vesterbacka et al. 2005) whilst in Sweden, 10,000 private wells are estimated to have concentrations greater than 100 Bq L^{-1}, which is the recommended action levels for private households.

In a study of private water supplies in southwestern England, a high proportion derived from granite areas exceeds the draft European Union action level of 1,000 Bq L^{-1}. It was also found that radon concentrations varied significantly over the course of a week and between samples taken several months apart. For water from groundwater sources, mean values (by source type) at the tap were generally lower than those at the source. This is consistent with loss of radon due to degassing as a result of water turbulence within the supply system and natural radioactive decay while the water is resident in the household supply system. All the water sources sampled showed large variability in radon concentration over the summer sampling period, whereas less pronounced variability was observed during the winter sampling. Maximum values were observed during the summer.

In SW Poland there is a correlation between radon potential and radon in groundwater where high concentrations are associated with enhanced levels of ^{226}Ra, high emanation coefficients, intense rock fracturing and intensive weathering of bedrocks (Przylibski et al. 2004). In the Extremadura region of Spain, the highest radon in groundwater (>1,000 Bq L^{-1}) occurs in areas of granitic and Precambian igneous bedrock, where there is a positive correlation between ^{222}Rn in groundwater and natural gamma radiation (Galán López and Martín Sánchez 2008).

Building materials generally contribute only a small percentage of the indoor air ^{222}Rn concentrations, although this may be 20–50% of the radon in an average UK dwelling (Gunby et al. 1993). Groves-Kirkby et al. (2008) review this topic and noted that in areas of generally low indoor radon concentrations, indoor radon may be mainly derived by emanation from building materials, with an estimated 20, 10–70, and 20 Bq m^{-3} of indoor radon being derived from this source in Belgium, Germany and the Netherlands respectively. However, in some areas, concrete, blocks, or wallboard made using radioactive shale or waste products from uranium mining will make a larger contribution to the indoor radon. High radium content and radon exhalation rates in concrete and bricks used in some Hong Kong buildings with high indoor radon concentrations are probably caused by the granitic composition of aggregates (Man and Yeung 1998). In Sweden, 300,000 houses constructed with radioactive alum shale form the world's largest stock of buildings that have used building materials with enhanced radiation (Åkerblom et al. 2005a, b). The houses have radon concentrations of 100–400 Bq m^{-3} and gamma radiation levels of 0.3–1.2 μSvh^{-1} (NRPA 2000). High effective dose rates (7.1–16.7 mSvy^{-1}) for ^{222}Rn and ^{220}Rn and their progenies have been estimated for cave dwellings excavated in loess in the Yan'an area of China, reflecting both exhalation from the loess and relatively poor ventilation (Wiegand et al. 2000). More recent radon and thoron measurements in loess cave dwellings in Gansu province, China revealed arithmetic mean radon and thoron concentrations of 91 and 351 Bq m^{-3}, respectively producing an effective dose of 3.08 mSv/year, significantly lower than earlier estimates (Yamada et al. 2006). Geology, coating level and material, decorating material and ventilation were concluded to be the main factors affecting radon concentration in underground buildings in China (Li et al. 2006). Tunnels built in granite, tuff, quartz sandstone and limestone generally have higher radon concentrations in summer compared with winter, although the reverse was observed in andesitic porphyry tunnels. However, the difference was statistically significant only in limestone tunnels (Li et al. 2010). Granite derived soil used for building materials in Guangdong Province, China, has high ^{238}U (140 Bq kg^{-1}) and ^{226}Ra (134 Bq kg^{-1}) producing radium equivalent activities much higher than the national limit value for building material (Song et al. 2012).

11.6.3 Gamma Rays from the Ground and Buildings (Terrestrial Gamma Rays)

Everyone is irradiated by gamma rays emitted by the radioactive materials in the Earth. Building materials extracted from the Earth may also emit gamma radiation, so people can be irradiated indoors as well as outdoors. Terrestrial gamma rays originate principally from the radioactive decay of the natural potassium, uranium, and thorium. These elements are widely distributed in terrestrial materials which include rocks and soils. The average annual gamma radiation dose from all these sources to the population in Great Britain is about 350 μSv with a range of 120–1,200 μSv.

In general, the gamma radiation dose is proportional to the amount of U, Th, and K in the ground and in building materials. Within a masonry building, most of the gamma radiation is received from the building materials, whereas in wooden buildings a larger part of the dose is contributed from gamma radiation from the ground. The average person in the UK spends only 8% of his or her time outdoors so the contribution to total radiation dose from the ground is relatively small. The bulk of the radiation above the ground surface is derived from only the top 30 cm or so of soil or rock. Soils developed upon radioactive rocks generally have a much lower gamma radioactivity than the rock substrate. Whereas one can predict or identify areas of high geological gamma radioactivity, the resultant dose to the population depends on additional factors such as soil type, house construction, and lifestyle.

A single K isotope, ^{40}K, comprising only 0.0119% of the total K, is radioactive. ^{40}K undergoes branched decay producing ^{40}Ca and ^{40}Ar, the latter reaction producing high-energy gamma rays. Uranium consists of two main isotopes (^{238}U and ^{235}U). Because ^{235}U comprises only 0.72% of the total U it may, for practical purposes, be ignored. Isotopes in the later section of the ^{238}U decay series (Table 11.1) include the bismuth isotope (^{214}Bi), which contributes most of the gamma activity of the decay series. As ^{226}Ra is chemically very different from U, it is possible in natural processes for the two to become separated so that the ^{226}Ra, and its daughter products are unsupported by the parent U. In addition, radioactive elements in the rock fragments and derived minerals in the weathered overburden are diluted with organic matter and water. Thus there may not be a simple relationship between the measured gamma ray flux and the U content.

Many of the daughter isotopes of ^{238}U and ^{232}Th are gamma active. ^{208}Tl is the main gamma active daughter product derived from ^{232}Th, and this takes 70 years to reach secular equilibrium in the ^{232}Th decay series. Potassium gives rise to a prompt gamma ray in which the intensity is directly related to the potassium concentration. On average potassium is much more abundant than thorium which, in turn, is more abundant than uranium. However, the specific gamma activities are such that, on average, approximately equivalent gamma emissions are observed from potassium and the decay series of uranium and thorium.

Areas of high natural radiation include areas of monazite sands in Brazil, China, Egypt, and India; volcanic rocks in Brazil and Italy; uranium mineralization in France, the UK, and the United States; and radium-enriched karst soils developed over limestones in Switzerland, the UK, and the United States (UNSCEAR 2009).

11.6.4 Food and Drink

Radioactive materials even occur in food. Potassium-40 in particular is a major source of internal irradiation. Natural radioactivity in the human diet gives an average annual dose of 300 μSv each year of which 180 μSv is from ^{40}K. The range for all internal radiation sources in Great Britain is 100–1,000 μSv per annum. Shellfish concentrate radioactive materials so that, even when there is no man-made radioactivity, people who consume large quantities of mussels, cockles, or winkles can receive a dose from natural radioactivity in food that is about 50% higher than average. Apart from restricting intake of shellfish, there is very little possibility of reducing the small exposure to natural radioactivity from food.

Table 11.8 Sources of radiation in Great Britain for average person in the UK (NRPB 1989)

Source	% Annual dose
Natural sources	87.0
Radon (^{222}Rn) gas from the ground	47.0
Thoron (^{220}Rn) from the ground	4.0
Gamma rays from the ground and building materials	14.0
Food and drink	12.0
Cosmic rays	10.0
Artificial sources	13.0
Medical	12.0
Nuclear discharges	0.1
Work	0.2
Fallout	0.4
Miscellaneous	0.4

11.6.5 Cosmic Rays

Approximately 10% of the average annual radiation dose is from cosmic rays (Table 11.8), although this increases with latitude and altitude. The average dose from cosmic radiation received each hour rises from 0.03 μSv at sea level, to 0.1 μSv in Mexico City (altitude of 2,250 m), and 5 μSv at the cruising altitude for commercial jet aircraft (10,000 m). Polar and mountain dwellers, aircrews, and frequent air travelers therefore receive higher doses of cosmic radiation. Little can be done about cosmic radiation because it readily penetrates ordinary buildings and aircraft. The average annual dose from cosmic rays in Great Britain is 250 μSv, with a range of 200–300 μSv.

11.7 Measurement of Radon

11.7.1 Radon Testing in the Home

Although radon cannot be seen or smelled, it can be measured relatively easily with the proper equipment. The most common procedures for measuring radon make use of the fact that it is the only natural gas that emits alpha particles, so if a gas is separated from associated solid and liquid phases any measurements of its radioactive properties relate to radon or its daughter products. In the United States, radon in homes is usually measured using inexpensive do-it-yourself radon test kits, which are available by mail order and in many retail outlets or by hiring a U. S. EPA qualified or state-certified radon tester. The EPA recommends that all homes be tested for radon below the third floor.

Common short-term test devices are charcoal canisters, alpha track detectors, liquid scintillation detectors, electret ion chambers, and continuous monitors. A short-term testing

device remains in the home for 2–90 days, depending on the type of device. Because radon levels tend to vary from day-to-day and season-to-season, a long-term test is more likely than a short-term test to measure the home's year-round average radon level. If results are needed quickly, however, a short-term test followed by a second short-term test may be used to determine the severity of the radon problem. Long-term test devices, comparable in cost to devices for short-term testing, remain in the home for more than 3 months. A long-term test is more likely to indicate the home's year-round average radon level than a short-term test. Alpha track detectors and electret ion detectors are the most common long-term test devices.

Charcoal canister and liquid scintillation detectors contain small quantities of activated charcoal. Radon and its decay products are adsorbed onto the charcoal and are measured by counting with a sodium iodide detector or a liquid scintillation counter. Radon adsorbed at the beginning of the exposure decays away after a few days so the duration of the measurement is restricted and the device does not measure the true average exposure. Charcoal detectors are suitable only for short-term tests when results are required urgently and a less accurate measurement is acceptable. The result should be well below the action level before it can be concluded that the annual average concentration will also be below the action level. Ambiguous short-term measurements should be followed up by a long-term measurement (NRPB 2000).

Alpha (etched) track detectors contain a small sheet of plastic that is exposed for a period of 1–3 months. Alpha particles etch the plastic as they strike it. These marks are then chemically treated and are usually counted automatically under a microscope to determine the radon concentration. Etched track detectors are relatively cheap and suitable for long-term measurement and are usually deployed for a period of 3 months.

Electret ion detectors contain an electrostatically charged Teflon disk. Ions generated by the decay of radon strike and reduce the surface voltage of the disk. By measuring the voltage reduction, the radon concentration can be calculated. Allowance must be made for ionization caused by natural background radiation. Different types of electret are available for measurements over periods of a few days to a few months. The detectors must be handled carefully for accurate results.

Continuous monitors are active devices that need power to function. They require operation by trained testers and work by continuously measuring and recording the amount of radon in the home. These devices sample the air continuously and measure either radon or its decay products (NRPB 2000).

A rigorous procedure must be followed for short-term tests if relatively reliable results are to be obtained. For example, doors and windows must be closed 12 h prior to testing and throughout the testing period. The test should not be conducted during unusually severe storms or periods of unusually high winds. The test kit is normally placed in the lowest lived-in level of the home, at least 50 cm above the floor, in a room that is used regularly, but not in the kitchen or bathroom where high humidity or the operation of an exhaust fan could affect the validity of the test. At the end of the test period, the kit is mailed to a laboratory for analysis; results are mailed back in a few weeks. If the result of the short-term test exceeds 100 Bq m^{-3} then a long-term test is normally recommended.

Remediation of the home is recommended if the radon concentration exceeds certain levels (150 Bq m^{-3} in Luxembourg and the United States; 200 Bq m^{-3} in Australia, Israel, Syria, and the UK; 400 Bq m^{-3} in Austria, Belgium, Denmark, Finland, Greece, and Sweden).

Radon levels are generally highest in winter e.g., Bossew and Lettner (2007) so seasonal corrections are usually applied to estimate the average annual radon level. In workplaces, consideration needs to be taken of work practices and the building design and use. For small premises at least one measurement should made in the two most frequently occupied ground floor rooms. In larger buildings at least one measurement is required for every 100 m^2 floor area. Regional variation of seasonal correction factors have been observed in Ireland and some regions in the UK have seasonal correction factors which are different from national factors (Burke and Murphy 2011; Denman et al. 2007). Different seasonal variations are sometimes observed in building sites on different bedrocks (Moreno et al. 2008). In the UK, this problem has been resolved by applying temperature correction factors to indoor radon measurements because the difference between the indoor and outdoor temperature has a major impact on the rate of entry of radon into dwellings (Miles 1998). The temperature correction factors are derived from the difference between outdoor and indoor temperature for each specific measurement period (Miles 1998, 2001).

11.7.2 Indoor Radon Measurement Validation Scheme

Great care is required in the interpretation of radon monitoring results because the rate at which radon is released into buildings is controlled by a complex series of factors, which requires monitoring equipment to be located in the right place over a prolonged period to take account of temporal variations. Validation schemes are required to (1) ensure organizations measure radon within an acceptable degree of uncertainty; (2) determine that detectors are handled in an appropriate fashion both before and after the detectors have been with householders; and (3) ensure minimum standards in how results are interpreted and presented,

which includes requiring the use of the seasonal variation factors. In the United States, the EPA operates a voluntary National Radon Proficiency Program that evaluates radon measurement companies and the test services they offer. Both the UK HPA and the U. S. EPA recommend that testing services be purchased from certified organizations.

11.7.3 Measurement of Radon in Soil Gas

Measurement of radon in soil gas using pumped monitors is recommended as the most effective method for assessing the radon potential of underlying rocks, overburden, and soil. Instruments for the determination of soil gas radon are generally based upon either an extraction method, using a "pump monitor" device for transferring a sample of the soil gas to a detector, or simply emplacing the detector in the ground (passive methods). In the former method, a thin rigid tapered hollow tube is usually hammered into the ground to a convenient depth, which causes minimum disturbance to the soil profile. Detection of radon is usually based upon the zinc sulfide scintillation method or the ionization chamber. Alpha particles produce pulses of light when they interact with zinc sulfide coated on the inside of a plastic or metal cup or a glass flask (Lucas cell). These may be counted using a photomultiplier and suitable counting circuitry. Because the radon isotopes are the only alpha-emitting gases, their concentration may be determined accurately using relatively simple equipment. Because of the different half-lives of these isotopes and their immediate daughter products, it is possible to calculate the activities of radon and thoron. The equipment is relatively robust for field use and is designed for rapid changing of the cell when it becomes contaminated. The large number of instruments produced attests to its suitability for field use. The concentration of radon in soil gases is usually sufficient that the level may be determined relatively fast; a matter of a few minutes generally suffices.

Radon can also be measured by emplacing alpha track detectors in the ground. Holes may be dug with an auger or drill to a depth of at least 0.5 m and preferably 1.0 m. Holes are normally lined with plastic piping in which the detector is emplaced and the top of the pipe sealed. The detectors are normally taped to the bottom of a plastic cup, which is inverted before burial. The detector is then recovered 3–4 weeks later. This procedure is used when long-term monitoring is required to overcome problems of short-term variation in radon concentration.

Although alpha track detectors overcome many of the problems associated with temporal variation in radon fluxes, they are time-consuming to emplace, requiring two visits to each site, with all the problems in reoccupying the site. More important, they require a laboratory processing stage. In practice they are generally not favored for primary investigations, although they do have an important role to play at later stages. They are also sensitive to thoron, but the presence of a polyethylene film seal, at a distance of about 5 cm from the detector, reduces the amount of short half-lived thoron while having a negligible effect on radon. The polyethylene film allows radon but not water vapor to diffuse. Water droplets on the surface of the film may also affect the recorded alpha counts and water vapor absorbers may need to be introduced into the sampling device.

Papastefanou (2007) reviews the instrumentation available for measuring radon in soil gas and groundwaters.

Determination of radon potential from soil gas radon concentrations generally produces better results when soil permeability is also measured (Barnet et al. 2008; Kemski et al. 2001).

11.7.4 Measurement of Radon in Water

Radon has a high partition coefficient (gas to water) so that the passage of fine gas bubbles through water provides an efficient means of extraction. The gas may be drawn into an evacuated Lucas cell. Alternatively a sealed re-circulating system may be set up. Very careful attention must be paid to the timing of both degassing and counting and careful calibration of the procedure with standardized radon solutions is required. Other methods require expensive equipment and laboratory processing. For example, the radon daughter ^{214}Bi photopeak emission can be measured using either a sodium iodide scintillation crystal or a high-resolution, lithium-drifted germanium semiconductor detector.

11.7.5 Measurement of Radon in Solid Materials

One of the solid daughter products of radon is ^{214}Bi. This emits high-energy gamma radiation at 1.76 MeV. Gamma spectrometric determinations of uranium in the field and laboratory often make use of this photopeak on the assumption that the decay chain is in equilibrium and therefore this measurement provides an effective total radon determination. If the parent uranium mineral is resistant to weathering (e.g., thorium- and REE-rich uranium oxides and silicates, monazites, zircons, etc.) then the radium will tend to be in secular equilibrium with the uranium. In such minerals the radon loss is normally low and gamma spectrometric measurements give a good indication of the uranium contents. The measurement of radon release from solid samples requires an alternative method. Radon release from disaggregated samples (soils, stream sediments, and unconsolidated aquifer sands) may be determined by agitating a slurry of the material with distilled water in a

sealed glass container, allowing a period of about 20–30 days for the generation of radon from radium, and then measuring the radon in the aqueous phase using a liquid scintillation counter. Emanation of radon from solid rock samples can be determined using a similar method.

11.8 Factors Controlling Release and Transfer of Radon Gas

Most radon remains in rocks and soil and only some of that near a free surface is released. Soil generally releases more radon than rock, as its constituents are more comminuted. The rate of release of radon from rocks and soils is largely controlled by the uranium concentration and by the types of minerals in which the uranium occurs. Once radon gas is released from minerals, the most important factors controlling its migration and accumulation in buildings include (1) transmission characteristics of the bedrock including porosity and permeability; (2) the nature of the carrier fluids, including carbon dioxide gas, surface water, and groundwater; (3) weather; (4) soil characteristics including permeability; (5) house construction characteristics; and (6) lifestyle of house occupants.

11.8.1 Mineralogical Effects

The main mineralogical factors affecting the release of radon are the solubility, internal structure, and specific surface area of uranium-bearing minerals. Uranium is very seldom homogeneously distributed throughout rocks and soils. Most of the uranium in rocks can be attributed to discrete uranium-bearing minerals, even when there is only a few mg kg^{-1} of uranium present. Because radon is a gas with a limited half-life, its chances of escaping from the parent mineral are much greater if it is generated from grain margins. Other important controls are the openness of and imperfections in the internal structure of the mineral and the specific surface area of the mineral grains.

The release of radon is generally controlled by alpha particle recoil mechanisms, which tend to expel radon from radium derived from uranium-bearing minerals. Most of the radon remains within the mineral to decay again to solid products. Only a very small proportion of the radon generated can be released by recoil. The location of the radium atoms in the mineral grains and the direction of the recoil of the radon atoms will determine whether the newly formed radon atoms enter pore spaces between mineral grains. Factors such as the specific surface area, the shape, degree of fracturing, imperfections, and even radiation-induced damage of the host uranium-bearing mineral affect the efficiency of radon expulsion. Because uranium minerals have high densities, the recoil range is usually low. However, if radium is present in intergranular films then recoil ranges varying between 20 and 70 μm would occur. The fraction of radon, produced by radium decay, that escapes from rock or soil (called the emanation coefficient) is dependent on the surface area of the source material. Emanation coefficients are greater for rocks than minerals, whereas soils usually have the highest values. If water is present in the pore space, however, the moving radon atom slows very quickly and is more likely to stay in the pore space. However, the emanation power of soil rises with water content, probably because radon atoms ejected from soil grains have a higher probability of remaining in the pore space and migrating out of the sample if the pore space is filled with water; being twice as high as for dry soil when soil moisture is 10% (Bossew 2003). Chau et al. (2005) suggested that the radon mass exhalation rate reaches a maximum at about 6% by weight, after which it decreases with both increasing and decreasing water content of soil samples (Chau et al. 2005). The radon emanation fraction is reported to increase markedly with grain size and reaches a constant value of 50% when moisture content is 0% and the radium is uniformly distributed on the grain surface (Sakoda et al. 2010). A dramatic increase in radon emanation occurs at smaller grain sizes with increasing moisture content because radon emanation depends greatly on the pore size between a Ra-bearing grain and a neighboring grain (Sakoda et al. 2010) Differences in the uranium-bearing minerals, and especially in the solubility of the major uranium-bearing minerals, control the amount of radon released. In some granites, for example, much of the uranium is found in the mineral uraninite (uranium oxide), which is easily weathered, especially near the surface. Uranium is more soluble in water so it is removed from the original mineral site, but the relatively insoluble radium, which is the immediate parent of radon gas, remains in a mixture of iron oxides and clay minerals. This material is a highly efficient radon generator because of the high specific surface area of the radium-bearing phase (Akerblom and Lindgren 1997). Radon generated from the radium can escape into the fluid phase with high efficiency thus facilitating its rapid migration to the surface. An inert gas, radon is relatively unaffected by chemical buffering reactions that often control the generation of other gases in rocks and their weathering products. In contrast, uranium in other granites may occur in chemically resistant high-thorium uraninite, zircon, monazite, and apatite, all of which liberate less radon. The mineral associations typically found in sedimentary rocks differ significantly from those in granites. In carboniferous limestone of northern England, for example, uranium is relatively uniformly distributed and associated with finely divided organic matter in the matrix of bioclastic limestones (usually <10 mg kg^{-1}U), although it may also be concentrated in stylolites, which typically

contain 20–60 mg kg^{-1}U. Even though the overall concentration of U in the limestones is below 2 mg kg^{-1}, high radon emissions are probably derived from radium deposited on the surfaces of fractures and cavities. The high specific surface area of the radium permits efficient release of radon and high migration rates are promoted by the high permeability of the limestone. In addition, uranium and radium are concentrated in residual soil overlying limestone. Radium is sometimes preferentially concentrated in soil organic material, which has a high emanation coefficient (Greeman and Rose 1996). In black shales in the UK, uranium is located mainly in the fine-grained mud matrix, where it may be present at levels up to 20 mg kg^{-1}U, and also in organic-rich bands at concentrations up to 40 mg kg^{-1}U. Much higher uranium concentrations have been reported from the Chattanooga shale in the United States (20–80 mg kg^{-1}), the Dictyonema shale in Estonia (30–300 mg kg^{-1}), and the alum shale in Sweden and Norway (50–400 mg kg^{-1}). Uranium is rare in detrital phases and may also be remobilized and adsorbed on iron oxides. In sandstones, uranium is concentrated in primary detrital minerals, such as apatite and zircon, which can contain high concentrations of U (>100 mg kg^{-1}). Uranium may also be adsorbed onto Fe oxides in the matrix of sandstone or its weathering products. Emission of radon from sandstones is restricted by the relatively low specific surface area of the uranium minerals and appears to be more dependent upon fracturing of the rock.

11.8.2 Transmission Characteristics of Bedrock

Although the generation of high levels of radon is ultimately dependent upon the concentration of uranium and upon the nature of the parent mineral, the transmission of radon gas to the surface is largely independent of these characteristics (Åkerblom and Mellander 1997). Once the radon is released from the parent mineral into the space between mineral grains (the intergranular region) other factors take over. The most notable of these are (1) the fluid transmission characteristics of the rock including permeability, porosity, pore size distribution, and the nature of any fractures and disaggregation features and (2) the degree of water retention (saturation) of the rocks. Faults and other fractures permit the efficient transmission of radon gas to the surface. The presence of faults with their enhanced fluid flow frequently results in high radon in soil gases (Ball et al. 1991).

11.8.3 Carrier Fluids

Radon readily diffuses into pores and cavities from mineral surfaces. However, its relatively short half-life (3.8 days) limits the distance over which diffusion may occur. In highly permeable dry gravel, radon has decayed to 10% of its original concentration over a diffusion length of 5 m (UNSCEAR 2009). In more normal soils, which are generally moist, this distance would be substantially less. Diffusive ^{222}Rn in soil gas can be determined from the specific ^{226}Ra activity, specific density, effective porosity, and radon emanation coefficients of soils and rocks (Washington and Rose 1992). In caves, radon concentrations of approximately 100 Bq m^{-3} would be expected if radon were generated by diffusion from solid limestone with 2.2 mg kg^{-1}U. However, the enhanced concentration of radon in caves suggests that structurally controlled convective transport of radon in fluids along faults, shear zones, caverns, or fractures is more significant than diffusive transport. Transportation of radon in this way may exceed 100 m.

Following radon release, migration in carrier fluids, such as carbon dioxide gas or water, is considered to be the dominant means of gas transmission to the surface. Radon release is higher in rocks that have a high surface area in contact with groundwater. Once released from the uranium minerals, radon migration is dependent upon the fluid flow characteristics of the rock and soil. In water, convective or pressure gravity flow mechanisms can influence migration of the radon, whereas in a gas the transport may be controlled by the diffusion characteristics of the carrier gas. Water flow below the water table is generally relatively slow as is ground-water transport in the soil aquifer (<1–10 cm per day). Thus all hydraulically transported radon will have decayed over a distance of less than 1–2 m. Radon is likely to be carried away more quickly by fluids in areas of permeable rocks such as limestones. The carrier effect may also be important for other rock types. Carbon dioxide may collect radon gas in the unsaturated zone and transport it along fractures, fissures, and faults. In situations where the carbon dioxide flux is high, such as in active volcanic areas, radon may be either diluted or enhanced because of rapid transport from the generation zone to the surface.

Radon in surface water is not generally accompanied by dissolved radium. In surface stream waters, the radon concentration appears to be more closely related to the radium concentration of the stream sediment. However, the radon concentration in surface streams is usually far too low for more than a very small degree of transfer across the air/water interface to occur, unless a gas phase is introduced. Radon dissolved in subsurface fluids migrates over long distances along fractures and caverns depending on the velocity of fluid flow. Radon is soluble in water and may thus be transported for distances of up to 5 km in streams flowing underground in limestone. Radon remains in solution in the water until a gas phase is introduced, for example, by turbulence or by pressure release. If emitted directly into the gas phase, as may happen above the water table, the presence of a carrier gas, such as carbon dioxide, would tend to induce

migration of the radon. This appears to be the case in certain limestone formations, where underground caves and fissures enable the rapid transfer of the gas phase.

11.8.4 Weather

The principal climatic factors affecting radon concentrations are barometric pressure, rainfall, and wind velocity. In the absence of a less permeable humic or clay rich topsoil, radon concentration in soil gas varies directly with barometric pressure, and to a lesser extent, inversely with wind speed. Where the topsoil is finer grained and more humic, the effects of barometric pressure and wind velocity are reduced to a marginal role. The extent to which rainfall affects radon concentrations depends on the permeability of the soil. For permeable soils, radon concentrations are only affected during precipitation when saturation of small pore spaces with moisture effectively prohibits the rapid outgassing of radon from the soil. This causes the buildup of radon below the moisture-saturated surface layer and increases of an order of magnitude are sometimes observed. Prolonged rainfall may penetrate deeply and seal the pore spaces in the soils to a considerable depth.

A similar buildup of radon is often observed during the night when dew forms on the surface and this can result in a twofold increase in soil gas alpha activity. Sealing of the pore spaces by near-surface moisture can result in temporary entrapment of radon in soil gases, with a significant increase in total gamma activity from radon daughters (^{214}Bi). Dry conditions cause clay-rich soils to dry out and to fracture, allowing easier egress for the soil gases and hence an increase in radon activity at the soil surface. Seasonal variation in soil pore radon concentration was observed by Rose et al. (1990) who found that the radon concentration tended to be lower in the winter and higher in the summer, often varying by a factor of 3–10. The variation was attributed largely to changes in the soil moisture content with more radon held in solution in the soil pore water during the winter. The variation is greater in the soil above 70 cm depth than below this depth, presumably due to greater short-term fluctuations in soil moisture content. This suggests that radon in soil gas measurements should be taken at depths greater than 70 cm in order to reduce the effects of temporal variations caused by rainfall.

Although barometric pressure and rainfall obviously cause temporal variation in radon concentration (indicated by alpha activity), it is encouraging to note that the soil gas radon fluxes in areas that are not mineralized appear to be relatively uniform. Various rock types have been tested and the site variation is often less than that between adjacent rock types. This is particularly important for radon potential mapping based on the measurement of radon in soil gas.

11.8.5 Soil Characteristics

The principal soil properties that influence the concentration of radon in soil gas, including the rate of release of radon and its transfer through soils, are soil permeability and soil moisture. In general soil permeability depends on such factors as soil texture, structure, median pore diameter, pore size distribution, pore volume, packing density, soil bulk density, and grain size. Soil mineralogy is an important factor controlling soil gas radon concentrations; in some cases, organically bound ^{226}Ra can be a principal source of ^{222}Rn in soil gas (Greeman et al. 1990; Greeman and Rose 1996). Radon volume activity increases with the percentage of coarse material in the soil thus confirming the general correlation between radon fluxes and soil permeability. In general, coarse gravelly soils will tend to have higher radon fluxes than impermeable clay soils. However, humic and clay soils may be impermeable in the winter when saturated with water or filled with ice if the ground is frozen, and during very dry periods they may crack and behave in a permeable manner. Soil permeability and rainfall (soil saturation) exert a considerable control on radon concentrations in houses. Soil permeability generally closely reflects the permeability of the underlying rocks and superficial deposits such as glacial till, alluvium, or gravel. Radon diffuses more slowly through water than air, so water-saturated soils impede the diffusion of radon enough for it to decay to harmless levels before it has diffused more than 5–10 cm. Consequently, radon from water-saturated soils is unlikely to enter buildings unless it is transported in other gases such as carbon dioxide or methane.

It is important to remember that whereas the top meter of the soil profile is generally removed during the construction of foundations for a dwelling, only a few centimeters of topsoil are removed from the remainder of the subfloor space. Indeed, in many cases the soil profile beneath a dwelling will not be unduly influenced by temporal variations in rainfall. The influence of the geochemistry and permeability of the bedrock or overburden beneath a dwelling on the potential for radon emissions from the ground may be greater than near-surface soil properties.

Only 10–50% of the radon produced in most soils escapes from the mineral grains and enters the pores. Soils in the United States generally contain between 5 and 80 Bq L^{-1} of radon. Drier, highly permeable soils and bedrock—such as limestones, coarse glacial deposits, and fractured or cavernous bedrock, and hill slopes—are usually associated with relatively high levels of indoor radon. The permeability of the ground permits radon-bearing air to move greater distances before it decays, and thus contributes to high indoor radon even if the radon content of soil gas is in the normal range (5–50 Bq L^{-1}).

11.9 Radon Migration Pathways

When considering *natural migration pathways*, it should be noted that although the general direction and position of planar discontinuities and openings including bedding planes, joints, shear zones, and faults can be determined by detailed structural mapping, the precise location of such migration pathways is often difficult to establish, especially if the area is covered with soil or drift. In the United States, high radon is associated with U-enriched shear zones in granites, which are characterized by high radon in soil gas and groundwater. Indeed, some of the highest indoor radon levels in the United States are associated with sheared fault zones. Similar observations have been made in southwest England. Radon and other gases are known to concentrate and migrate upward along faults and through caves and other solution cavities. However, natural cavities such as potholes and swallow holes in limestone would also be difficult to locate precisely due to their irregular and relatively unpredictable disposition.

Radon has been used to identify the location of faults and frequently reaches a maximum in the direct vicinity of faults (Barnet et al. 2008; Ielsch et al. 2010; Neri et al. 2011; Pereira et al. 2010; Swakon et al. 2005). However, a consistent decrease of indoor radon away from mapped faults is not observed in the UK (Appleton 2004) and it is likely that elevated radon concentrations will be associated mainly with active faults. Radon has also been used for earthquake and vulcanological investigations (Avino et al. 1999; Ciotoli et al. 1998; Ghosh et al. 2009; Hishinuma et al. 1999; Papastefanou 2010)

Artificial pathways underground include mine workings and disused tunnels and shafts. Radon concentrations in old uranium and other mine workings are commonly 10,000–60,000 Bq m^{-3} and can be as high as 7,100,000 Bq m^{-3} even when uranium is a very minor component of the metalliferous veins (Gilmore et al. 2001). High radon is known to be associated with gassy ground overlying coal-bearing rock strata. In addition, relatively randomly orientated and distributed blasting and subsidence fractures will affect areas underlain by mined strata. The sites and disposition of recent coal mine workings in some countries may be obtained from mine records, although these may not be reliable. Other artificial pathways related to near-surface installations include electricity, gas, water, sewage, and telecommunications services, the location of which may be obtained from the local service agencies. Land drains provide another potential migration pathway. The detection and prediction of migration pathways is difficult and may be imprecise, although a detailed geological and historical assessment together with appropriate radon gas monitoring and a detailed site investigation should provide a reasonable assessment of the source and radon gas migration pathways. Information on the local geology may be obtained from maps, memoirs, boreholes, and site investigation records.

11.10 Factors Affecting Radon in Buildings

The design, construction, and ventilation of the home affect indoor radon levels. Radon can enter a home through cracks in solid floors and walls below construction level; through gaps in suspended concrete and timber floors and around service pipes; and through crawl spaces, cavities in walls, construction joints, and small cracks or pores in hollow-block walls (Fig. 11.3). Radon concentrations are generally highest in basements and ground floor rooms that are in contact with the soil or bedrock. Air released by well water during showering and other household activities may also contribute to indoor radon levels, although this generally makes a relatively small contribution to the total radon level.

When constructing a house with a basement in the United States, a hole is dug, footings are set, and coarse gravel is usually laid down as a base for the basement slab. The gap between the basement walls and the ground outside is backfilled with material that often is more permeable than the original ground. Radon moves into this permeable material and the gravel bed underneath the slab from the surrounding soil. The backfill material is typically rocks and soil from the foundation site but may be imported material with different radon emanation characteristics to the local rocks and soils. Therefore, the amount of radon in the permeable material depends on the amount of uranium in the local or imported rock as well as the type, permeability, and moisture content of the soil. The backfill layer will need to have a thickness of at least 1 m or have a very high Ra^{226} concentration for it to be a significant source of radon in the building.

In a typical masonry building in which radon occurs at the UK national average level of 20 Bq m^{-3}, approximately 60% of radon comes from the ground on which the building stands, 25% from building materials, 12% from fresh air, 2% from the water supply, and 1% from the gas supply. These figures apply to the average house in the UK, but can vary substantially, and the proportion of radon entering a home from the ground will normally be much higher in homes with high radon levels. The dominant mechanism of radon ingress is pressure-induced flow through cracks and holes in the floor. Slightly negative pressure differences between indoor and outdoor atmospheres caused by wind outside and heating inside the building draw radon contaminated air into the building, especially through the floor. Energy-conserving measures such as double-glazing restrict the fresh supply of air and lessen the dilution of radon indoors. Conversely, they may also reduce the pressure difference

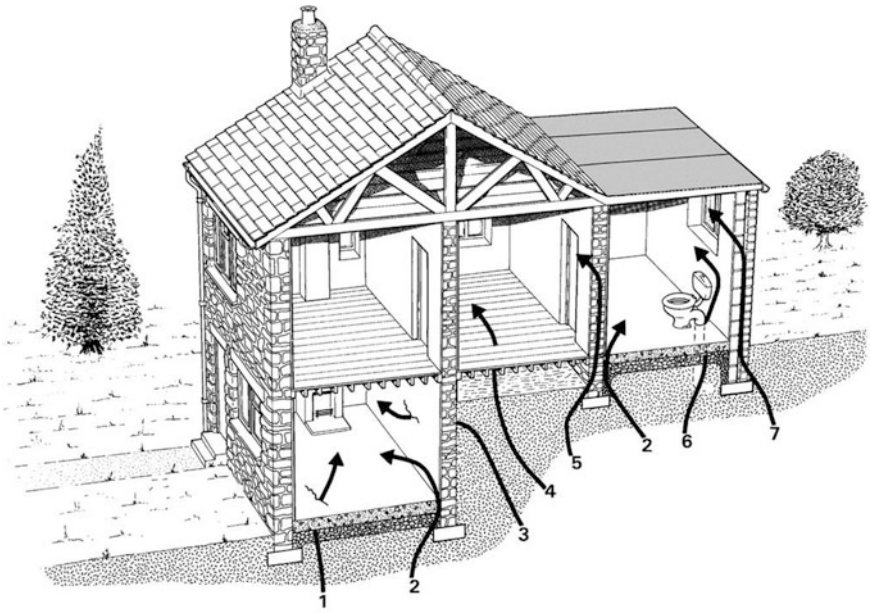


Fig. 11.3 Routes by which radon enters a dwelling (Reproduced from BR211, BRE 1999 with permission from C. Scivyer, BRE

between indoors and outdoors and thus reduce the influx of radon from the ground. Poor ventilation may increase radon concentrations, but it is not the fundamental cause of high indoor radon levels. Household energy efficiency improvements that decrease ventilation (e.g. better sealed windows and doors) could lead to an increase in exposure to radon.

Indoor radon concentrations are generally about 1,000 times lower than radon in the soil underlying the house. Most houses draw less than one percent of their indoor air from the soil with the remainder from outdoors where the air is generally quite low in radon. In contrast, houses with low indoor air pressures, poorly sealed foundations, and several entry points for soil air may draw as much as 20% of their indoor air from the soil. Consequently, radon levels inside the house may be very high even in situations where the soil air has only moderate amounts of radon.

Clavensjö and Åkerblom (1994) suggested that the ^{222}Rn concentration in a building or room that results from the transport of soil air may be calculated by the following formula:

$$C_{building} = \frac{C_t \cdot L}{V_{building} \cdot (n + \lambda)}$$

$C_{building}$ = ^{222}Rn concentration in the building/room (Bq m^{-3})
C_t = ^{222}Rn concentration in the soil air entering from the ground (Bq m^{-3})
L = volume soil air entering from the ground (m^3 h^{-1})
n = rate of air change in the building (air changes h^{-1})
λ = decay constant, for ^{222}Rn, 7.55 $\times$ 10^{-3} (h^{-1})
$V_{building}$ = building/room volume (m^3)

In high-rise residential buildings without basements in which most people are living in apartments, the contribution to indoor radon from the ground may be expected to be insignificant. In Hong Kong, where the geometric mean indoor radon concentration for residential buildings is quite high (48 Bq m^{-3}), the relatively small contribution from soil gas may be redistributed through the high-rise residential buildings by centralized air conditioning systems (Mui et al. 2008), although the main source of indoor radon in Hong Kong is likely to be concrete containing granite aggregate.

11.11 Radon Potential Mapping Methods

Accurate mapping of radon-prone areas helps to ensure that the health of occupants of new and existing dwellings and workplaces is adequately protected. Radon potential maps have important applications, particularly in the control of radon through planning, building control, and environmental health legislation. Radon potential maps can be used (1) to assess whether radon protective (preventive) measures may be required in new buildings; (2) for the cost-effective targeting of radon monitoring in existing dwellings and workplaces to detect those buildings with radon above reference levels which need to be remediated; (3) to allow measurement campaigns and public awareness to be targeted on areas at greatest risk and (4) to provide radon risk assessments for homebuyers and sellers. The radon map data helps local authorities to communicate quickly and effectively with existing homeowners or developers

planning to build new homes in high radon areas. In the UK, this is achieved by the online reports services developed by BGS and the HPA: (i) for existing homes with a valid postcode, a Radon Risk Report can be ordered online from the UK Radon website (http://www.ukradon.org/); (ii) for existing large homes and other large buildings and plots of land, a GeoReport can be obtained from BGS (http://shop.bgs.ac.uk/georeports/). It is important, however, to realize that radon levels often vary widely between adjacent buildings due to differences in the radon potential of the underlying ground as well as differences in construction style and use. Whereas a radon potential map can indicate the relative radon risk for a building in a particular locality, it cannot predict the radon risk for an individual building. In the UK, radon potential maps generally indicate the probability that new or existing houses will exceed a radon reference level, which in the UK is called the action level (200 Bq m^{-3}) (Miles and Appleton 2005; Miles et al. 2007), In other countries, geological radon potential maps predict the average indoor radon concentration (United States) or give a more qualitative indication of radon risk (Germany and the Czech Republic).

Two main procedures have been used for mapping radon-prone areas. The first uses radon measurements in existing dwellings to map the variation of radon potential between administrative or postal districts or grid squares. The second is geological radon potential mapping in which each geological feature is assigned to a radon potential class based on the interpretation of one or more of the following types of data: (1) radon concentrations in dwellings (indoor radon), (2) concentration, mineralogical occurrence, and chemical state of uranium and radium in the ground (radiometric and geochemical data), (3) rock and soil permeability and moisture content, (4) concentration of radon in soil gas, and (5) building architecture (construction characteristics). Because the purpose of maps of radon-prone areas is to indicate radon levels in buildings, maps based on actual measurements of radon in buildings are generally preferable to those based on other data.

Procedures for monitoring and surveys of radon in dwellings are described in Nazaroff (1988) and Miles (2001). In the UK, measurements are made with passive integrating detectors over a period of 3 months whereas short-term "screening" measurements taken over a 2- to 7-day period are commonly used for mapping in the United States. Measurements carried out over less than a year should be corrected for seasonal variations. In the United States, houses with basements typically have higher indoor radon than those with slab-on-grade construction because basements tend to have more entry points for radon and a lower internal pressure relative to the soil than non-basement homes. Architecture type is one factor within the Radon Index Matrix used to estimate geological radon potential in the United States (Gundersen and Schumann 1996).

Bungalows and detached houses tend to have higher indoor radon than terraced houses or flats in the same area of the UK. Building material, double-glazing, draught-proofing, date of building, and ownership also have a significant impact on indoor radon concentrations. Radon potential mapping is sometimes based on indoor radon data that have been normalized to a mix of houses typical of the housing stock as this removes possible distortion caused by construction characteristics. Maps based on results corrected for temperature but not normalized to a standard house mix reflect such factors as the greater prevalence of detached dwellings in rural areas, and hence the higher risk of high radon levels in rural areas compared with cities where flats are usually more prevalent. Radon potential estimates based on radon levels in the actual housing stock are more appropriate for the identification of existing dwellings with high radon.

Requirements for mapping radon-prone areas using indoor radon data are similar whether the maps are made on the basis of grid squares or geological units. These requirements include (1) accurate radon measurements made using a reliable and consistent protocol, (2) centralized data holdings, (3) sufficient data evenly spread, and (4) automatic conversion of addresses to geographical coordinates. It appears that Great Britain is the only country that currently meets all of these requirements for large areas (Miles and Appleton 2005). In countries where lesser quality or quantity of indoor radon data are available, there is greater reliance on proxy data for radon potential mapping (e.g., Czech Republic, Germany, Sweden, and United States). Where there are no existing houses or indoor radon measurements, proxy data (such as soil gas radon concentrations) are required to map radon potential.

Mapping levels of radon in administrative areas has the advantage of simplifying any subsequent administrative action. Use of grid squares allows an appropriate size of area to be chosen and simplifies the analysis. Use of geological boundaries may help to delineate differences in radon potential with greater spatial accuracy than other types of boundary. Whereas a wide variety of factors affect the concentration of radon in buildings, regional variations are related principally to the geological characteristics of the ground. Indoor radon surveys in the UK have confirmed the association of high levels of radon in dwellings with uraniferous granites, uraniferous sedimentary rocks, permeable limestones, and phosphatic ironstones, as well as fault and shear zones. Similar observations have been made in the Czech Republic, Germany, Luxembourg, Sweden, and the United States.

It is important to remember that however indoor radon data are grouped (whether by grid square or geological unit), a wide range of indoor radon levels is likely to be

found. This is because there is a long chain of factors that influence the radon level found in a building, such as radium content and permeability of the ground below it, and construction details of the building (Miles and Appleton 2000). Radon potential does not indicate whether a building constructed on a particular site will have a radon concentration that exceeds a reference level. This can only be established through measuring radon in the building.

Uncertainties relating to house-specific factors, proximity to geological boundaries and measurement error impact on the radon mapping process (Hunter et al. 2009, 2011; Bossew 2009).

11.11.1 Non-Geological Radon Potential Mapping

Radon measurements in existing dwellings are used to map the radon potential of countries (Fig. 11.4), administrative districts, or grid squares without taking into consideration the geological controls on radon in dwellings. Because the factors that influence radon concentrations in buildings are largely independent and multiplicative, the distribution of radon concentrations is usually lognormal, so lognormal modeling can be used to produce accurate estimates of the proportion of homes above a reference level (Miles 1998; Bossew 2010), although other transformations have also been investigated (Murphy and Organo 2008). In the UK, radon maps were used to show the fraction of the housing stock above the action level in each 5-km grid square (Fig. 11.5) (Lomas et al. 1996). Where house radon data are plentiful, maps using grid squares smaller than 5 km can be made. In some cases, this method can show up variations that are obscured by general geological grouping, such as variations in radon potential within a geological unit. Investigations in southwestern England revealed that the finer the grid, the closer the correlation with the geological controls of radon in dwellings.

Radon potential mapping using indoor radon measurements has been carried out in other European countries including Ireland, Luxembourg and France (Dubois 2005; Fennell et al. 2002; Organo and Murphy 2007), but the maps are not as detailed as the 5-km grid maps of the UK. This is mainly due to the relatively low measurement density and restricted coverage. In India and China, for example, average indoor radon concentrations of approximately 23 Bq m^{-3} in both countries are derived from national surveys of only 1,500 and 11,000 dwellings respectively (Cheng et al. 2002; Jin et al. 1996; Ramachandan and Sathish 2011) compared with over 4,500,000 available for radon mapping in the UK.

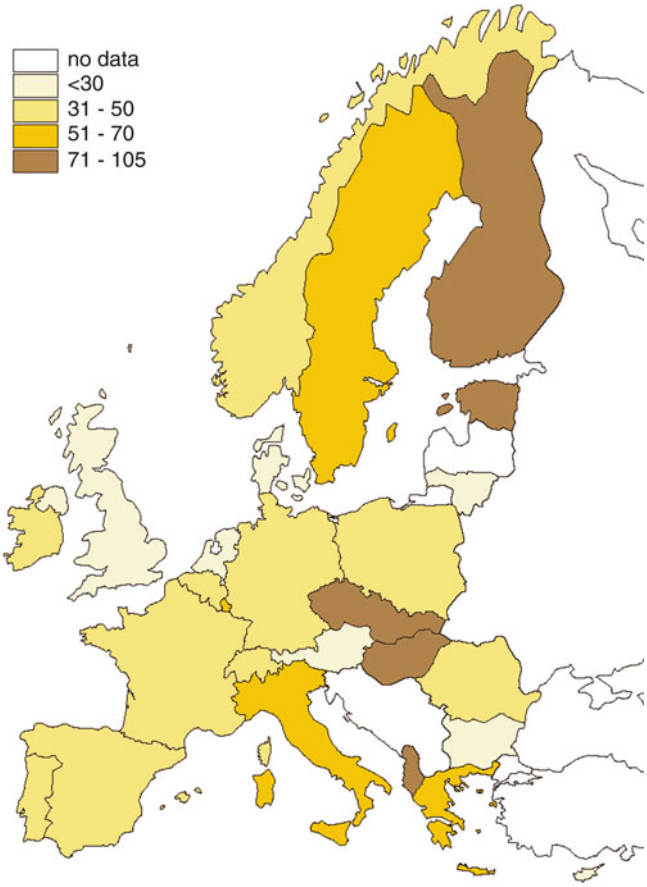


Fig. 11.4 Geometric mean radon concentrations (Bq m^{-3}) indoors in Europe (Compiled from data in UNSCEAR 2000)

In the United States most measurements of indoor radon have been made using short-term charcoal monitors, so these cannot be used directly to estimate long-term average radon levels. Although individual short-term measurement results are poor indicators of radon potential, aggregations of them can be corrected for bias and can provide useful information where no long-term results are available. A statistical technique known as Bayesian analysis improves estimates of mean radon level in areas where the data are sparse. A U. S. EPA survey of radon in homes covered about 6,000 homes across the United States, all measured using long-term etched track detectors. The data distribution is very sparse, given the size of the country, but it can provide estimates of the mean radon levels and distributions for the whole country and states, although not for smaller areas.

A range of statistical and geostatistical methods have been employed in some countries to map the spatial variation in indoor radon concentrations (Bossew 2009; Bossew et al. 2008; Dubois et al. 2007, 2010; Friedmann and Gröller 2010; Raspa et al. 2010; Tuia and Kanevski 2008; Zhu et al. 2001).

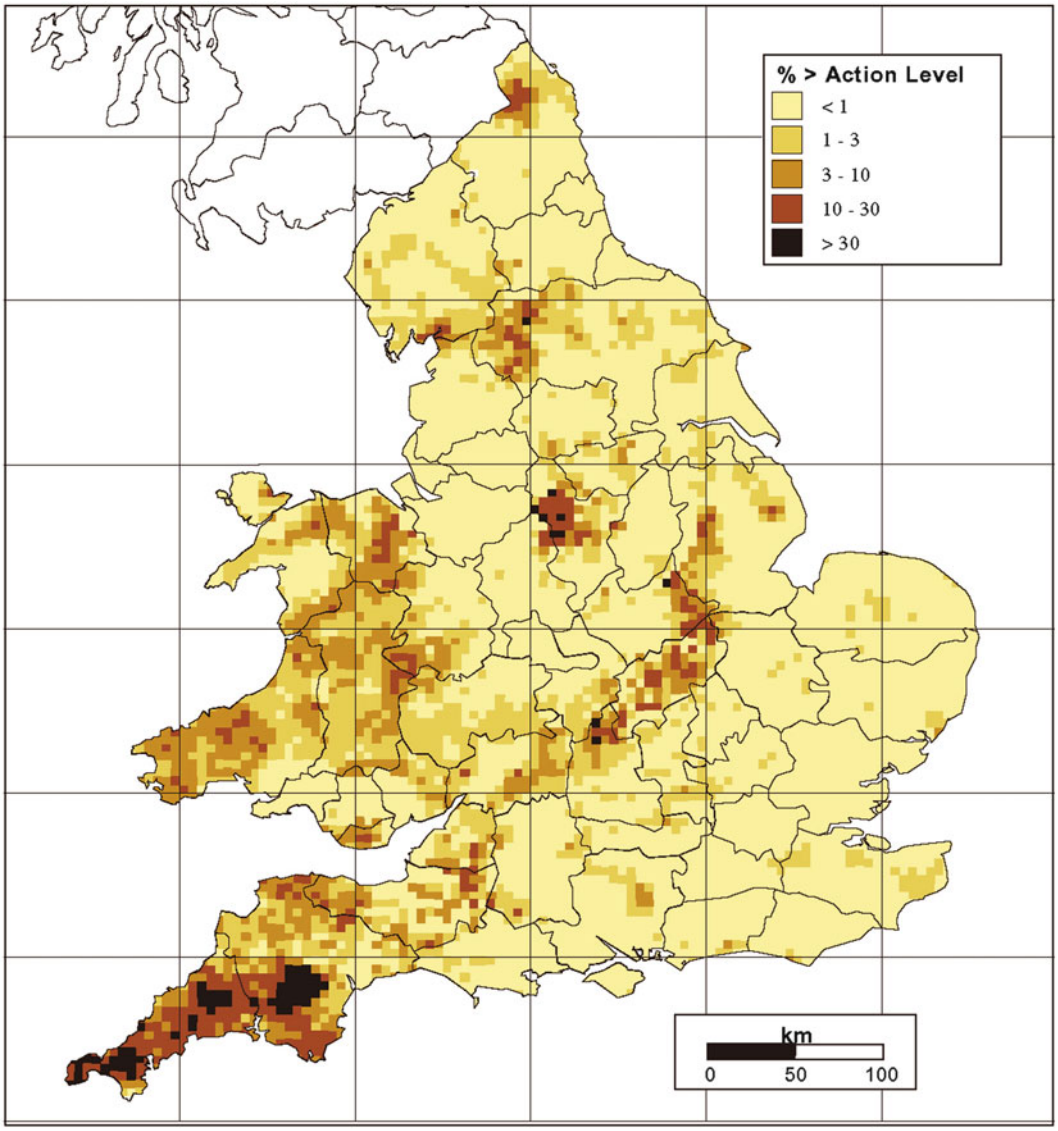


Fig. 11.5 Estimated proportion of homes exceeding the action level in each 5-km grid square of England and Wales (Adapted from Fig. 2–2 in Appleton et al. 2000a)

11.11.2 Geological Radon Potential Mapping

Approximately 25% of the total variation of indoor radon concentrations in England and Wales can be explained by the mapped bedrock and superficial geology. The proportion of the total variation explained by geology is higher (up to 37%) in areas where there is strong contrast between the radon potential of sedimentary geological units and lower (14%) where the influence of confounding geological controls, such as uranium mineralisation, cut across mapped geological boundaries. The proportion of the variation that can be attributed to mapped geological units increases with the level of detail of the digital geological data (Appleton and Miles 2010). Bossew et al. (2008) showed that the fraction of the variation in indoor radon concentrations in Austria explained by geology was 11.2% when the approximately 24,000 short-term indoor radon measurements were grouped into 111 relatively detailed geological units. As a consequence, the most accurate and detailed radon potential maps are generally those based on house radon data and geological boundaries provided that the indoor radon data can be grouped by sufficiently accurate geological boundaries. In the absence of an adequate number of high quality indoor radon measurements, proxy indicators such as soil gas radon data or information on U content may be used to assess geological radon potential. The reliability of maps based on proxy data increases with the number of classes as well as the quantity and quality of available data. Radon

potential maps based on indoor radon data grouped by geological unit have the capacity to accurately estimate the percentage of dwellings affected together with the spatial detail and precision conferred by the geological map data (Miles and Ball 1996). The reliability and spatial precision of mapping methods is, in general, proportional to the indoor radon measurement density. It is, however, reassuring that even when the measurement density is as low as the minimum for 5-km grid square mapping (i.e., 0.2–0.4/ km^2), geological radon potential mapping discriminates between geological units in a logical way. These relationships can be explained on the basis of the petrology, chemistry, and permeability of the rock units and are confirmed in adjoining map sheets with higher measurement densities (Miles and Appleton 2000).

Geological radon potential maps of the UK have been produced at 1:625,000, 1:250,000, and 1:50,000. Each geological unit within a map sheet or smaller area, such as a 5-km grid square, has a characteristic geological radon potential that is frequently very different from the average radon potential for the grid square shown (Fig. 11.6). Lithological variations within geological units can cause geological radon potential mapping to miss significant areas of higher radon potential identified by 1-km grid square mapping. Geological and grid square mapping are likely to be most powerful when used in a complementary fashion by integrating maps produced by the two methods and by grouping results both by geological unit and by grid square (Appleton and Miles 2002).

The factors that influence radon concentrations in buildings are largely independent and multiplicative, the distribution of radon concentrations is usually lognormal. Therefore lognormal modelling was used to produce accurate estimates of the proportion of homes above a reference level in the UK (Miles et al. 2007). Radon maps based on indoor radon data grouped by geological unit have the capacity to accurately estimate the percentage of dwellings affected together with the spatial detail and precision conferred by the geological map data. When indoor radon measurements are grouped by geology and 1-km squares of the UK national grid, the cumulative percentage of the variation between and within mapped geological units is shown to be 34–40% (Appleton and Miles 2010). This confirms the importance of radon maps that show the variation of indoor radon concentrations both between and within mapped geological boundaries. Combining the grid square and geological mapping methods gives more accurate maps than either method can provide separately (Fig. 11.7) (Miles and Appleton 2005).

In Austria, indoor radon data, normalised to a defined 'standard room' to reduce variations related to house type, are combined with geological information to produce a national radon potential map (Friedmann and Gröller 2010).

A combination of indoor radon measurements and geological information are used to produced radon potential maps in Canada (Chen 2009; Chen et al. 2011), Radon potential maps for highly populated areas are produced using indoor radon measurements where enough indoor radon data are available, whilst in areas where few or no indoor radon measurements are available, radon potential maps could be developed from various data sources with a multi-factor scoring system including geological information on soil permeability, soil gas radon concentration and ground uranium concentration, although both these approaches are constrained by the sparsity of indoor radon measurements and geological information relevant to radon (Chen 2009).

Uranium and radium concentrations in surface rocks and soils are useful indicators of the potential for radon emissions from the ground. Uranium can be estimated by gamma spectrometry either in the laboratory or by ground, vehicle, or airborne surveys (IAEA 2003; Dickson 2004). The close correlation between airborne radiometric measurements and indoor radon concentrations has been demonstrated in the United States in Virginia and New Jersey, Nova Scotia in Canada, and also in parts of the UK (Appleton and Ball 2001; Appleton et al. 2008, 2011a, b; Scheib et al. 2006), Ireland (Appleton et al. 2011a, b), Italy (Verdoya et al. 2009), Norway (Smethurst et al. 2008) and Thailand (Wattanikorn et al. 2007). Areas with high permeability tend to have significantly higher indoor radon levels than would be otherwise expected from the equivalent ^{226}Ra concentrations, reflecting an enhanced radon flux from permeable ground.

Duval and Otton (1990) identified a linear relationship between average indoor radon levels and surface radium content for soils of low to moderate permeability. However, areas with high permeability (>50 cmh^{-1}) had significantly higher indoor radon levels than would otherwise be expected from the ^{226}Ra concentrations, which reflects an enhanced radon flux from permeable ground. Grasty (1997) demonstrated that any estimate of natural gamma ray flux from the uranium decay series (i.e., radium) in the ground must take into consideration the radon coefficient of the soil as well as its radon diffusion coefficient, which depends largely on soil moisture. Clay soils tend to have higher eU when wet whereas sandy soils have lower eU (Grasty 1997).

Sweden was the first country to make use of airborne gamma ray spectrometry data to produce maps of radon potential. Radon potential is estimated and mapped on the basis of available data including (1) geology, (2) airborne radiometric surveys (covering 65% of Sweden), (3) results from radiometric surveys of the ground, (4) results from radon surveys in buildings, (5) results from earlier geotechnical investigations (e.g., permeability and groundwater level), (6) field surveys including gamma spectrometry,

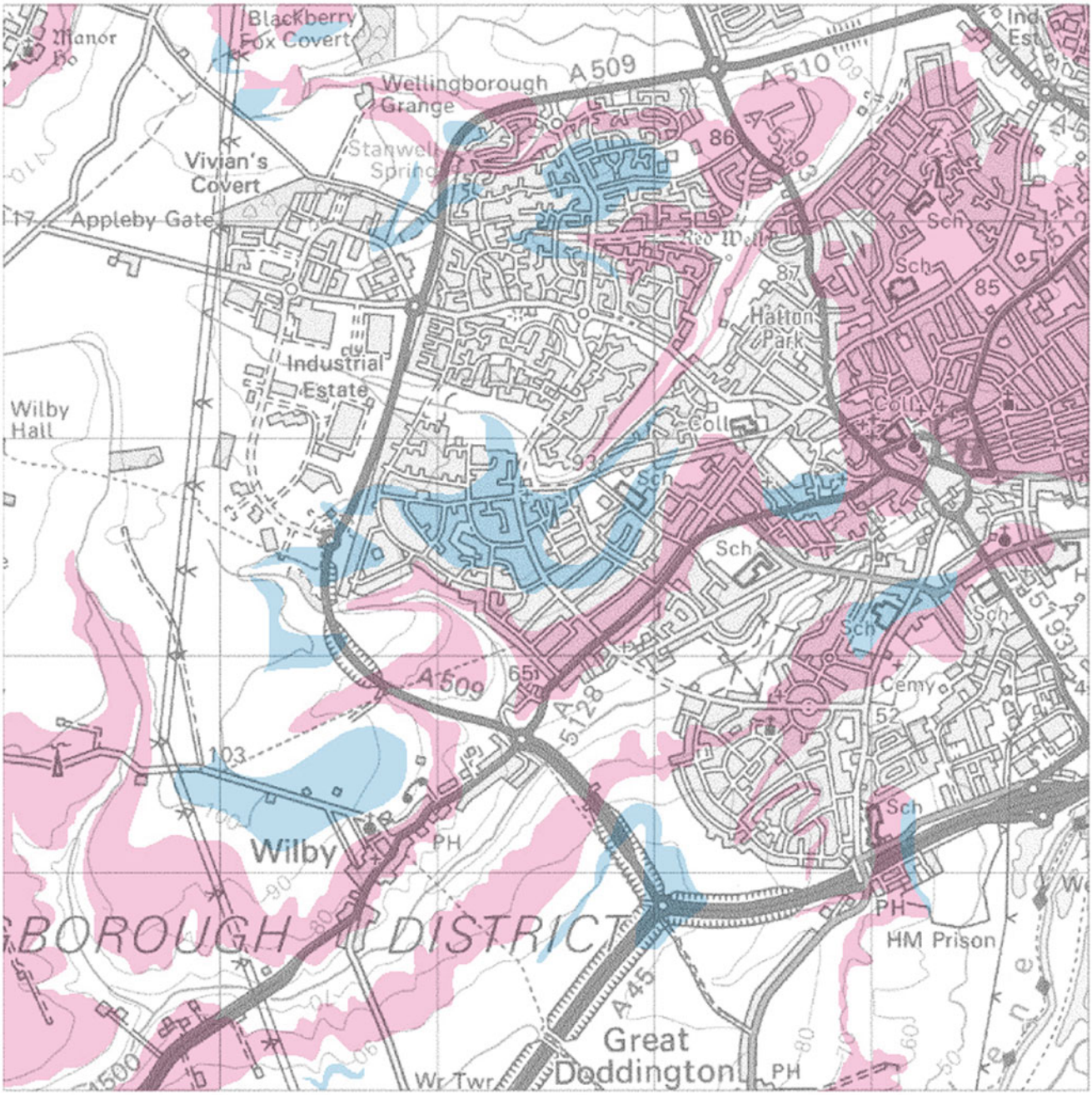


Fig. 11.6 Geological radon potential map of the 5-km grid square (485,265) that encompasses the western sector of Wellingborough, England. The 1:50,000 scale map illustrates the distribution of geological units with <3% (*white*), 3–5% (*blue*), and 10–20% (*pink*) of dwellings above the UK radon action level. The 5-km grid square has an average radon potential of 3.9% (NRPB 1998 data) (Topography based on ordnance survey 1:50,000 scale colour raster data with permission of The Controller of Her Majesty's Stationery Office Crown Copyright. Ordnance Survey Licence number GD272191/2004.)

and (7) orientation soil gas radon measurements. Åkerblom (1987) established a simple threefold radon risk classification based on geology, permeability, and soil gas radon (Table 11.9). These criteria are used at a mapping scale of 1:50,000 or larger in conjunction with airborne gamma spectrometry surveys to produce provisional radon risk maps both for indoor radon and groundwater (Akerblom and Lindgren 1997). A similar classification was used for the radon risk map of Estonia (Petersell et al. 2005)

Radon risk mapping of the Czech Republic at a scale of 1:500,000 (Fig. 11.8) is based upon a number of data sets for airborne radiometry, geology, pedology, hydrogeology, ground radiometry, and soil gas radon. Rock and soil permeability were obtained from hydrogeological and pedological

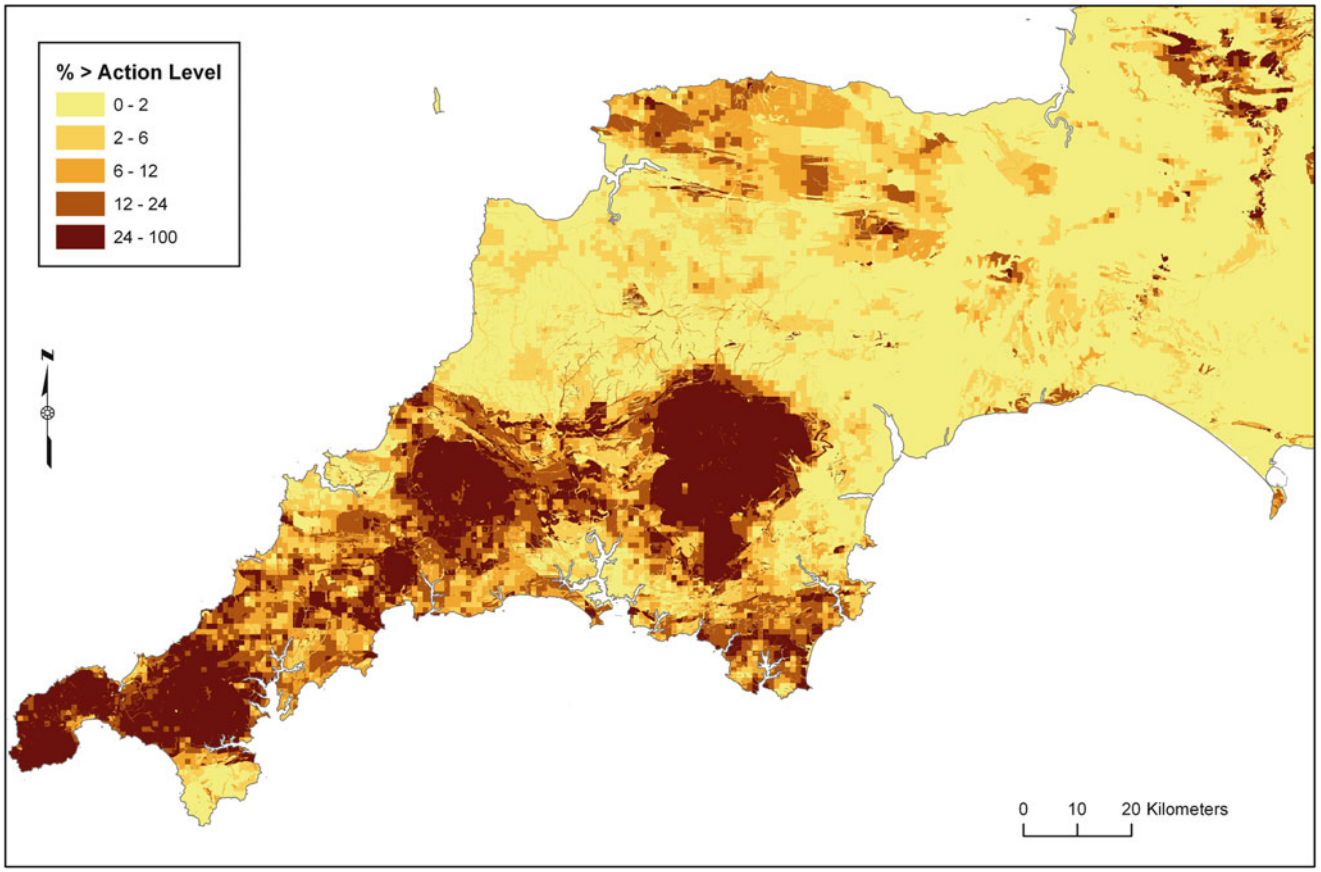


Fig. 11.7 Radon potential map of southwest England based on geology and indoor radon measurements (Adapted from Fig. 8 in Miles and Appleton 2005)

Table 11.9 Criteria used in Sweden for classifying high-radon and low-radon ground (After Clavensjö and Åkerblom 1994)

Bed-rock or overburden	^{226}Ra (Bq kg^{-1})	^{222}Rn in soil gas (Bq L^{-1})
High radon ground		
Bare rock	>200	not relevant
Gravel, sand, coarse till	>50	>50
Sand, coarse silt	>50	>50
Silt	>70	>60
Clay, fine till	>110	>120
Low radon ground		
Bare rock	<60	
Gravel, sand, till	<25	<20
Silt	<50	<20
Clay, fine till	<80	<60

maps and reports. Radon risk categories (low, medium, and high) were established for geological and lithological units and were based upon a rigid set of rules accepted by the Ministry of Environment. Radon risk maps are currently produced at the 1:50,000 scale, and these can be used for the identification of dwellings exceeding the guidance level to an accuracy of 70–80% (Miksová and Barnet 2002; Barnet et al. 2008). However, the maps are not recommended for the prediction of the requirement for radon protective measures in new buildings for which soil gas radon site assessments are required.

Airborne radiometric survey data were used to produce the first radon potential maps in the United States. The U. S. EPA radon map was developed using five factors to determine radon potential (indoor radon measurements, geology, aerial radioactivity, soil permeability, and foundation type). Radon potential assessment is based on geologic provinces adapted to county boundaries for the Map of Radon Zones (Fig. 11.9). The purpose of the map is to assist national, state, and local organizations to implement radon-resistant building codes. In the United States, high geological radon potential is associated with granites, limestones, black shales, and glacial tills and gravels derived from these bedrocks in the Appalachians; sandy and clay tills derived from sandstones, limestones, and black shales in the northern Great Plains and Great Lakes areas; and uraniferous granites, permeable limestones, sedimentary, and metamorphic rocks together with derived colluvial and alluvial deposits in the Rocky Mountains and parts of the western Great Plains.

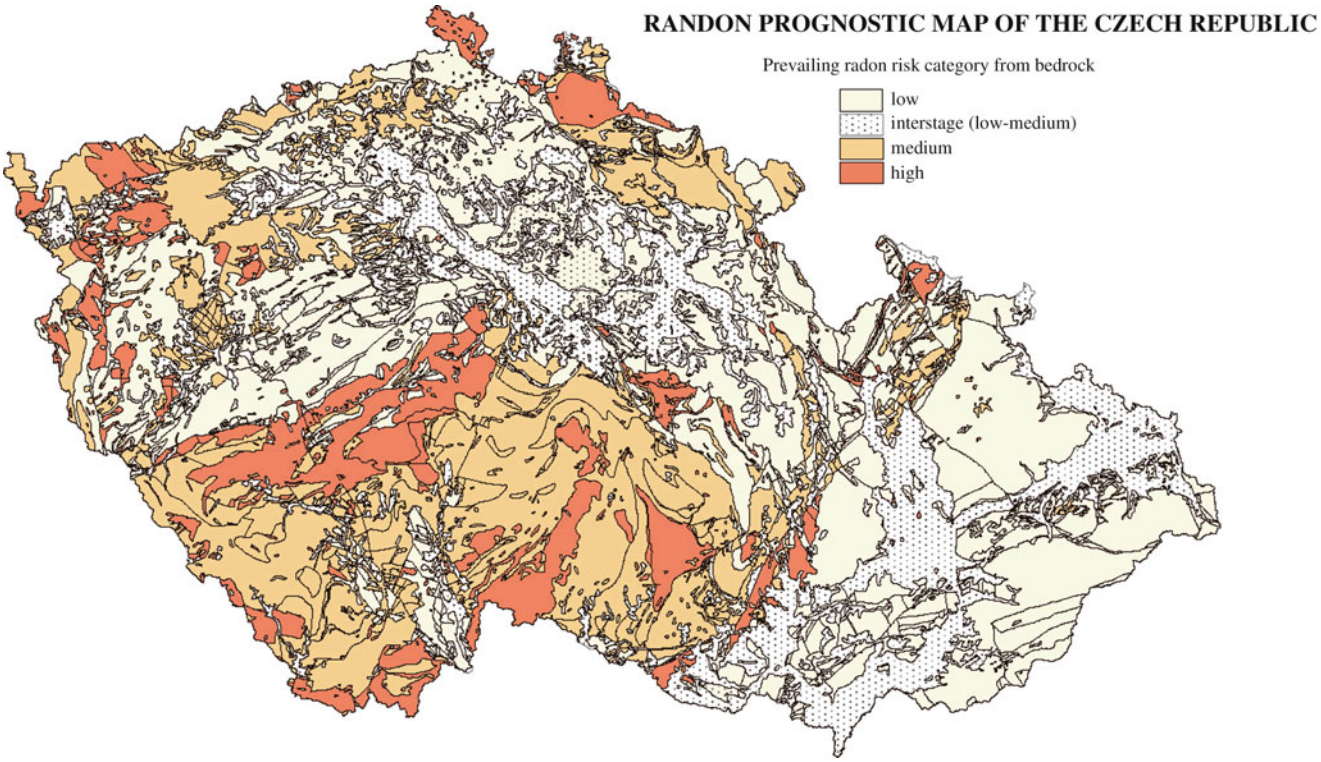


Fig. 11.8 Radon prognostic (risk) map of the Czech Republic (Reproduced with permission from C. Barnet, Czech Geological Survey.)

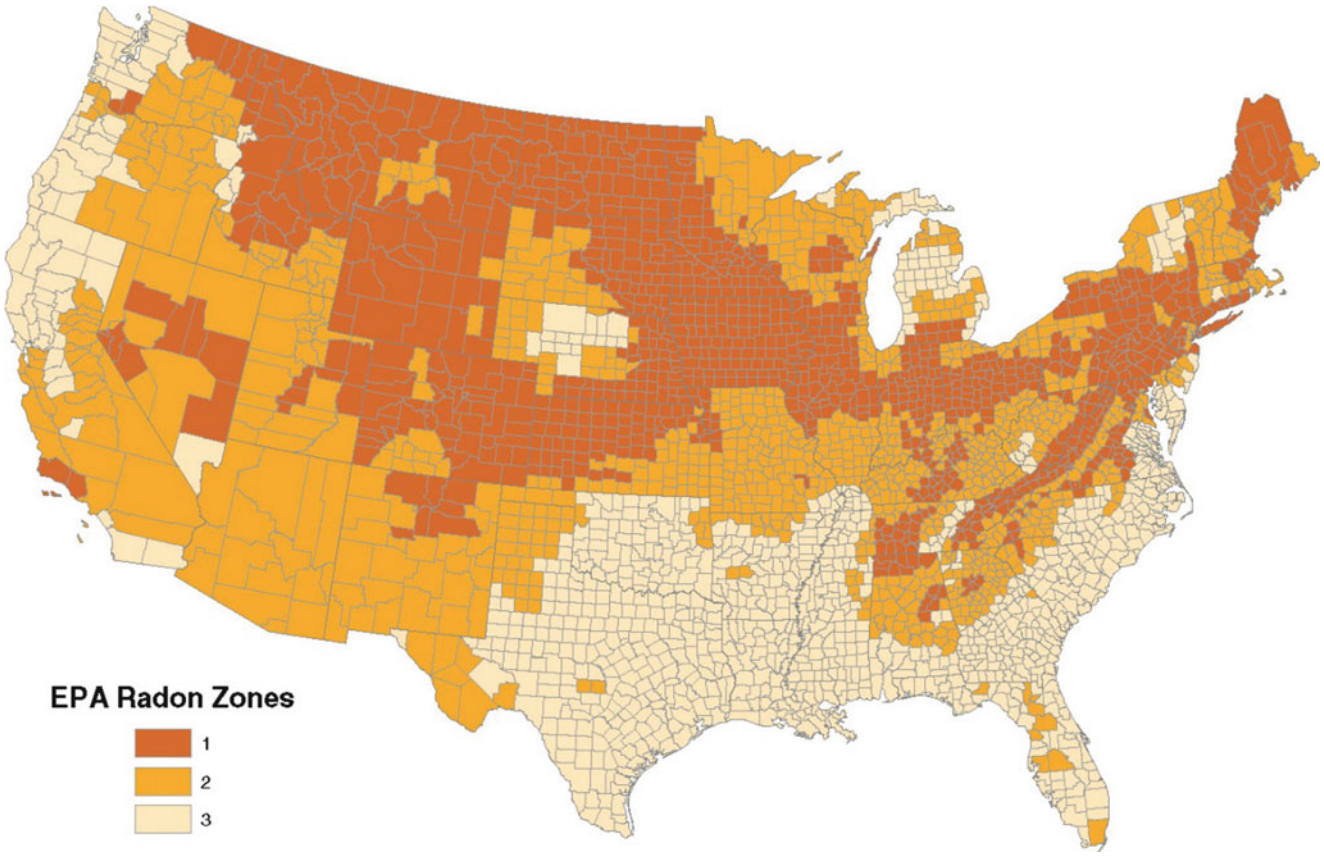


Fig. 11.9 U. S. EPA Map of Radon Zones (excluding Alaska and Hawaii). Zone 1, 2, and 3 counties have a predicted average indoor radon screening concentration of >148, 74–148, and <74 Bq m^{-3}, respectively (Map based on state radon potential maps available at http://www.epa.gov/radon/zonemap.html; state and county boundaries SRI ArcUSA 1:2 M.)

In Northern Ireland (Appleton et al. 2008, 2011c) linear regression analysis of airborne and soil geochemical parameters revealed that the most significant independent variables were eU, a parameter derived from airborne gamma spectrometry measurements of radon decay products in the top layer of soil and exposed bedrock, and the permeability of the ground. The radon potential map generated from airborne gamma spectrometry data agrees in many respects with the map based on indoor radon data and geology but there are several areas where radon potential predicted from the airborne radiometric and permeability data is substantially lower. This under-prediction could be caused by the radon concentration being lower in the top 30 cm of the soil than at greater depth, because of the loss of radon from the surface rocks and soils to air.

Smethurst et al. (2008) found a simple linear relationship between uranium concentration in the ground and the proportion of dwellings with radon concentrations above 200 Bq m^{-3} and demonstrated the spatial correlation between high concentrations of equivalent uranium and radon concentrations in dwellings, permitting the use of airborne radiometric data in radon hazard mapping, especially to classify areas with few indoor radon measurements. In areas of high permeable fluvio-glacial deposits, the airborne data underestimate the level of the radon hazard caused by the high ground permeability, thus confirming the observations of Scheib et al. (2006) and Appleton et al. (2008, 2011a, b) that in some geological domains, the permeability of the ground needs to be factored in to modeling of indoor radon from airborne radiometric data. In both Norway and Northern Ireland, additional areas of high potential radon hazard, not yet confirmed by indoor radon measurements, were identified.

The probability of homes in the Republic of Ireland having high indoor radon concentrations is currently estimated on the basis of known in-house radon measurements averaged over 10 x 10 km grid squares (Fennell et al. 2002). Airborne gamma-ray spectrometer data was used to predict radon potential in two distinct areas of Ireland using linear regression models derived from in-house radon measurements in conjunction with geological and ground permeability data The best agreement between the percentage of dwellings exceeding the Reference Level (RL) for radon concentrations in Ireland (% > RL), estimated from indoor radon data, and modelled RP in the Tralee-Castleisland area is produced using models based on airborne gamma-ray spectrometry equivalent uranium (eU) and ground permeability data. Good agreement was obtained between the % > RL from indoor radon data and radon potential (RP) estimated from eU data in the Cavan area using terrain specific models. In both areas, RP maps derived from eU data are spatially more detailed than the published 10-km grid map (Appleton et al. 2011b).

A harmonised methodology to derive a single map of the geogenic radon potential is being developed in order to improve regulatory tools for radon risk management in France, The method aims to determine the capacity of geological units to produce radon and facilitate its transfer, based on the interpretation of existing geological data. Geological units are classified according to their uranium content in order to create a radon source potential map. This initial map is enhanced by overlaying parameters, such as fault lines, which control the preferential pathways of radon through the ground and which can increase the radon levels in soils. The geogenic radon map provides a more precise zoning than the scale of the map of radon priority areas currently based solely on administrative boundaries (Ielsch et al. 2010).

After uranium and radium concentration, the permeability and moisture content of rocks and soils is probably the next most significant factor influencing the concentration of radon in soil gas and buildings. Radon diffuses farther in air than in water, so in unsaturated rocks and overburden with high fluid permeability, higher radon values are likely to result from a given concentration of uranium and radium than in less permeable or water-saturated materials. Weathering processes can also affect permeability. Enhanced radon in soil gas is also associated with high-permeability features such as fractures, faults, and joints. The fracturing of clays, resulting in enhanced permeability, combined with their relatively high radium content and their emanation efficiency may also result in higher radon concentrations in dwellings. The permeability of glacial deposits exerts a particularly strong influence on the radon potential of underlying bedrock.

It has been demonstrated in a number of countries, including Canada, Germany, the UK, the United States, and Sweden, that soil gas radon measurements combined with an assessment of ground permeability can be used to map geological radon potential in the absence of sufficient indoor radon measurements. Significant correlations between average indoor and soil gas radon concentrations, grouped according to geological unit, have been recorded in the Czech Republic (Fig. 11.10), Germany, the UK, and the United States.

Soil gas radon measurements combined with an assessment of ground permeability has been used to map geological radon potential in the absence of sufficient indoor radon measurements. 10–15 soil gas radon measurements are generally required to characterize a site or geological unit. Ground classifications based on geology, permeability and soil gas radon measurement have been developed in Sweden (Clavensjö and Åkerblom 1994), Finland, Germany (Kemski et al. 2008), Estonia and the Czech Republic (Barnet et al. 2008).

In Germany, Kemski and Klingel (2008) demonstrated that soil gas and indoor radon activity concentrations are connected by a transfer factor and that this factor can vary depending on regional differences in building characteristics, especially between the western and eastern part of Germany. Risk prediction maps for radon in houses can be derived from soil gas radon map data and the regional distribution of the radon transfer factor to produce a map indicating the probability of exceeding indoor radon threshold values. The probability maps can be used for indoor radon prediction, especially in regions without a sufficient number of indoor radon data (Kemski and Klingel 2008).

Where low correlations have been measured between radon in soil gas and radon in adjacent houses, the probable causes include: (1) the small number of houses with variable design in the study, (2) single rather than multiple soil gas measurements, (3) short-term indoor radon measurements, and (4) a mixture of summer and winter measurements. Spot measurements of soil gas radon and short-term indoor data are known to be relatively unreliable and it is now generally accepted that 10–15 soil gas radon measurements are required to characterize a site or geological unit.

11.12 Radon Site Investigation Methods

Radon migrates into buildings as a trace component of soil gas. Therefore the concentration of radon in soil gas should provide a good indication of the potential risk of radon entering a building if its construction characteristics permit the entry of soil gas. There is a growing body of evidence that supports the hypothesis that soil gas radon is a relatively reliable indirect indicator of indoor radon levels at the local as well as the national scale (Figs. 11.10 and 11.11). However, in some cases, soil gas radon data may be difficult to

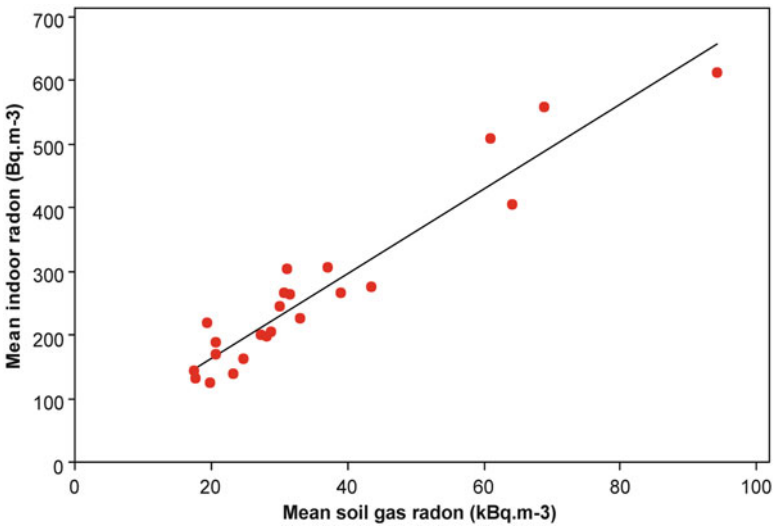


Fig. 11.10 Relationship between average soil gas radon (Bq L^{-1}) and average indoor radon (Bq m^{-3}) for major rock types of the Czech Republic (Based on data in Table 11.1, Barnet et al. 2010.)

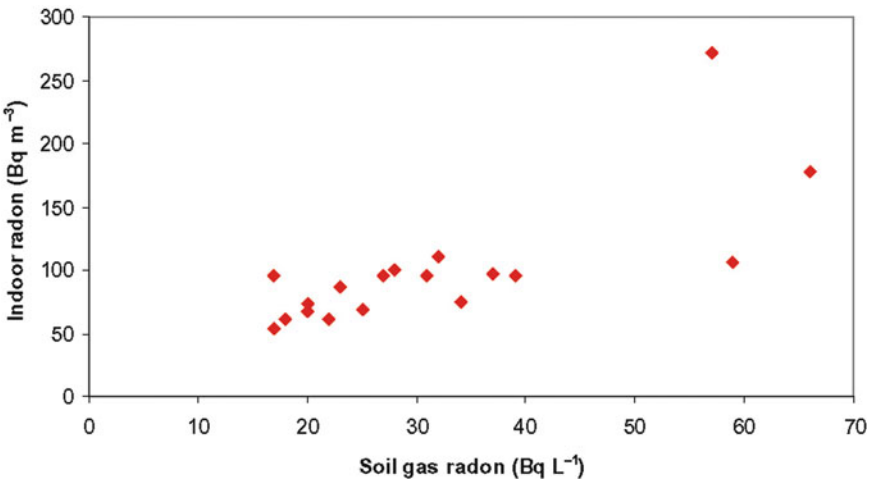


Fig. 11.11 Relationship between average soil gas radon concentration (Bq L^{-1}) and the geological radon potential (GEORP = estimated proportion of dwellings exceeding the UK radon action level, 200 Bq m^{-3}. Data for dwellings sited on the Jurassic Northampton Sand Formation grouped by 5-km grid square) (Reproduced from Appleton et al. 2000a.)

interpret due to the effects of large diurnal and seasonal variations in soil gas radon close to the ground surface as well as variations in soil gas radon on a scale of a few meters.

Soil gas radon data may be difficult to interpret due to the effects of large diurnal and seasonal variations in soil gas radon close to the ground surface and variations in soil gas radon on a scale of a few meters. The former problem may be overcome by sampling at a depth greater than 70 cm or by the use of passive detectors with relatively long integrating times, although this may not be a practical option if site investigation results are required rapidly. Small-scale variability in soil gas radon may be overcome by taking 10–15 soil gas radon measurements on a 5- to 10-m grid to characterize a site. Radon in soil gas varies with climatic changes including soil moisture, temperature, and atmospheric pressure. Weather conditions should be as stable as possible during the course of a soil gas radon survey. A range of methods such as controlled gas extraction, air injection procedures, or water percolation tests can be used to estimate gas permeability at a specific site. In the absence of permeability measurements, more qualitative estimates of permeability can be based on visual examination of soil characteristics, published soil survey information, or on the relative ease with which a soil gas sample is extracted.

In some areas and under some climatic conditions, site investigations using soil gas radon cannot be carried out reliably, for example, when soil gas cannot be obtained from waterlogged soils or when soil gas radon concentrations are abnormally enhanced due to the sealing effect of soil moisture. These conditions are particularly common in winter. Problems with the determination of permeability and its incorporation into a radon site investigation procedure have been encountered in the Czech Republic where the quality of the permeability classification obtained at a site is very reliant on the personal experience of the technical staff carrying out the site investigation. If soil gas radon concentrations cannot be determined because of climatic factors, measurement of radon emanation in the laboratory or gamma spectrometric measurement of eU can be used as radon potential indicators in some geological environments. However, few data are available and the methods have not been fully tested.

The Swedish National Board of Housing, Building, and Planning has adopted a ground classification based on geology, permeability, and soil gas radon measurement. (Clavensjö and Akerblom 1994) This procedure is used to predict radon emissions expected on a particular construction site. Finland has adopted a similar radon risk classification of building ground based on radioactivity and permeability. Measurement of radon emanation coefficients and radium concentrations by gamma spectrometry is also used to investigate radon characteristics of the ground in new building areas where buildings are to be constructed on unconsolidated sediments or directly onto bedrock (Table 11.9) (Clavensjö and Åkerblom 1994). In Germany an empirical ranking classification has been developed for radon potential based on median soil gas radon and permeability measured by air injection through the soil gas probe. All new development sites in the Czech Republic require a site investigation comprising a geological survey and measurement of radon in soil gas. The radon risk classification (Table 11.10) is based upon soil gas radon concentration limits and is broadly similar to classifications used in Finland, the UK, the United States, and Sweden. The Czech Radon Risk Classification For Foundation Soils (Table 11.10) has now been replaced by a classification (Fig. 11.12) based on the determination of the building site radon index (Barnet et al. 2008).

Table 11.10 Czech Republic radon risk classes based on radon in soil gas and rock-overburden permeability (After Barnet 1994)

	Rock-overburden permeability		
	High	Medium	Low
Radon risk	Radon concentration in soil gas (Bq L^{-1})		
High	>30	>70	>100
Medium	10–30	20–70	30–100
Low	<10	<20	<30

Radon may also be a problem underground, such as in tunnels. Radon emanation from borehole core samples can be determined to derive values of radon emanation per unit of surface area, which is an effective and simple way of assessing radon hazard in tunnels at the site investigation stage (Talbot et al. 1997).

11.13 Geological Associations

Geology exerts a strong control on indoor radon and radon hazard. Relatively high concentrations of radon are associated with particular types of bedrock and unconsolidated deposits, for example some granites, uranium-enriched phosphatic rocks and black shales, limestones, sedimentary ironstones, permeable sandstones and uraniferous metamorphic rocks. Fault shear zones, glacial deposits derived from uranium-bearing rock and even clay-rich sediments from which high radon emanation reflects large surface area and high permeability caused by cracking, may be radon prone along with uranium and radium enriched soils derived from limestone (especially karstic terrain), uranium mining residues, and mine tailings.

Rock types that are high radon sources in the United States include:

1. Uraniferous metamorphic rocks and granites—some of the highest indoor levels in the United States, particularly in the Rocky and Appalachian ranges and the Sierra Nevada are associated with fault shear zones in these rocks

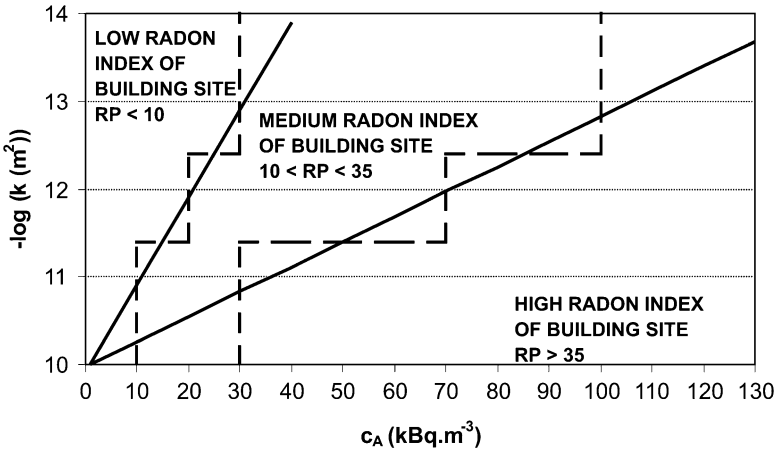


Fig. 11.12 Czech radon potential classification for assessing the radon risk of building sites (copy of Fig. 6–2, Barnet et al. 2008; reproduced with permission from I. Barnet, Czech Geological Survey)

2. Marine black shales are sources of high radon throughout the United States and especially in the central region from Ohio to Colorado
3. Glacial deposits derived from uranium-bearing rock and sediment, especially in the northwestern Midwest, where high radon emanation reflects large surface area and high permeability caused by cracking when dry
4. Soils derived from carbonate, especially karstic terrain, which are high in uranium and radium
5. Uranium mining residues and mine tailings in the states of the western United States (e.g., Colorado)
6. Phosphate ore close to the surface and in mining waste on the surface, can result in high radon concentrations, especially in Polk County, Florida.

The maximum ^{226}Ra concentration in phosphate ores is typically about 50 times greater than the average concentration in soil. Releases from coal residues and the burning of natural gas and coal complete the list of major contributors to atmospheric radon in the USA (Gundersen et al. 1992).

Enhanced levels of radon in houses and soil gas in the UK are associated with the following geological associations. The *uranium* association comprises rocks and their weathering products containing enhanced levels or uranium or radium. The *permeable rock* association comprises permeable rocks, unconsolidated overburden, and their weathering products. Areas underlain by less permeable rocks, unconsolidated overburden, and soils, especially where these have low uranium concentrations, are generally characterized by low radon in houses and soil. Low radon is also associated with permeable sandstones containing low quantities of uranium.

The uranium association comprises granites in southwestern England characterized by high uranium concentrations, a deep weathering profile, and uranium in a mineral phase that is easily weathered. Although the uranium may be removed through weathering, radium generally remains *in situ* (Ball and Miles 1993). Radon is easily emanated from the host rock and high values of radon have been measured in groundwaters and surface waters (110–740 Bq L^{-1}) and also in soil gas (frequently >400 Bq L^{-1}). There is a clear correspondence between areas where more than 30% of the house radon levels are above the action level and the major granite areas (Ball and Miles 1993). In SW England, the Grampian and Helmsdale areas of Scotland (Scheib et al. 2009) and the Mourne Mountains in Northern Ireland (Appleton et al. 2011c) there is a correlation between areas where it is estimated that more than 30% of the house radon levels are above 200 Bq m^{-3} and the major granite areas. Groundwater wells with >500 B L^{-1} are associated with uranium rich granite bedrock in Sweden, especially when fractures are coated with ^{226}Ra, and also with areas containing many pegmatitic intrusions (Akerblom and Lindgren 1997)

The depositional and diagenetic environment of many black shales leads to enrichment of uranium. For example, some Lower Carboniferous shales in northern England and NE Wales contain 5–60 mg kg^{1} uranium. Weathering and secondary enrichment can substantially enhance U levels in soils derived from these shales. It is found that on average 15–20% of houses (rising to more than 65% in some areas) sited on uraniferous black shales with >60 mg kg^{1}U and high soil gas radon (32 Bq L^{-1}; Ball et al. 1992) are above the UK radon action level.

Uranium-enriched phosphatic horizons occur in the Carboniferous Limestone, the Jurassic oolitic limestones, and in the basal Cretaceous Chalk in the UK and these sometimes give rise to high radon in soil gases and houses. Many iron deposits are phosphatic and slightly uraniferous and a large proportion (>20%) of houses underlain by the Northampton Sand Formation (NSF) ironstone in England are affected by high levels of radon (Figs. 11.5 and 11.11). Phosphatic pebbles from the Upper Jurassic, and Lower and Upper Cretaceous phosphorite horizons in England contain 30–119 mg kg^{-1}U. Radon in dwellings is a significant problem in areas where these phosphatic rocks occur close to the surface, especially if the host rocks are relatively permeable.

The NSF consists of ferruginous sandstones and oolitic ironstone with a basal layer up to 30 cm thick containing phosphatic pebbles. Whereas the ferruginous sandstones and ironstones mainly contain low concentrations of U (<3 mg kg^{-1}), the phosphatic pebbles contain up to 55 mg kg^{-1}. It is, however, probable that the mass of the NSF, which in many cases contains disseminated radium, may contribute more to the overall level of radon emissions than the thin U-enriched phosphate horizons.

High levels of radon occur in both soil gas and houses underlain by Carboniferous Limestone in the UK as well as in caves and mines. There are 10% to more than 30% of houses built on the limestones that have radon concentrations greater than the UK action level (Appleton et al. 2000a). Much of the radon is thought to emanate from uranium- and radium-enriched residual soils that overlie the highly permeable limestones.

Chalk is a particularly abundant limestone in the south of England, but its radon emanation characteristics are different from the Carboniferous and Jurassic limestones. Chalk still retains its primary porosity, although most of the water and gas flow is through fissures. The proportion of dwellings with radon above the action level is much lower than over the Carboniferous Limestone, but higher levels of radon occur where the chalk is covered with congeliturbate and residual clay-with-flint deposits.

Thick, permeable Cretaceous sand formations in south-western England, including the glauconitic Lower and Upper Greensand and the Upper Lias Midford Sands, all emanate high levels of soil gas radon (mean values 20–48 Bq L^{-1}), and are characterized by a high proportion of houses above the action level (13 and 22% for the Upper Greensand and Midford Sands, respectively). In contrast, impermeable mudstones and clays in England and Wales are generally characterized by low to moderate soil gas radon (about 20 Bq L^{-1}) and less than 1% of homes exceed the action level.

Similar associations between high radon and Lower Carboniferous limestones, Namurian uraniferous and phosphatic black shales, and some granites and highly permeable fluvioglacial deposits have also been recorded in Ireland (Cliff and Miles 1997).

In the Czech Republic, the highest indoor and soil gas radon levels are associated with the Variscan granites, granodiorites, syenites, and phonolites of the Bohemian massif. Syenites contain 12–20 mg kg^{-1}U and the phonolites have 10–35 mg kg^{-1}U and soil gas radon levels up to more than 450 Bq L^{1}. High radon is associated with Paleozoic metamorphic and volcanic rocks and also with uranium mineralization in the Pribram area (Barnet et al. 2002; Miksová and Barnet 2002; Barnet et al. 2008, 2010)

In Germany the highest radon occurs over the granites and Paleozoic basement rocks. Median soil gas radon for some granites ranges from 100 to 200 Bq L^{-1} (Kemski et al. 2001, 2009). In contrast, the highest radon potential in Belgium is associated with strongly folded and fractured Cambrian to Lower Devonian bedrocks in which uranium preferentially concentrated in ferric oxyhydroxides in fractures and joints is considered to be the main source of radon (Zhu et al. 2001). In France some of the highest radon levels occur over peraluminous leucogranites or metagranitoids in a stable Hercynian basement area located in South Brittany (western France). These rocks are derived from uraniferous granitoids with average uranium contents of over 8 mg kg^{-1} (Ielsch et al. 2001a, b). Soil gas and indoor radon concentrations were found to be controlled by lithology, structure, and uranium mineralization in India (Singh et al. 2002). High radon is associated with alum shale in both Sweden (Tell et al. 1993) and Belgium (Poffijn et al. 2002). In Korea, the mean values of soil gas radon concentrations were highest in granite gneiss and banded gneiss and lowest in soils over shale, limestone, and phyllite schist (Je et al. 1999). High indoor radon concentrations are associated with uranium-rich granite (Wang et al. 2005) in Zhuhai City, China with the two highest indoor radon concentrations found in workshops producing potable water from granite groundwater containing high radon concentrations (1,300–1,980 Bq L^{-3}) (Chen et al. 2005). Radon from thermal spring waters in granite terrains contributes a major proportion of indoor radon in hot spring hotels located in Guangdong province, China (Song et al. 2011). High concentrations of radon and radium are also found in groundwaters associated with granite in Fujian province, China (Zhuo et al. 2001). Wang et al. (2009, 2011) found that there is a strong correlation between radioactivity level and geological lithology with the highest exhalation rates and soil gas radon concentrations being associated with biotite granite (151 kBq m^{-3}) and lower concentrations with weathered granite (38 kBq m^{-3}) and Quaternary sediments (7 kBq m^{-3}). The impact of unconsolidated deposits mainly reflects their permeability. For example, peat and lacustrine clays strongly reduce radon potential associated with the underlying bedrock, whereas permeable sand and gravel and river terrace deposits tend to enhance radon potential. In Sweden fragments and mineral grains of uranium-rich granites, pegmatites, and black alum shales are dispersed in till and glaciofluvial deposits leading to high radon in soils and dwellings, especially when the glaciofluvial deposits are highly permeable sands and gravels (Clavensjö and Åkerblom 1994). In Norway, high radon levels are associated with highly permeable fluvioglacial and fluvial deposits derived from all rock types and also with moderately permeable unconsolidated sediments (mainly basal till) containing radium-rich rock fragments (Smethurst et al. 2008; Sundal et al. 2004).

11.14 Administrative and Technical Responses

11.14.1 Environmental Health

Environmental health responses include provision of guidance for radon limitation including recommendations for dose limits and action levels, establishment of environmental health standards for houses and workplaces, and enforcement of Ionizing Radiations Regulations to control exposure to radon in workplaces (Åkerblom 1999; Appleton et al. 2000b; NRPA 2000; WHO 2009). National authorities are recommended to set a reference level for radon that represents the maximum accepted radon concentration in a residential dwelling. Remedial actions may be recommended or required for homes in which radon exceeds the reference level. Various factors such as the distribution of radon concentrations, the number of existing homes with high radon concentrations, the arithmetic mean indoor radon level and the prevalence of smoking are usually considered when setting a national radon reference level. WHO (2009) proposed a reference level of 100 Bq m^{-3} to minimize health hazards due to indoor radon exposure although if this concentration cannot be reached, the reference level should not exceed 300 Bq m^{-3}. This represents approximately 10 mSv per year according to recent calculations and recommendations by the International Commission on Radiation Protection (ICRP 2007, 2009) which also recommends 1,000 Bq m^{-3} as the entry point for applying occupational radiological protection requirements. There are substantial variations in action levels (or their equivalents) in countries that perceive a radon problem. International and national recommendations for radon limitation in existing and future homes, given as the annual average of the gas concentration in Bq m^{-3}, range from 150 to 1,000 for existing dwelling and from 150 to 250 for new dwellings (Åkerblom 1999). The majority of countries have adopted 400 and 200 Bq m^{-3}, respectively, for the two reference levels (HPA 2009).

The reasons for these different reference levels appear largely historical but are also due to a combination of environmental differences, different construction techniques, and varying levels of political and environmental concern. There would be advantages in harmonizing standards because the existence of different levels may lead to confusion among the public. The ICRP considers that one common international standard is unlikely to be achieved, and that this is less important than achieving reasonable reductions in radon levels in radon prone areas.

In addition to variations in house radon standards between countries, there also appears to be some variation in standards applied within the field of radiological protection. For example, some observers have suggested that if the highest natural radon levels recorded in some houses in Cornwall and Devon were reached, for example, at a nuclear installation, it would be closed down immediately. Equally, it is reported that all houses in the Chernobyl area had to be evacuated under Soviet law if they were contaminated to a level equivalent to the current UK action level of 200 Bq m^{-3}. However, international and national radiological protection authorities are united in acknowledging the need for a distinction in the ways radiation is approached in these different circumstances.

Recommendations differ from country to country. In the UK, testing for radon is recommended by government in radon Affected Areas where more than 1% of dwellings exceed the action level of 200 Bq m^{-3}. In 2010, HPA guidance on radon gas concentrations introduced a new 'target level' of 100 Bq m^{-3} above which remediation work should be seriously considered especially by high risk groups such as smokers and ex smokers. In contrast, the U. S. EPA recommends that all homes be tested for radon because (1) high levels of indoor radon have been found in every state, (2) radon levels vary so much from place to place, and (3) because dwellings differ so radically in their vulnerability to radon. The U. S. EPA estimates that 1 in every 15 homes has radon levels higher than 4pCiL[1] (148 Bq m^{-3}), the level above which the EPA recommends that corrective action is taken. The U.S. EPA recommends that no action is required below 150 Bq m^3, action within a few years between 170 and 750 Bq m^{-3}, urgent action between 750 and 7,500 Bq m^{-3}, and immediate action above 7,500 Bq m^{-3}.

The European Commission Recommendation (2001/928/Euratom) on the protection of the public against exposure to radon in drinking water supplies recommends 1,000 Bq/L as an action level for public and commercial water supplies above which remedial action is always justified on radiological protection grounds. Water supplies that support more than 50 people or distribute more than 10 m^3 per day, as well as all water that is used for food processing or commercial purposes, except mineral water, are covered by the Europe Commission Recommendation. The 1,000 Bq/L action level also applies to drinking water distributed in hospitals, residential homes, and schools and should be used for consideration of remedial action in private water supplies. The U. S. EPA recommends that states develop multimedia mitigation (MMM) programs to address the health risks from radon in indoor air while individual water systems should reduce radon levels in drinking water to 148 Bq L^{-1} (4,000 pCi L^{-1}) or lower. The EPA is encouraging states to adopt this option because it is the most cost-effective way to achieve the greatest radon risk reduction. If a state chooses not to develop an MMM program, individual water systems

would be required to either reduce radon in their system's drinking water to 11 Bq L^{-1} (300 pCi L^{-1}) or develop individual local MMM programs and reduce levels in drinking water to 148 Bq L^{-1} (4,000 pCi L^{-1}). The regulations will not apply to private wells, because the EPA does not regulate them. A guideline value of 2 μ gL^{-1} uranium (equivalent to approximately 0.02 Bq L^{-1}) is recommended by the World Health Organization (WHO), although this is based on its toxicity, which is more detrimental to health than its radioactivity (WHO 1996). The U. S. EPA established maximum contaminant levels (MCL) of 30 μg L^{-1} uranium, and 0.185 Bq L^{-1} for radium-226 and radium-228 in community water supplies. No specific value for uranium is given in the EU Directive for drinking water (CEC 1998), which establishes a 0.1 $mSvy^{-1}$ total indicative dose guidance level for radionuclides, excluding tritium, ^{40}K radon, and radon decay products. Action levels ranging from 7.4 to 160 Bq L^{-1} uranium in drinking water have been reported for Austria, Finland, and France.

11.14.2 Radon Monitoring

The overall aim of most countries that have identified a radon problem is to map radon-prone areas and then identify houses and workplaces with radon concentrations that exceed the radon reference level. In the UK, for example, radon affected areas are delineated by measuring radon in a representative sample of existing dwellings. Householders are then encouraged to have radon measured in existing and new dwellings in affected areas and local authority environmental health departments are generally responsible for ensuring that radon in workplaces is monitored in appropriate areas.

11.14.3 Protective Measures

Provisions have been made in the building regulations to ensure that new dwellings are protected against radon where a significant risk of high radon concentrations in homes has been identified on the basis of house radon surveys. Nine European countries (Czech Republic, Denmark, Finland, Ireland, Latvia, Norway, Slovak Republic, Sweden, and the UK) have regulations and guidelines for construction requirements to prevent elevated radon concentrations in new buildings. Austria, Germany, Greece, and Switzerland plan to introduce such regulations. In most of the countries with regulations, enforced radon protection in new buildings is specified in the national building codes. Implementation of regulations is normally shared by national and local authorities. Eight European countries (Czech Republic, Denmark, Finland, Ireland, Norway, Slovak Republic, Sweden, and the UK) have regulations and guidelines for radon prevention in the planning stages of new development (e.g., where construction permits are applied for dwellings, offices, and factories). Austria and Germany are considering the introduction of guidance and/or regulations for dealing with radon at the planning stage. In the Czech Republic, Ireland, Slovak Republic, and Sweden regulations require an investigation of radon risk at construction sites before building is permitted (Åkerblom 1999). Cost-effectiveness analysis in the UK suggest that radon protective (preventive) measures, such as a sealed membrane under the ground floor in new homes would be justified in all areas (HPA 2009).

11.14.4 Remedial Measures

In the UK, owners of workplaces may be forced to carry out remedial measures whereas householders in dwellings with radon above the action level are generally only advised to take action to reduce the radon level. Guidance on reducing radon in dwellings is provided, but the cost of installing remedial measures in a dwelling is normally the householder's responsibility. Grant aid may be available.

The principal ways of reducing the amount of radon entering a dwelling are similar to those used for protective measures in new dwellings. These are

1. Install an airtight barrier across the whole of the ground floor to prevent radon getting through it and also seal voids around service inlets
2. Subfloor ventilation of underfloor cavities, i.e., drawing the air away from underneath the floor so that any air containing radon gas is dispersed outside the house
3. Subfloor depressurization (radon sump)
4. Positive pressurization (i.e., pressurize the building in order to prevent the ingress of radon)
5. Ventilation (i.e., avoid drawing air through the floor by changing the way the dwelling is ventilated)

In the United States, the cost of radon mitigation in a typical home ranges from about $500 to about $2,500. Fitting radon resistant measures at the time of construction would cost $350–$500. Similar costs apply in the UK.

Radon gas may be easily removed from high-radon groundwaters by aeration and filter beds will remove daughter products. Various aeration technologies are available including static tank, cascade, or forced aeration in a packed tower. Radon removal technologies used in the United States include removal of ^{222}Rn by spray jet aeration, packed tower aeration, and multistage bubble aeration. Packed tower aeration is simple and cheap and is recommended for large drinking water supplies. Removal of ^{222}Rn by granular activated carbon is efficient but ^{238}U decay products, including U, Po, Bi, and Pb (^{210}Pb), are adsorbed onto the activated

carbon, which produces a disposal problem. The U. S. EPA recommends that the most practical treatment methods for radionuclide removal are ion exchange and lime-soda softening for radium, aeration and granular activated carbon for radon, and anion exchange and reverse osmosis for uranium.

Monitoring and remediation in existing homes are not cost-effective at present in the UK but might become so in areas with mean indoor radon concentrations of 60 Bq m^{-3} or above if the UK Action Level was reduced from 200 to 100 Bq m^{-3} (HPA 2009).

See Also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements) • Chapter 10 (Volcanic Emissions and Health) • Chapter 23 (Environmental Epidemiology)

References

Åkerblom G (1987) Investigations and mapping of radon risk areas. In: Proceedings of international symposium on geological mapping, Trondheim, 1986: in the service of environmental planning. Norges Geologiske Undersoekelse, Oslo, pp 96–106

Åkerblom G (1999) Radon legislation and national guidelines 99:18. Swedish Radiation Protection Institute

Åkerblom G, Falk R, Lindgren J, Mjönes L, Östergren I, Söderman A-L, Nyblom L, Möre H, Hagberg N, Andersson P, Ek B-M (2005a) Natural radioactivity in Sweden, exposure to external radiation, radiological protection in transition. In: Proceedings of the XIV regular meeting of the Nordic Society for Radiation Protection. NSFS Rättvik, Sweden, pp 207–210

Åkerblom G, Falk R, Lindgren J, Mjönes L, Östergren I, Söderman A-L, Nyblom L, Möre H, Hagberg N, Andersson P, Ek B-M (2005b) Natural radioactivity in Sweden, exposure to internal radiation, radiological protection in transition. In: Proceedings of the XIV regular meeting of the Nordic Society for Radiation Protection. NSFS Rättvik, Sweden, pp 211–214

Akerblom G, Lindgren J (1997) Mapping of groundwater radon potential, Uranium exploration data and techniques applied to the preparation of radioelement maps; proceedings. International Atomic Energy Agency, Vienna, pp 237–255

Åkerblom G, Mellander H (1997) Geology and radon. In: Durrani SA, Ilic´ R (eds) Radon measurements by etched track detectors. World Scientific, River Edge, pp 21–49

Appleton JD, (2004) Influence of faults on geological radon potential in England and Wales. In: Barnet I, Neznal M, Pacherová P (eds) Radon investigations in the Czech Republic X and the seventh international workshop on the geological aspects of radon risk mapping. Czech Geological Survey & Radon Corporation, Prague

Appleton JD, Ball TK (2001) Geological radon potential mapping. In: Bobrowsky PT (ed) Geoenvironmental mapping: methods, theory and practice. Balkema, Rotterdam, pp 577–613

Appleton JD, Miles JCH (2002) Mapping radon-prone areas using integrated geological and grid square approaches. In: Barnet I, Neznal M, Mikšová J (eds) Radon investigations in the Czech Republic IX and the sixth international workshop on the geological aspects of radon risk mapping. Czech Geological Survey, Prague, pp 34–43

Appleton JD, Miles JCH (2010) A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk. J Environ Radioact 101:799–803

Appleton JD, Miles JCH, Talbot DK (2000a) Dealing with radon emissions in respect of new development: evaluation of mapping and site investigation methods for targeting areas where new development may require radon protective measures. British geological survey research report, RR/00/12

Appleton JD, Miles JCH, Scivyer CR, Smith PH (2000b) Dealing with radon emissions in respect of new development: summary report and recommended framework for planning guidance. British geological survey research report, RR/00/07

Appleton JD, Miles JCH, Green BMR, Larmour R (2008) Pilot study of the application of tellus airborne radiometric and soil geochemical data for radon mapping. J Environ Radioact 99:1687–1697

Appleton JD, Cave MR, Miles JCH, Sumerling TJ (2011a) Soil radium, soil gas radon and indoor radon empirical relationships to assist in post-closure impact assessment related to near-surface radioactive waste disposal. J Environ Radioact 102:221–234

Appleton JD, Doyle E, Fenton D, Organo C (2011b) Radon potential mapping of the Tralee-Castleisland and Cavan areas (Ireland) based on airborne gamma-ray spectrometry and geology. J Radiol Prot 31:221–235

Appleton JD, Miles JCH, Young M (2011c) Comparison of Northern Ireland radon maps based on indoor radon measurements and geology with maps derived by predictive modelling of airborne radiometric and ground permeability data. Sci Total Environ 409: 1572–1583

Avino R, Capaldi G, Pece R (1999) Radon in active volcanic areas of Southern Italy. Nuovo Cimento Della Societa Italiana Di Fisica C-Geophys Space Phys 22:379–385

Ball TK, Miles JCH (1993) Geological and geochemical factors affecting the radon concentration in homes in Cornwall and Devon, UK. Environ Geochem Health 15(1):27–36

Ball TK, Cameron DG, Colman TB, Roberts PD (1991) Behavior of radon in the geological environment – a review. Q J Eng Geol 24(2):169–182

Ball TK, Cameron DG, Colman TB (1992) Aspects of radon potential mapping in Britain. Radiat Prot Dosimetry 45:211–214

Barnet I (1994) Radon risk classification for building purposes in the Czech Republic. In: Barnet I, Neznal M (eds) Radon investigations in Czech Republic V. Czech Geological Survey, Prague, pp 18–24

Barnet I, Miksová J, Fojtíková I (2002) The GIS analysis of indoor radon and soil gas in major rock types of the Czech Republic. In: Barnet I, Neznal M, Mikšová J (eds) Radon investigations in the Czech Republic IX and the sixth international workshop on the geological aspects of radon risk mapping. Czech Geological Survey, Prague, pp 5–11

Barnet I, Pacherova P, Neznal M, Neznal M (2008) Radon in geological environment – Czech experience, special papers. Czech Geological Survey, Prague, p 70

Barnet I, Pacherová P, Preusse W, Stec B (2010) Cross-border radon index map 1: 100 000 Lausitz – Jizera – Karkonosze – Region (northern part of the Bohemian Massif). J Environ Radioact 101:809–812

Bossew P (2003) The radon emanation power of building materials, soils and rocks. Appl Radiat Isot 59:389–392

Bossew P (2009) Uncertainty of block estimates introduced by misallocation of point samples: on the example of spatial indoor radon data. J Environ Radioact 100:274–279

Bossew P (2010) Radon: exploring the log-normal mystery. J Environ Radioact 101:826–834

Bossew P, Lettner H (2007) Investigations on indoor radon in Austria, part 1: seasonality of indoor radon concentration. J Environ Radioact 98:329–345

Bossew P, Dubois G, Tollefsen T (2008) Investigations on indoor radon in Austria, part 2: geological classes as categorical external drift for spatial modelling of the radon potential. J Environ Radioact 99:81–97

BRE (1999) Radon: guidance on protective measures for new dwellings. Building research establishment report, BR 211
Burke O, Murphy P (2011) Regional variation of seasonal correction factors for indoor radon levels. Radiat Meas 46:1168–1172
CEC (Council of the European Community) (1998) Council directive 98/83/EC on the quality of water intended for human consumption. Off J Europ Comm, L 330/32 of 5.12.98
Chau ND, Chrusciel E, Prokolski L (2005) Factors controlling measurements of radon mass exhalation rate. J Environ Radioact 82:363–369
Chen J (2009) A preliminary design of a radon potential map for Canada: a multi-tier approach. Environ Earth Sci 59:775–782
Chen D, You X, Hu R (2005) Indoor radon survey in indoor environments in Zhuhai city, China. Radiat Meas 39:205–207
Chen J, Ford K, Whyte J, Bush K, Moir D, Cornett J (2011) Achievements and current activities of the Canadian radon program. Radiat Prot Dosimetry 146:14–18
Cheng JP, Guo QJ, Ren TS (2002) Radon levels in China. J Nucl Sci Technol 39:695–699
Ciotoli G, Guerra M, Lombardi S, Vittori E (1998) Soil gas survey for tracing seismogenic faults: a case study in the Fucino basin, central Italy. J Geophys Res-Solid Earth 103:23781–23794
Clavensjö B, Åkerblom G (1994) The radon book. The Swedish Council for Building Research, Stockholm
Cliff KD, Miles JCH (eds) (1997) Radon research in the European Union, EUR 17628. National Radiological Protection Board, Chilton, UK
Cohen BL (1997) Problems in the radon vs. Lung cancer test of the linear no-threshold theory and a procedure for resolving them. Health Phys 72:623–628
Cross FT (1992) A review of experimental animal radon health insights and implications. In: Dewey DC et al (eds) Radiation research, a twentieth century perspective. Academic, New York, pp 333–339
Darby SC et al (1995) Radon and cancers other than lung cancer in underground miners – a collaborative analysis of 11 studies. J Natl Cancer Inst 87(5):378–384
Darby SC, Whitely E, Silcocks P et al (1998) Risk of cancer associated with residential radon exposure in South-West England: a case–control study. Br J Cancer 78:394–408
Darby S, Hill D, Auvinen A, Barrios-Dios JM, Baysson H, Bochicchio F, Deo H, Falk R, Forastiere F, Hakama M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, Makelainen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Rosario AS, Tirmarche M, Tomasek L, Whitley E, Wichmann HE, Doll R (2005) Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case–control studies. Br Med J 330:223–226
Denman AR, Crockett RGM, Groves-Kirkby CJ, Phillips PS, Gillmore GK, Woolridge AC (2007) The value of seasonal correction factors in assessing the health risk from domestic radon—a case study in Northamptonshire, UK. Environ Int 33:34–44
Dubois G (2005) An overview of radon surveys in Europe. Office for Official Publications of the European Community, Luxembourg, p 168
Dubois G, Bossew P, Friedmann H (2007) A geostatistical autopsy of the Austrian indoor radon survey (1992–2002). Sci Total Environ 377:378–395
Dubois G, Bossew P, Tollefsen T, De Cort M (2010) First steps towards a European atlas of natural radiation: status of the European indoor radon map. J Environ Radioact 101:786–798
Duport P (2002) Is the radon risk overestimated? Neglected doses in the estimation of the risk of lung cancer in uranium underground miners. Radiat Prot Dosimetry 98(3):329–338
Duval JS, Otton JK (1990) Radium distribution and indoor radon in the Pacific Northwest. Geophys Res Lett 17(6):801–804
Fennell SG, Mackin GM, Madden JS, McGarry AT, Duffy JT, O'Colmáin M, Colgan PA, Pollard D (2002) Radon in dwellings the Irish national radon survey. Survey Radiological Protection Institute of Ireland, Dublin
Friedmann H, Gröller J (2010) An approach to improve the Austrian radon potential Map by Bayesian statistics. J Environ Radioact 101:804–808
Galán López M, Martín Sánchez A (2008) Present status of 222Rn in groundwater in Extremadura. J Environ Radioact 99:1539–1543
Ghosh D, Deb A, Sengupta R (2009) Anomalous radon emission as precursor of earthquake. J Appl Geophys 69:67–81
Gillmore GK, Sperrin M, Phillips P, Denman A (2000) Radon hazards, geology, and exposure of cave users: a case study and some theoretical perspectives. Ecotoxicol Environ Saf 46:279–288
Gillmore GK, Phillips P, Denman A, Sperrin M, Pearce G (2001) Radon levels in abandoned metalliferous mines, Devon, southwest England. Ecotoxicol Environ Saf 49:281–292
Gillmore GK, Phillips PS, Denman AR, Gilbertson DD (2002) Radon in the Creswell crags Permian limestone caves. J Environ Radioact 62:165–179
Gilmore GK, Phillips P, Denman A, Sperrin M, Pearce G (2001) Radon levels in abandoned metalliferous mines, Devon, Southwest England. Ecotoxicol Environ Saf 49(3):281–292
Grasty RL (1997) Radon emanation and soil moisture effects on airborne Gamma-Ray measurements. Geophysics 62(5):1379–1385
Gray A, Read S, McGale P, Darby S (2009) Lung cancer deaths from indoor radon and the cost effectiveness and potential of policies to reduce them. BMJ 338:a3110
Greeman DJ, Rose AW (1996) Factors controlling the emanation of radon and thoron in soils of the eastern U. S. A. Chem Geol 129: 1–14
Greeman DJ, Rose AW, Jester WA (1990) Form and behavior of radium, uranium, and thorium in central Pennsylvania soils derived from dolomite. Geophys Res Lett 17(6):833–836
Green BMR, Lomas PR, O'Riordan MC (1992) Radon in dwellings in England, NRPB-R254. National Radiological Protection Board, Chilton
Groves-Kirkby CJ, Denman AR, Phillips PS, Tornberg R, Woolridge AC, Crockett RGM (2008) Domestic radon remediation of U.K. dwellings by sub-slab depressurisation: evidence for a baseline contribution from constructional materials. Environ Int 34: 428–436
Gunby JA, Darby SC, Miles JCH, Green BMR, Cox DR (1993) Factors affecting indoor radon concentrations in the United Kingdom. Health Phys 64:2–12
Gundersen LCS, Schumann ER (1996) Mapping the radon potential of the United States: examples from the Appalachians. Environ Int 22(suppl 1):S829–S844
Gundersen LCS, Schumann ER, Otton JK, Dubief RF, Owen DE, Dickenson KE (1992) Geology of radon in the United States. In: Gates AE, Gundersen LCS (eds) Geologic controls on radon, Special paper 271. Geological Society America, Boulder, pp 1–16
Hishinuma T, Nishikawa T, Shimoyama T, Miyajima M, Tamagawa Y, Okabe S (1999) Emission of radon and thoron due to the fracture of rock. Nuovo Cimento Della Societa Italiana Di Fisica C-Geophys Space Phys 22:523–527
HPA (2009) Radon and public health: report prepared by the subgroup on radon epidemiology of the independent advisory group on ionising radiation. Documents of the Health Protection Agency. Health Protection Agency, UK
Hunter N, Muirhead CR, Miles JCH, Appleton JD (2009) Uncertainties in radon related to house-specific factors and proximity to geological boundaries in England. Radiat Prot Dosim 136:17–22
Hunter N, Muirhead CR, Miles JCH (2011) Two error components model for measurement error: application to radon in homes. J Environ Radioact 102:799–805
ICRP (2007) The 2007 recommendations of the international commission on radiological protection, Annals ICRP. ICRP, Oxford
ICRP (2009) International commission on radiological protection statement on radon. ICRP 37:2–4

ICRP (International Committee on Radiological Protection) (1993) Protection against radon-222 at home and at work. Ann ICRP 23(2):1–65, ICRP Publication 65

Ielsch G et al (2001a) Radon (Rn-222) level variations on a regional scale: influence of the basement trace element (U, Th) geochemistry on radon exhalation rates. J Environ Radioact 53(1):75–90

Ielsch G, Thieblemont D, Labed V, Richon P, Tymen G, Ferry C, Robe MC, Baubron JC, Bechennec F (2001b) Radon (Rn-222) level variations on a regional scale: influence of the basement trace element (U, Th) geochemistry on radon exhalation rates. J Environ Radioact 53:75–90

Ielsch G, Cushing ME, Combes P, Cuney M (2010) Mapping of the geogenic radon potential in France to improve radon risk management: methodology and first application to region Bourgogne. J Environ Radioact 101:813–820

Je HK, Kang CG, Chon HT (1999) A preliminary study on soil-gas radon geochemistry according to different bedrock geology in Korea. Environ Geochem Health 21(2):117–131

Jin Y, Wang Z, Lida T, Ikebe Y, Abe S, Chen H, Wu L, Zeng Q, Du K, Li S (1996) A new subnationwide survey of outdoor and indoor Rn-222 concentrations in China. Environ Int 22:S657–S663

Kemski J, Klingel R (2008) Prediction of indoor radon in Germany on a regional scale. Kerntechnik 73:131–137

Kemski J, Siehl A, Stegemann R, Valdivia-Manchego M (2001) Mapping the geogenic radon potential in Germany. Sci Total Environ 272(1–3):217–230

Kemski J, Klingel R, Siehl A, Valdivia-Manchego M (2009) From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ Geol 56:1269–1279

Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, Klotz JB, Letourneau EG, Lynch CF, Lyon JI, Sandler DP, Schoenberg JB, Steck DJ, Stolwijk JA, Weinberg C, Wilcox HB (2005) Residential radon and risk of lung cancer – a combined analysis of 7 north American case–control studies. Epidemiology 16:137–145

Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, Klotz JB, Letourneau EG, Lynch CF, Lyon JL, Sandler DP, Schoenberg JB, Steck DJ, Stolwijk JA, Weinberg C, Wilcox HB (2006) A combined analysis of North American case–control studies of residential radon and lung cancer. J Toxicol Environ Health-Part a-Curr Issue 69:533–597

Laurier D, Valenty M, Tirmarche M (2001) Radon exposure and the risk of leukemia: a review of epidemiological studies. Health Phys 81(3):272–288

Lawrence CE, Akber RA, Bollhöfer A, Martin P (2009) Radon-222 exhalation from open ground on and around a uranium mine in the wet-dry tropics. J Environ Radioact 100:1–8

Li XY, Zheng B, Wang Y, Wang X (2006) A survey of radon level in underground buildings in China. Environ Int 32:600–605

Li X, Song B, Zheng B, Wang Y, Wang X (2010) The distribution of radon in tunnels with different geological characteristics in China. J Environ Radioact 101:345–348

Lomas PR, Green BMR, Miles JCH, Kendall GM (1996) Radon Atlas of England, NRPB-290. National Radiological Protection Board, Chilton

Lubin JH (1998) On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen's ecologic regression. Health Phys 75(1):4–10

Lubin JH, Boice JD (1997) Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies. J Natl Cancer Inst 89(1):49–57

Lubin JH et al (1994) Radon exposure in residences and lung-cancer among women – combined analysis of 3 studies. Cancer Causes Control 5(2):114–128

Man CK, Yeung HS (1998) Radioactivity contents in building materials used in Hong Kong. J Radioanal Nucl Chem 232(1–2):219–222

Miksová J, Barnet I (2002) Geological support to the national radon programme (Czech Republic). Bull Czech Geol Surv 77(1):13–22

Miles JCH (1998) Mapping radon-prone areas by log-normal modelling of house radon data. Health Phys 74:370–378

Miles JCH (2001) Temporal variation of radon levels in houses and implications for radon measurement strategies. Radiat Prot Dosimetry 93:369–376

Miles JCH, Appleton JD (2000) Identification of localised areas of England where radon concentrations are most likely to have 5% probability of being above the Action Level. Department of the Environment, Transport and the Regions report, DETR/RAS/00.001. DETR, London

Miles JCH, Appleton JD (2005) Mapping variation in radon potential both between and within geological units. J Radiol Prot 25:257–276

Miles JCH, Ball TK (1996) Mapping radon-prone areas using house radon data and geological boundaries. Environ Int 22(suppl 1):779–782

Miles JCH, Appleton JD, Rees DM, Green BMR, Adlam KAM, Myers AH (2007) Indicative atlas of radon in England and Wales. HPA, Chilton

Moreno V, Baixeras C, Font L, Bach J (2008) Indoor radon levels and their dynamics in relation with the geological characteristics of La Garrotxa, Spain. Radiat Meas 43:1532–1540

Mudd GM (2008) Radon releases from Australian uranium mining and milling projects: assessing the UNSCEAR approach. J Environ Radioact 99:288–315

Mui KW, Wong LT, Hui PS (2008) An approach to assessing the probability of unsatisfactory radon in air-conditioned offices of Hong Kong. J Environ Radioact 99:248–259

Murphy P, Organo C (2008) A comparative study of lognormal, gamma and beta modelling in radon mapping with recommendations regarding bias, sample sizes and the treatment of outliers. J Radiol Prot 28:293–302

NAS (1998) Health effects of exposure to radon (BEIR VI). National Academy of Sciences, Washington, DC

Nazaroff WW (1988) Measurement techniques. In: Radon and its decay products in indoor air. Wiley, New York, pp 491–504

Neri M, Giammanco S, Ferrera E, Patanè G, Zanon V (2011) Spatial distribution of soil radon as a tool to recognize active faulting on an active volcano: the example of Mt. Etna (Italy). J Environ Radioact 102:863–870

NRPB (1989) Living with radiation. National Radiological Protection Board, Chilton

NRPB (2000) Health risks from radon. National Radiological Protection Board, Chilton

Organo C, Murphy P (2007) The Castleisland radon survey – follow-up to the discovery of a house with extremely high radon concentrations in county Kerry (SW Ireland). J Radiol Prot 27:275–285

Papastefanou C (2007) Measuring radon in soil gas and groundwaters: a review. Ann Geophys 50:569–578

Papastefanou C (2010) Variation of radon flux along active fault zones in association with earthquake occurrence. Radiat Meas 45:943–951

Pereira AJSC, Godinho MM, Neves LJPF (2010) On the influence of faulting on small-scale soil-gas radon variability: a case study in the Iberian uranium province. J Environ Radioact 101:875–882

Petersell V, Åkerblom G, Ek B-M, Enel M, Mõttus V, Täht K (2005) Radon risk Map of Estonia: explanatory text to the radon risk map set of Estonia at scale of 1:500 000. Swedish Radiation Protection Authority, Tallinn/Stockholm, p 76

Poffijn A, Goes E, Michaela I (2002) Investigation of the radon potential of an Alum deposit. In: Barnet I, Neznal M, Miksova J (eds) Radon investigations in the Czech Republic IX and the sixth international workshop on the geological aspects of radon risk mapping, Czech geological survey, Prague

Przylibski TA, Mamont-Ciesla K, Kusyk M, Dorda J, Kozlowska B (2004) Radon concentrations in groundwaters of the Polish part of the Sudety Mountains (SW Poland). J Environ Radioact 75: 193–209

NRPA (Nordic Radiation Protection Authorities) (2000) Naturally occurring radioactivity in the Nordic countries – recommendations. The Radiation Protection Authorities in Denmark, Finland, Iceland, Norway, and Sweden

Ramachandan TV, Sathish LA (2011) Nationwide indoor 222Rn and 220Rn map for India: a review. J Environ Radioact 102:975–986

Raspa G, Salvi F, Torri G (2010) Probability mapping of indoor radon-prone areas using disjunctive kriging. Radiat Prot Dosimetry 138:3–19

Rees DM, Bradley EJ, Green BMR (2011) Radon in homes in England and Wales: 2010 data review, HPA-CRCE-015. HPA, Chilton

Rose AW, Hutter AR, Washington JW (1990) Sampling variability of radon in soil gases. J Geochem Explor 38:173–191

Sakoda A, Ishimori Y, Hanamoto K, Kataoka T, Kawabe A, Yamaoka K (2010) Experimental and modeling studies of grain size and moisture content effects on radon emanation. Radiat Meas 45: 204–210

Scheib C, Appleton JD, Jones DJ, Hodgkinson E (2006) Airborne uranium data in support of radon potential mapping in Derbyshire, Central England. In: Barnet I, Neznal M, Pacherová P (eds) Radon investigations in the Czech Republic XI and the eighth international workshop on the geological aspects of radon risk mapping. Czech geological survey, Prague, pp 210–219

Scheib C, Appleton JD, Miles JCH, Green BMR, Barlow TS, Jones DG (2009) Geological controls on radon potential in Scotland. Scott J Geol 45:147–160

Singh S, Kumar A, Singh B (2002) Radon level in dwellings and its correlation with uranium and radium content in some areas of Himachal Pradesh, India. Environ Int 28(1–2):97–101

Skeppstrom K, Olofsson B (2006) A prediction method for radon in groundwater using GIS and multivariate statistics. Sci Total Environ 367:666–680

Smethurst MA, Strand T, Sundal AV, Rudjord AL (2008) Large-scale radon hazard evaluation in the Oslofjord region of Norway utilizing indoor radon concentrations, airborne gamma ray spectrometry and geological mapping. Sci Total Environ 407:379–393

Smith BJ, Field RW, Lynch CF (1998) Residential Rn-222 exposure and lung cancer: testing the linear no-threshold theory with ecologic data. Health Phys 75(1):11–17

Somlai J, Gorjanacz Z, Varhegyi A, Kovacs T (2006) Radon concentration in houses over a closed Hungarian uranium mine. Sci Total Environ 367:653–665

Song G, Wang X, Chen D, Chen Y (2011) Contribution of 222Rn-bearing water to indoor radon and indoor air quality assessment in hot spring hotels of Guangdong, China. J Environ Radioact 102:400–406

Song G, Chen D, Tang Z, Zhang Z, Xie W (2012) Natural radioactivity levels in topsoil from the Pearl River delta zone, Guangdong, China. J Environ Radioact 103:48–53

Sundal AV, Henriksen H, Soldal O, Strand T (2004) The influence of geological factors on indoor radon concentrations in Norway. Sci Total Environ 328:41–53

Swakon J, Kozak K, Paszkowski M, Gradzinski R, Loskiewicz J, Mazur J, Janik M, Bogacz xJ, Horwacik T, Olko P (2005) Radon concentration in soil gas around local disjunctive tectonic zones in the Krakow area. J Environ Radioact 78:137–149

Talbot DK, Hodkin DL, Ball TK (1997) Radon investigations for tunnelling projects; a case study from St. Helier, Jersey. Q J Eng Geol 30(Part 2):115–122

Tell I et al (1993) Indoor radon-daughter concentration and gamma-radiation in urban and rural homes on geologically varying ground. Sci Total Environ 128(2–3):191–203

Tuia D, Kanevski M (2008) Indoor radon distribution in Switzerland: lognormality and extreme value theory. J Environ Radioact 99:649–657

UNSCEAR (2009) United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). UNSCEAR 2006 report, annex E. Sources-to-effects assessment for radon in homes and workplaces. United Nations, New York

UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) (2000) Sources, effects, and risks of ionizing radiation. United Nations, New York

USDOE (United States Department of Energy) (1988) Radiation inhalation studies of animals, DOE/ER-0396, Washington, DC

Verdoya M, Chiozzi P, De Felice P, Pasquale V, Bochiolo M, Genovesi I (2009) Natural gamma-ray spectrometry as a tool for radiation dose and radon hazard modelling. Appl Radiat Isot 67:964–968

Vesterbacka P, Makelainen I, Arvela H (2005) Natural radioactivity in drinking water in private wells in Finland. Radiat Prot Dosimetry 113:223–232

Wang N, Xiao L, Li C, Huang Y, Pei S, Liu S, Xie F, Cheng Y (2005) Determination radioactivity level of 238U, 232Th and 40K in surface medium in Zhuhai city by in-situ gamma-ray spectrometry. J Nucl Sci Technol 42:888–896

Wang N, Xiao L, Li C, Mei W, Hang Y, Liu D (2009) Level of radon exhalation rate from soil in some sedimentary and granite areas of China. J Nucl Sci Technol 46:303–309

Wang N, Xiao L, Liu CK, Liu S, Huang Y, Liu DJ, Peng M (2011) Distribution and characteristics of radon gas in soil from a high-background-radiation city in China. J Nucl Sci Technol 48:751–758

Washington JW, Rose AW (1992) Temporal variability of radon concentrations in the interstitial gas of soils in Pennsylvania. J Geophys Res 97(B6):9145–9159

Watson SJ, Al J, Oatway WB, Hughes JS (2005) HPA-RPD-001 – ionising radiation exposure of the UK population: 2005 review. HPA, Chilton

Wattananikorn K, Emharuthai S, Wanaphongse P (2008) A feasibility study of geogenic indoor radon mapping from airborne radiometric survey in northern Thailand. Radiat Meas 43:85–90

WHO (1996) Guidelines for drinking-water quality, vol 2, 2nd edn. World Health Organization, Geneva

WHO (2009) WHO handbook on indoor radon: a public health perspective. WHO Press, Geneva

Wiegand J et al (2000) Radon and thoron in cave dwellings (Yan'an, China). Health Phys 78(4):438–444

Yamada Y, Sun QF, Tokonami S, Akiba S, Zhou WH, Hou CS, Zhang SZ, Ishikawa T, Furukawa M, Fukutsu K, Yonehara H (2006) Radon-thoron discriminative measurements in Gansu province, China, and their implication for dose estimates. J Toxicol Environ Health-Part a-Curr Issue 69:723–734

Zhu HC, Charlet JM, Poffijn A (2001) Radon risk mapping in southern Belgium: an application of geostatistical and GIS techniques. Sci Total Environ 272(1–3):203–210

Zhuo WH, Iida T, Yang XT (2001) Occurrence of Rn-222, Ra-226, Ra-228 and U in groundwater in Fujian Province, China. J Environ Radioact 53:111–120

Arsenic in Groundwater and the Environment

12

P.L. Smedley and David G. Kinniburgh

Contents

12.1 Introduction

Awareness of the problems associated with arsenic in drinking water and the environment has grown significantly over the last two decades or so and today an enormous literature exists documenting its occurrence, behaviour and impacts in many places across the globe. The mobilisation of arsenic in the environment occurs through a complex combination of natural biogeochemical reactions and human interactions. Most recognised problems are generated by mobilisation and transport under natural conditions, but mobilisation has also been caused, or exacerbated, by mining, fossil-fuel combustion and use of synthetic arsenical compounds (pesticides, herbicides, crop desiccants and arsenic-based additives in livestock feed). Arsenical pesticides and herbicides have been used much less over the last few decades, and more recent restrictions have been imposed on the use of arsenic in wood preservation (e.g. European Communities' Directive 2003/2/EC), but the legacy of such sources may still pose a localised threat to the environment.

Human exposure to arsenic occurs through a number of pathways, including air, food, water and soil. The relative impacts of these vary depending on local circumstances but of the potential pathways, drinking water poses one of the greatest threats to human health as borne out by the large number of documented case histories from around the world. The concentrations of arsenic in drinking water are very variable, depending on nature of source (surface water, groundwater, rainwater) and local conditions, and observed ranges vary over several orders of magnitude. Excepting localised sources of anthropogenic contamination, the highest aqueous arsenic concentrations are usually found in groundwaters because of the high solid/solution ratios found in aquifers. Groundwaters therefore pose the greatest overall threat to health. Groundwaters with arsenic concentrations sufficiently high to be detrimental to humans, or with already detectable health impacts have been reported in Argentina, Bangladesh, Burkina Faso, Cambodia, Chile, China, India,

P.L. Smedley (✉) • D.G. Kinniburgh
British Geological Survey, Oxfordshire, UK
e-mail: pls@bgs.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_12,

Hungary, Laos, Mexico, Nepal, Romania, Spain, Taiwan, Thailand and Vietnam and occasional problems are found in many other countries.

Some of the groundwater arsenic problems have been recognised for a considerable time. Probably the earliest cases of health effects from arsenic contamination of drinking water were recognised in a mining area of Poland in the 1890s. Here, contamination of water supplies by oxidation of arsenic-bearing sulphide minerals produced localised health problems (see Tseng et al. 1968). In central Argentina, arsenic-related health problems were first documented in 1917 and problems in Taiwan were also first identified in the 1910s. In each of these cases, solutions to the problems were sought primarily through engineering which for the most part provided alternative supplies of surface water or of treated water. As a result, geochemical investigations into the processes controlling the arsenic mobilisation in these areas were not carried out until much more recently and awareness of the problems observed in impacted regions such as the Bengal Basin, Cambodia and Vietnam have developed many years later.

European and US EPA regulations implement the current recommended WHO provisional guideline value for arsenic in drinking water of 10 μg L^{-1}. However, many countries continue to use the WHO pre-1993 guideline value of 50 μg L^{-1} as their national standard, in part because of the difficulties of testing low concentrations and in part due to difficulties with compliance. Concentrations of 10 μg L^{-1} and 50 μg L^{-1} are therefore both still used as reference points for the testing and reporting of arsenic.

This chapter reports the current state of knowledge on the occurrence of arsenic in natural waters and their host rocks across the world and attempts to describe what are currently understood to be the main biogeochemical processes that control its mobilisation in the environment (see also Chaps. 2, 24, and 25, this volume).

12.2 Sources of Arsenic in the Natural Environment

12.2.1 Minerals

Arsenic occurs as a major constituent in more than 200 minerals, including elemental arsenic, arsenides, sulphides, oxides, arsenates and arsenites. Most are ore minerals or their alteration products. These minerals are relatively rare in the natural environment. Among the most common occurrences in ore zones are arsenian pyrite ($Fe(S,As)_2$), arsenopyrite (FeAsS), löllingite ($FeAs_2$), realgar (AsS), orpiment (As_2S_3), cobaltite (CoAsS), niccolite (NiAs) and scorodite ($FeAsO_4.2H_2O$). Arsenian pyrite ($Fe(As,S)_2$) is probably the most important source of arsenic in ore zones (Deditius et al. 2008; Savage et al. 2000) and concentrations in excess of 10 wt.% (up to 19 wt.%) As have been reported (e.g. Barker et al. 2009; Blanchard et al. 2007; Reich and Becker 2006). Reich and Becker (2006) concluded from first-principles calculations that pyrite and marcasite polymorphs can contain up to 6 wt.% As in solid solution, beyond which unmixing to separate pyrite/marcasite and arsenopyrite occurs. The oxidation state of As in arsenian pyrite is most typically As^{1-} in substitution for S, although recent occurrences of As^{3+}-bearing pyrite ($(Fe,As)S_2$) have been documented in some hydrothermal gold deposits, signalling precipitation in more oxidising conditions than in the more common As^{1-} form (Chouinard et al. 2005; Deditius et al. 2008).

Other rock-forming sulphide minerals, including chalcopyrite and galena, also contain often high concentrations of arsenic (Table 12.1) as well as a range of other trace elements (transition metals, Cd, Pb, Ag, Au, Sb, P, W and Mo). Zonation of trace elements within sulphide grains is common.

Table 12.1 Typical ranges of arsenic concentrations in rock-forming and accessory minerals

Mineral	Arsenic concentration range (mg kg^{-1})
Sulphide minerals	
Pyrite	100–130,000
Pyrrhotite	5–100
Marcasite	20–126,000
Galena	5–10,000
Sphalerite	5–17,000
Chalcopyrite	10–5,000
Oxide minerals	
Haematite	up to 160
Fe(III) oxyhydroxide	up to 76,000
Magnetite	2.7–41
Ilmenite	<1
Silicate minerals	
Quartz	0.4–1.3
Feldspar	<0.1–2.1
Biotite	1.4
Amphibole	1.1–2.3
Olivine	0.08–0.17
Pyroxene	0.05–0.8
Carbonate minerals	
Calcite	1–260
Dolomite	<3
Siderite	<3
Sulphate minerals	
Gypsum/anhydrite	<1–6
Barite	<1–12
Jarosite	34–2,000
Other minerals	
Apatite	<1–1,000
Halite	<3–30
Fluorite	<2

Barker et al. (2009), Di Benedetto et al. (2006), Savage et al. (2005), Smedley and Kinniburgh (2002) and references therein

Besides being an important component of ore bodies, pyrite is also formed in low-temperature sedimentary environments under sulphide-reducing conditions. Such authigenic pyrite plays an important role in present-day geochemical cycles, being present in the sediments of rivers, lakes, oceans and aquifers. Authigenic realgar has also been observed in shallow aquifer sediments (O'Day et al. 2004). In sulphidic solutions, soluble arsenic will likely be incorporated into pyrite through surface arsenite adsorption (Bostick and Fendorf 2003). Pyrite is unstable in aerobic systems and oxidises to form iron oxides with release of associated arsenic. A lower stability of arsenian pyrite compared to low-As pyrite also potentially favours As release (Blanchard et al. 2007; Savage et al. 2000). The presence of pyrite in mineralised veins is responsible for the production of acid mine drainage, and for the common presence of arsenic problems around coal mines and areas of intensive coal burning.

Arsenic is also found in large concentrations in many oxide minerals and hydrous metal oxides, either as part of the mineral structure or as adsorbed species. Concentrations in Fe oxides can reach several weight percent (Table 12.1), particularly where they form as the oxidation products of primary As-bearing iron sulphide minerals. Adsorption of arsenic to hydrous iron oxides is particularly strong, depending on redox speciation, pH, solute compositions and iron-oxide form; adsorbed loadings can be great even where concentrations in solution are low. Adsorption to hydrous Al and Mn oxides may also be important if present in quantity (Goldberg 1986; Stollenwerk 2003). Sorption to the edges of clays may occur, although the loadings are much smaller on a weight basis than for the iron oxides (Smedley and Kinniburgh 2002).

Concentrations in phosphate minerals can be relatively high, for example up to 1,000 mg kg^{-1} in apatite (Table 12.1). However, phosphate minerals are much less abundant than oxide minerals and so make a correspondingly small contribution to the arsenic concentration in most sediments. Arsenic tends to be present at much smaller concentrations in the common rock-forming minerals, although low concentrations are invariably present. Most common silicate minerals (including quartz, feldspar, micas, amphiboles) contain around 1 mg kg^{-1} or less (Table 12.1). Concentrations of arsenic in carbonate minerals are also usually low, of the order of a few mg kg^{-1}, although Di Benedetto et al. (2006) found up to 257 mg kg^{-1} in Italian travertines.

12.2.2 Rocks, Sediments and Soils

Arsenic occurs ubiquitously but at variable concentrations in rocks, unconsolidated sediments and soils. Average crustal abundance is around 1.5 mg kg^{-1}. Igneous rocks usually have concentrations of around 5 mg/kg or less. Volcanic glasses have on average around 6 mg kg^{-1} (Smedley and Kinniburgh 2002). Metamorphic rocks tend to have concentrations that reflect those of their igneous and sedimentary precursors. Most also contain around 5 mg kg^{-1} or less. Pelitic rocks (slates, phyllites) typically have the highest concentrations (Table 12.2), with on average of around 18 mg kg^{-1} (see also Chap. 2, this volume).

Concentrations in sedimentary rocks are typically in the range 5–10 mg kg^{-1} (Webster 1999). Average sediments are enriched in arsenic relative to igneous and metamorphic rocks because they contain greater quantities of minerals with high adsorbed arsenic loads. Of the sediments, sands and sandstones have lower concentrations which reflect the dominance of low-arsenic minerals (i.e. quartz and feldspars). Average sandstone arsenic concentrations are around 4 mg kg^{-1}.

Argillaceous deposits have a broader range of concentrations than sandstones (Table 12.2), being on average around 13 mg kg^{-1}. Higher values reflect the larger proportion of sulphide minerals, oxides, organic matter and clays present. Black shales have concentrations at the upper end of the range, principally because of their enhanced pyrite content (Table 12.2). Marine clays appear to have higher concentrations than non-marine equivalents. This may also be due to a higher proportion of pyrite in offshore pelagic sediments.

German organic-rich shales have arsenic concentrations of 100–900 mg kg^{-1} (Riedel and Eikmann 1986). Some coal samples have very high concentrations, up to 35,000 mg kg^{-1} (Belkin et al. 2000) (Table 12.2), although lower concentrations of 2.5–17 mg kg^{-1} are more typical. Some of the highest observed arsenic concentrations are found in ironstones and Fe-rich rocks. Concentrations of arsenic up to 2,900 mg kg^{-1} have been reported for such rocks (Table 12.2, references in Smedley and Kinniburgh 2002). Phosphorites are also relatively enriched in arsenic, with concentrations up to ca. 400 mg kg^{-1}. Data for carbonate rock types usually indicate relatively low arsenic concentrations (Tables 12.1 and 12.2).

Unconsolidated sediments have compositions which do not differ significantly from their indurated equivalents. Much evidence indicates that muds and clays have higher concentrations than most sands and carbonates. Values are typically 3–10 mg kg^{-1}, depending on texture and mineralogy (Table 12.2). High concentrations reflect abundance of pyrite or iron oxides. There is often a significant positive correlation between the iron and arsenic concentrations in sediments. High concentrations are also common in mineralised areas. Placer deposits in streams can have concentrations reflecting the abundance of sulphide minerals. Average arsenic concentrations for stream sediments in England and Wales have been reported in the range 5–8 mg kg^{-1} (Table 12.2, references in Smedley and Kinniburgh 2002), although a higher median value (13.4 mg kg^{-1}) has been computed

Table 12.2 Typical ranges of arsenic concentrations in rocks, sediments, soils and other superficial deposits

Rock/sediment type	As concentration range (mg kg^{-1})
Igneous rocks	
Ultrabasic rocks	0.03–15.8
Basic rocks	0.06–113
Intermediate rocks	0.09–13.4
Acidic rocks	0.2–15
Metamorphic rocks	
Quartzite	2.2–7.6
Hornfels	0.7–11
Phyllite/slate	0.5–143
Schist/gneiss	<0.1–18.5
Amphibolite and greenstone	0.4–45
Sedimentary rocks	
Marine shale/mudstone	3–15 (up to 490)
Shale (Mid-Atlantic Ridge)	48–361
Non-marine shale/mudstone	3.0–12
Sandstone	0.6–120
Limestone/dolomite	0.1–20.1
Phosphorite	0.4–188
Iron formations and Fe-rich sediment	1–2,900
Evaporites (gypsum/anhydrite)	0.1–10
Coals	0.3–35,000
Unconsolidated sediments	
Alluvial sand (Bangladesh)	1.0–6.2
Alluvial mud/clay (Bangladesh)	2.7–14.7
River bed sediments (Bangladesh)	1.2–5.9
Lake sediments	0.5–44
Glacial till	1.9–170
World average river sediments	5
Stream and lake silt	<1–72
Stream sediment, England and Wales	5–8
Loess silts, Argentina	5.4–18
Continental margin sediments	2.3–8.2
Soils	
Mixed soils	0.1–55
Peaty and bog soils	2–36
Peat	up to 9
Acid sulphate soils	1.5–45
Soils near sulphide deposits	2–8,000
Contaminated superficial deposits	
Mining-contaminated lake sediment	80–1,104
Mining-contaminated reservoir sediment	100–800
Mine tailings	396–2,000
Soils and tailings-contaminated soil	120–52,600
Industrially polluted inter-tidal sediments	0.38–1,260
Orchard soils	366–732
Soils below chemicals factory	1.3–4,770

From sources summarised in Smedley and Kinniburgh (2002)

from a more recent survey in eastern England (BGS data, unpublished). In the Bengal Basin, river sediments with concentrations averaging 2.0, 2.8 and 3.5 mg kg^{-1} were found for samples from the Ganges, Brahmaputra and Meghna rivers respectively (Datta and Subramanian 1997).

Cook et al. (1995) found concentrations in Canadian lake sediments ranging between 0.9 and 44 mg kg^{-1} but noted that the highest concentrations were present up to a few kilometres down-slope of mineralised areas. They also found concentrations in glacial till of 1.9–170 mg kg^{-1}

(Table 12.2) and noted the highest concentrations down-ice of mineralised areas.

In soils, concentrations of arsenic are of the order of 5–15 mg kg^{-1}, with an average of 7.2 mg kg^{-1} quoted for world soils and 7.4 mg kg^{-1} for American soils (Table 12.2, references in Smedley and Kinniburgh 2002). Peats and bog soils can have higher concentrations (average 13 mg kg^{-1}; Table 12.2), but this is principally because of an increased prevalence of sulphide mineral phases under the reduced conditions. Shotyk (1996) found up to 9 mg As kg^{-1} in peat profiles and in a profile with lower mineral content, i.e. purer peat, the As content was 1 mg kg^{-1} or lower.

Acid sulphate soils can be generated by the oxidation of pyrite in sulphide-rich terrains including mineralised areas and dewatered mangrove swamps and as such can be relatively enriched in arsenic. Acid sulphate soils from Canada have arsenic concentrations up to 45 mg kg^{-1} (Dudas 1984). Gustafsson and Tin (1994) found similarly high concentrations (up to 41 mg kg^{-1}) in acid sulphate soils from the Mekong delta of Vietnam.

Additional arsenic inputs to soils may be derived locally from industrial sources such as smelting and fossil-fuel combustion products and agricultural sources such as pesticides and phosphate fertilisers. Concentrations in the range 366–732 mg kg^{-1} have been found in orchard soils following historical application of arsenical pesticides to fruit crops (Table 12.2). Long-term use of phosphate fertilisers may also add arsenic to soil.

Concentrations of arsenic in sediments and soils contaminated by the products of mining activity, including tailings and effluent, can have arsenic concentrations orders of magnitude higher than under natural conditions. Tailings piles and tailings-contaminated soils can contain up to several thousands of mg kg^{-1} (Table 12.2). The values reflect not only increased abundance of primary arsenic-rich sulphide minerals, but also secondary iron arsenates and iron oxides formed as reaction products of the original ore minerals. The primary sulphide minerals are susceptible to oxidation in the tailings pile and the secondary minerals have varying solubility under oxidising conditions in groundwaters and surface waters. Scorodite ($FeAsO_4.2H_2O$) is a common sulphide oxidation product and its solubility is likely to control arsenic concentrations in such environments. Secondary arsenolite (As_2O_3) is also often represented in such environments.

12.3 Arsenic in Groundwater

12.3.1 Aqueous Speciation

Analytical capability for aqueous arsenic species has developed rapidly and modern automated hyphenated chromatographic techniques (e.g. Bednar et al. 2004) have improved on traditional methods using hydride generation, repeat measurements and determination of species by difference. Sensitive detection can be achieved with methods such as ICP-MS and this coupled with field-based (solid-phase extraction) or laboratory-based (e.g. HPLC) techniques allows for the direct measurement of aqueous organic forms alongside the more abundant inorganic As(III) and As(V). Field-based measurements counteract potential problems with instability in sample storage although significant problems with matrix effects using solid-phase extraction (Bednar et al. 2004) mean that this technique can be unsuitable for analysis of e.g. SO_4-rich acid minewaters, saline geothermal waters or alkaline, SO_4- and oxyanion-rich groundwaters, at least without large sample dilutions.

Arsenic can occur in the environment in a number of oxidation states (−3, −1, 0, +3 and +5) but in natural waters is found principally in inorganic form as arsenite (As(III)) or arsenate (As(V)), depending on redox conditions. Arsenic can form aqueous complexes with carbonate ligands, and As(III) with reduced sulphur (thioarsenites), which may be significant in some groundwaters.

Organic arsenic species produced by microbiological methylation reactions may be present in surface waters and may even dominate in organic-rich soil solutions, but are rarely present in quantity in groundwaters (Bednar et al. 2004; Chen et al. 1995; Das et al. 1995; Del Razo et al. 1990; Kondo et al. 1999; Lin et al. 1998). Exceptions potentially occur locally in cases of industrial or agricultural pollution. In practice concentrations of organic compounds in soil solutions rarely exceed a few micrograms per litre (Huang and Matzner 2006).

The ratio of As(III) to As(V) in groundwaters depends upon the abundance of the redox-active solids, including organic carbon and iron/manganese oxide, the flux of potential oxidants (oxygen, nitrate and sulphate) and on microbial activity. In geothermal systems, As(III) usually forms the main solute component but this can oxidise rapidly in surface conditions when mixed with surface water and especially when facilitated by bacteria (Wilkie and Hering 1998). Arsenic(III) also dominates in strongly reducing aquifers, typified by Fe(III) and sulphate reduction, as expected from the redox sequence (Berg et al. 2007; BGS and DPHE 2001; Chen et al. 1994; Das et al. 1995; Postma et al. 2007; Smedley et al. 2003; Stollenwerk et al. 2007; Tandukar et al. 2005; van Geen et al. 2006; Zheng et al. 2005). By contrast, in oxidising systems, solute arsenic is dominated by As(V) (Altamirano Espinoza and Bundschuh 2009; Bhattacharya et al. 2005; Del Razo et al. 1990; Robertson 1989; Smedley et al. 2002, 2007). The strong oxidising capacity of Mn oxides has been well-documented and implicated in arsenic redox speciation in several systems (Amirbahman et al. 2006; Oscarson et al. 1981a, b; Ying et al. 2011).

In natural systems, thermodynamic equilibrium in As species is often not achieved although abundant evidence indicates that reaction kinetics can be accelerated significantly by microbial activity, which have been shown to both oxidise

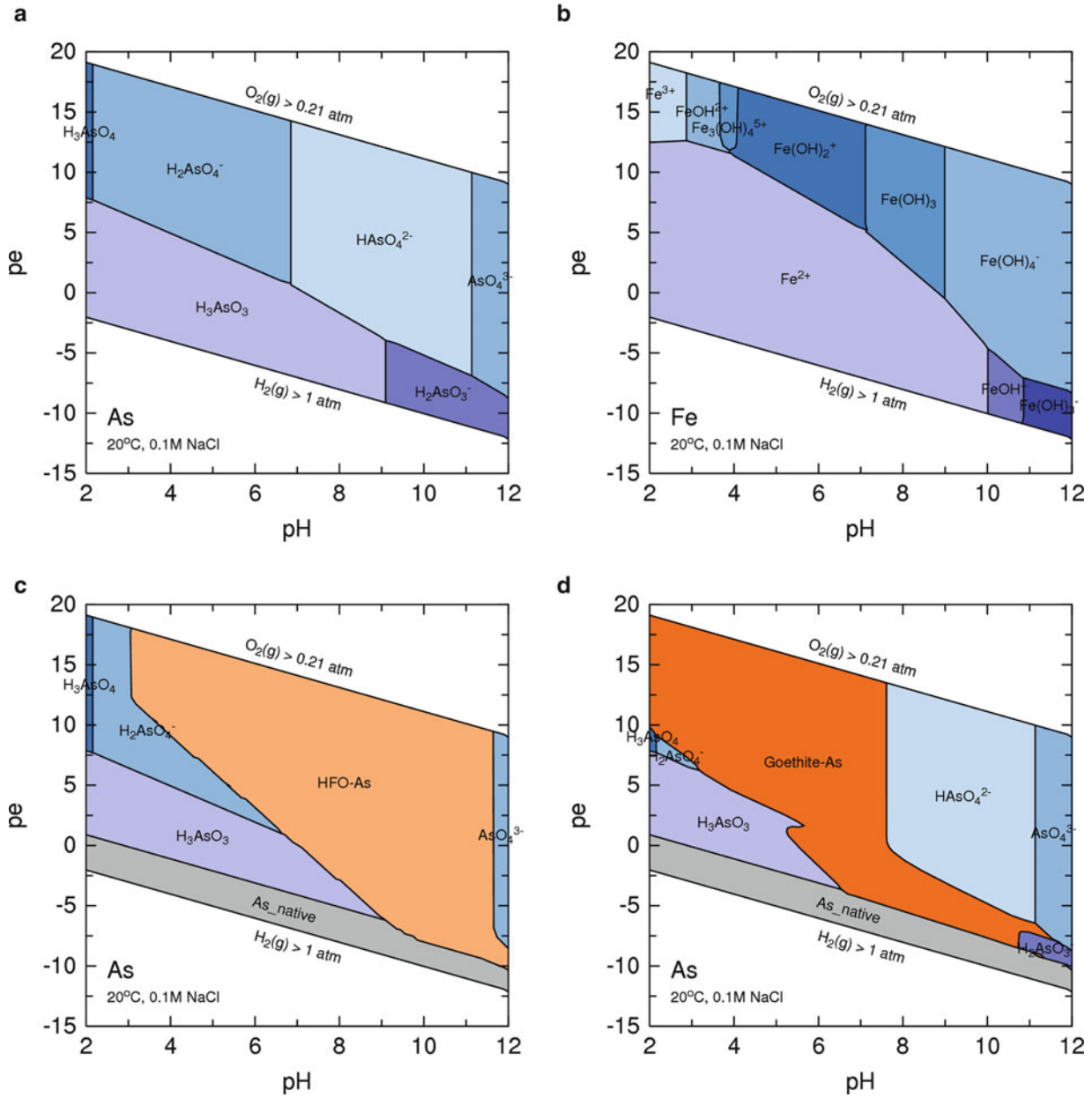


Fig. 12.1 pe-pH predominance diagrams for the As-Fe system. (**a**) Aqueous As species only; (**b**) aqueous Fe species only; (**c**) As species where hydrous ferric oxide (*HFO*) has been allowed to precipitate when stable and sorb As, and (**d**) the same as (**c**) except that goethite has been assumed to be the stable iron-oxide mineral. Adsorbed species concentrations have been calculated using the CD-MUSIC model. Where appropriate, total concentrations of Fe = 3e − 3 M and As = 1e − 4 M. A background concentration of 0.1 M NaCl was assumed in all cases. No reduced Fe minerals such as magnetite have been allowed to precipitate. All calculations were made by Phreeqc and plotted using PhreePlot

arsenite (Hering and Kneebone 2002; Wilkie and Hering 1998) and reduce arsenate (Oremland and Stolz 2003).

Sorption plays a key role in arsenic speciation and iron oxides in particular have a strong control on aqueous concentrations. Figure 12.1 shows pe-pH predominance diagrams for the As-Fe system. Figure 12.1a, b both show the redox and pH sensitivity of both aqueous As and Fe, with a range of protonated and hydrolysed species present over the natural pH range. Iron tends to be insoluble in oxic, near-neutral pH systems and forms a variety of iron oxides depending on the precise conditions and their evolution. Hydrous ferric oxide (HFO), a high-surface-area form of iron oxide, often forms when Fe is precipitated rapidly. This oxide is able to adsorb As on its surface and the adsorbed As (HFO-As) can become the dominant form of As (Fig. 12.1c). HFO is subject to both acid dissolution at low pH and reductive dissolution at low pe (Fig. 12.1c) which result in the release of As to solution. At high pH, As is desorbed from HFO due to increasing electrostatic repulsion of the negatively charged oxide surface and the AsO_4^{3-} species, again leading to higher concentrations in solution. HFO tends to transform slowly to more stable

forms of iron oxide with lower specific surface area, such as goethite, with time. Goethite still has a significant adsorption affinity for As, albeit somewhat different from HFO due to the different crystal structure. Goethite is less prone than HFO to both acid dissolution and reductive dissolution and so the Goethite-As field occupies a correspondingly large area on the predominance diagram at low pH and low pe (Fig. 12.1d). However, the lower specific surface area of goethite compared with HFO means that adsorption is not so great at high pH and so in goethite-dominated systems the aqueous As species tend to predominate at high pH.

Although these predominance diagrams illustrate some important geochemical interactions of Fe and As, they by no means tell the whole story. There are many other important interactions that have been omitted from these diagrams. For example, reduced iron oxides such as magnetite can form at low pe and these too will tend to adsorb As (Dixit and Hering 2003). Also, significant concentrations of sulphide may be present in reduced systems leading to the potential formation of pyrite and other Fe-S minerals, again with scope for the adsorption of As on the mineral surface or the incorporation of As into the mineral structure as a solid solution. Nonetheless, the diagrams go some way to explaining observed occurrences of high-arsenic groundwater in some strongly reducing aquifers and at extremes of pH.

12.3.2 Arsenic Abundance and Distribution in Groundwater

Aquifers have high solid:solution ratios, typically 3–20 kg L^{-1} and as a result, groundwaters within them are especially vulnerable to water-rock interaction and arsenic release. In addition, aquifers more often have the physical and chemical conditions favourable for arsenic mobilisation and transport than is the case in surface waters. Despite this, the occurrence of high arsenic concentrations in groundwaters is the exception rather than the rule. Background concentrations in groundwater are in most countries less than 10 μg L^{-1} and often substantially lower. Nonetheless, values quoted in the literature show a very large range from <0.5 to 15,000 μg L^{-1}, i.e. more than four orders of magnitude.

Mobilisation of arsenic in solution is favoured especially under oxidising conditions at high pH and under strongly reducing conditions. Evaporative concentration can also increase arsenic (and other element) concentrations substantially and may be important in some arid areas. Additional arsenic problems are encountered in some geothermal areas and in many areas of sulphide mineralisation and mining. Documented cases of industrial arsenic pollution (including those from agriculture) also exist. While these may be severe locally, occurrences are relatively rare and can usually be anticipated.

Investigations worldwide have revealed a number of major aquifers with significant groundwater arsenic problems. The hydrogeological and geochemical conditions in these affected aquifers vary, although some common unifying features are apparent. Particular areas of risk are large alluvial and delta plains and sedimentary inland basins, the latter especially in arid and semi-arid areas. Geologically young (Quaternary) aquifers are particularly prone to developing and preserving high-arsenic groundwater. Alluvial and delta plains with recognised groundwater arsenic problems include the Bengal Basin (Bangladesh, India), Mekong Valley (Cambodia, Laos, Vietnam), Yellow River Plain (China), Terai (Nepal), Red River delta (Vietnam), Chianan and Lanyang Plains (Taiwan) and Great Hungarian Plain (Hungary and Romania). Those from inland and closed basins include the Chaco-Pampean Plain (Argentina), and other plains of Chile, Mexico and south-western USA (Fig. 12.2). In many of these areas, significant numbers of wells have arsenic concentrations of several hundreds of micrograms per litre, with occasional sources in the milligram-per-litre range. Arsenic-related health problems have been recognised in the resident populations in numerous areas. Reconnaissance surveys of groundwater quality in other areas such as parts of Burma and Pakistan have also revealed small numbers of wells with concentrations of arsenic above drinking-water limits. Documentation of the affected aquifers of Burma is so far limited. The regions of the world having major aquifers with recognised arsenic problems are outlined below and categorised in terms of their environmental conditions. Distributions of arsenic in the environment related to geothermal activity as well as mining and mineralisation are also described.

12.3.3 Alluvial Plains and Deltas

12.3.3.1 Bengal Basin (Bangladesh, India)

The change from use of surface water or dug-well water to groundwater accessed by hand-pumped tubewells began in India and Bangladesh during the 1960s and accelerated significantly from the 1980s onwards. This led to a vast increase in the access of rural populations to what was considered a superior and safe source of drinking water from the readily available groundwater resources contained in the shallow aquifers. The discovery in 1983 in West Bengal (Garai et al. 1984) and in 1993 in Bangladesh (BGS and DPHE 2001) of problems with arsenic in the groundwater changed the situation significantly. Access of rural populations to improved supplies of drinking-water was down from the 90% coverage in the early 1990s to 68% in 2004 (Rosenboom 2004). The discoveries led to the instigation of major programmes of well testing and resource assessment, health surveillance, water-supply mitigation, and research.

Concentrations of arsenic in the groundwaters of the Bengal Basin span some four orders of magnitude, with occasional extremes above 1 mg L^{-1}. Recognised health

Fig. 12.2 Occurrence of documented arsenic problems in groundwater (arsenic >50 μg L^{-1}) in major aquifers and environmental problems related to mining and geothermal sources

problems include skin disorders and internal cancers. The high arsenic concentrations occur in groundwaters within unconsolidated grey micaceous alluvial and deltaic Holocene sands and silts associated with the Ganges, Brahmaputra and Meghna river systems. The presence of poorly-permeable surficial overbank deposits (BGS and DPHE 2001; Chakraborti et al. 2001; McArthur et al. 2004; van Geen et al. 2008), particularly in the lower sections of the basin, together with both solid organic matter and DOC in the groundwater, has resulted in development of strongly reducing aquifer conditions, aided by microbial reactions. Groundwater compositions are typified by high concentrations of dissolved Fe, Mn, HCO_3, NH_4, P and DOC as well as high As, and low concentrations of NO_3 and SO_4, though the correlations between these parameters are often poor. Methane has also been detected in some groundwaters (Ahmed et al. 1998). Many studies have observed that dissolved arsenic concentrations peak in the typical depth range 15–50 m (BGS and DPHE 2001; Harvey et al. 2002; Klump et al. 2006; McArthur et al. 2004; Zheng et al. 2005), although high concentrations (>50 μg L^{-1}) are observed at depths up to about 150 m below surface.

A national survey of arsenic in groundwater (BGS and DPHE 2001), using some 3,500 groundwater samples, found that 27% of samples from the Holocene shallow aquifer (<150 m depth) contained arsenic at concentrations exceeding 50 μg L^{-1}, and 46% exceeded 10 μg L^{-1} (Fig. 12.3). Subsequent arsenic analyses from the Bangladesh Arsenic Mitigation and Water Supply Program (BAMWSP 2005) showed that of almost five million tube wells tested nationally by the mid 2000s, some 30% had arsenic concentrations above 50 μg L^{-1}. The results from these two datasets compare reasonably well. Each demonstrates a very variable distribution of arsenic regionally across Bangladesh, with the greatest proportion of exceedances occurring in groundwaters from the southern and south-eastern parts of the country.

A UNICEF/DPHE survey of groundwaters from southern Bangladesh (Rosenboom 2004), found of some 317,000 tubewells tested, 66% contained arsenic with concentrations above 50 μg L^{-1}; 10% had <10 μg L^{-1}. Van Geen et al. (2003b) also found, from a survey of 6,000 tube wells in Araihazar, south-central Bangladesh, that some 75% of tube wells (depth range 15–30 m) had arsenic concentrations above 50 μg L^{-1}. The higher percentage exceedances of these studies reflect their location in some of the worst-affected areas.

The BGS and DPHE (2001) study estimated, on the basis of the population at the time (125 million estimate), that up to 35 million people were drinking groundwater with arsenic concentrations above 50 μg L^{-1}, and 57 million were drinking water with concentrations above 10 μg L^{-1}. This represents by far the most severely affected population recognised globally. Figures will have been modified somewhat by subsequent mitigation efforts although this is to be offset by population growth (161 million 2012 estimate).

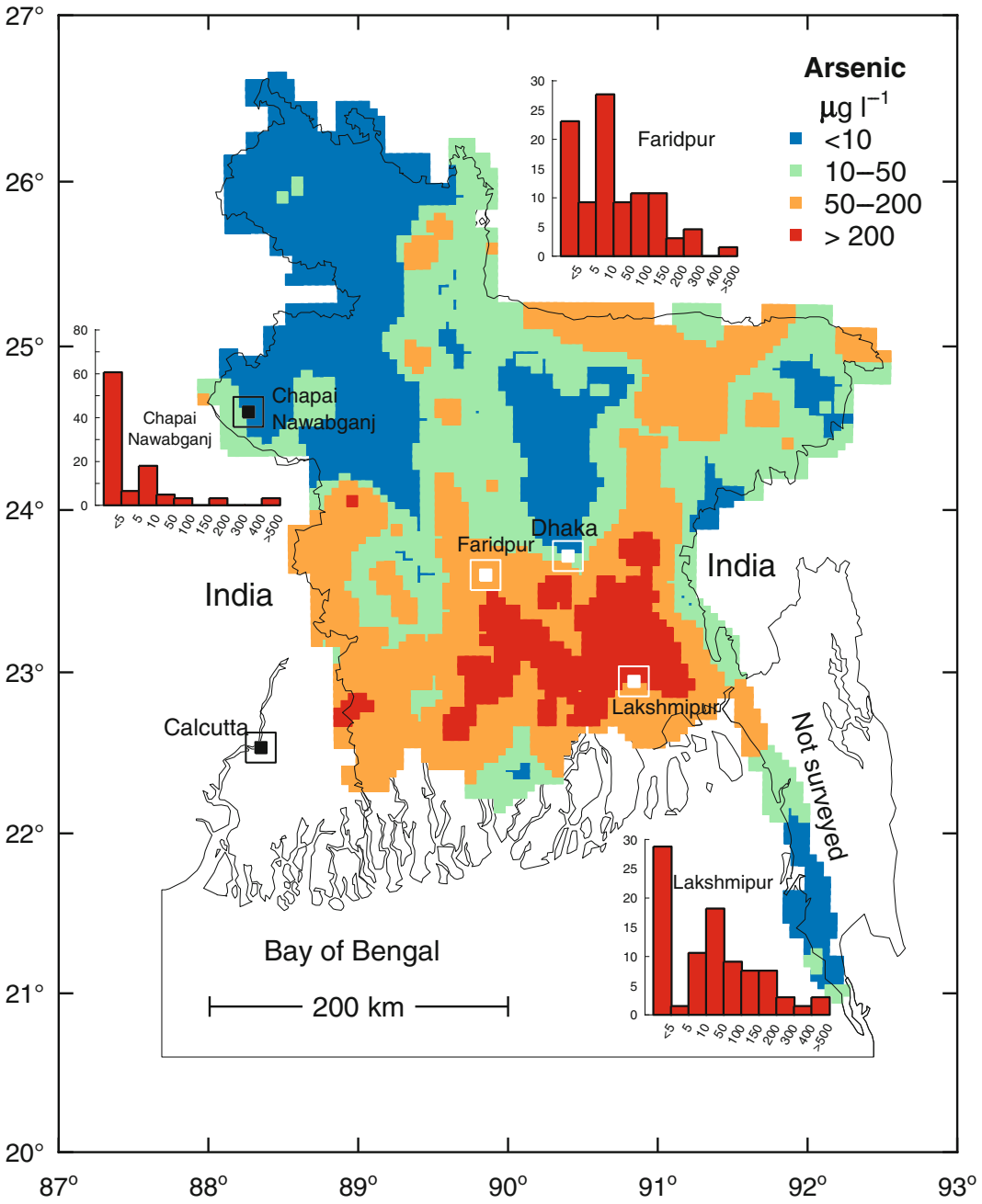


Fig. 12.3 Smoothed contour map of arsenic concentrations in shallow groundwater (<150 m) from Bangladesh and histograms of the distributions of arsenic in three special study areas (BGS and DPHE 2001). Low concentrations indicated in *blue* in northern Bangladesh denote the locations of the Tista Fan, and Madhupur and Barind Tracts

In India, the first evidence of arsenic-related health problems emerged in West Bengal (Bhattacharya et al. 1997; Chowdhury et al. 1997; Das et al. 1994), where it is estimated that about 6.5 million people are drinking water with arsenic concentrations greater than 50 µg L^{-1} (Mukherjee et al. 2006). High-arsenic groundwaters have subsequently also been identified in alluvial aquifers in the states of Bihar, Tripura, Uttar Pradesh, Jharkhand and Assam (Chakraborti et al. 2003; Mukherjee et al. 2006). Sedimentary aquifers potentially at risk from high arsenic concentrations in groundwater also occur in the neighbouring states of Meghalaya and Mizoram. In 2003, the Indian Bureau of Standards revised the national limit for arsenic in drinking water from 50 to 10 µg L^{-1}, in recognition of the severity of the problem and response to changing international regulations.

High-arsenic groundwaters have been found in up to nine districts of West Bengal, the five worst affected being Malda, Murshidabad, Nadia, 24 North Parganas and 24

South Parganas (Chakraborti et al. 2004; McArthur et al. 2004; Rahman et al. 2001). These cover up to 38 000 km^2. High arsenic concentrations occur typically in the depth range 10–80 m (McArthur et al. 2004). Chakraborti et al. (2001) found from 90,000 water analyses in the Holocene aquifers of West Bengal that some 34% exceeded 50 μg L^{-1} and 55% exceeded 10 μg L^{-1}.

In both Bangladesh and West Bengal, groundwaters from older Pleistocene aquifers usually have low arsenic concentrations (<10 μg L^{-1}). These sediments, dominantly orange-brown and more oxic than the overlying Holocene deposits, occur at variable depths. BGS and DPHE (2001) distinguished the aquifers using a 150-m depth cut-off, although van Geen et al. (2003b) documented Pleistocene deposits in the Araihazar area of Bangladesh at depths as shallow as 30 m (range 30–120 m). Pleistocene deposits also occur at surface within the distinct structural units of the Madhupur and Barind Tracts of Bangladesh (BGS and DPHE 2001) and give rise to the low mapped arsenic concentrations in shallow groundwaters there (Fig. 12.3).

Dug wells also usually have low arsenic concentrations, and offer a potential alternative supply of drinking water. However, Rosenboom (2004) found that 11% of shallow dug wells tested in the UNICEF/DPHE survey had concentrations above 50 μg L^{-1}. Dug wells are also more prone to bacterial contamination and sustainability problems during the dry season.

Arsenic in the Holocene aquifer sediments typically has concentrations in the range <1–20 mg/kg (BGS and DPHE 2001; Chakraborti et al. 2001; Swartz et al. 2004) which are not unusual by world standards. The arsenic has been found or postulated to occur in association with Fe(III)- or mixed Fe(II)–Fe(III) oxides (BGS and DPHE 2001; Harvey et al. 2002; Horneman et al. 2004; McArthur et al. 2001; Swartz et al. 2004), phyllosilicate minerals (Breit et al. 2001) and sulfide minerals (Nickson et al. 2000; Polizzotto et al. 2006). Most studies concur that iron oxides are an important control on arsenic mobilisation in the Bengal Basin groundwaters, although opinions are divided on the precise mechanisms involved. Competition for adsorption sites between dissolved arsenic species and other constituents such as phosphate is also likely to exert a control. Paucity of sulphate in the groundwaters over large parts of the region appears contributory to the limited sequestration of arsenic by iron sulphide under the strongly reducing conditions.

Redox changes in the aquifers involve reduction of As(V) to As(III) with resultant changes in sorption behaviour. Reductive dissolution of iron oxide and reductive desorption of arsenic are processes likely to result in increased solute arsenic concentrations, exacerbated by competition for binding sites between arsenic and especially phosphate. Such reactions appear to be only part of a complex sequence of diagenetic transformations in the sediments arising from the onset of reducing conditions. Reactions involve changes in iron oxide structure, redox status and surface properties over time, with transformation to phases such as magnetite as well as precipitation of siderite and vivianite having been invoked (e.g. BGS and DPHE 2001; Horneman et al. 2004; Swartz et al. 2004). Microbial mediation of the redox reactions and arsenic cycling has been demonstrated (Islam et al. 2004).

Several studies have found much higher ratios of Fe(II)/FeT in the Holocene Bengal sediments compared to the underlying Pleistocene deposits (Horneman et al. 2004; Polizzotto et al. 2006; van Geen et al. 2006). Iron oxides in the latter appear dominated by goethite and haematite and explain the prominent colour change from grey to orange-brown in passing from the Holocene to the Pleistocene sediments (e.g. van Geen et al. 2003b). Holocene sediments have also been found to contain much higher concentrations of phosphate-extractable (van Geen et al. 2006; Zheng et al. 2005) and oxalate-extractable (BGS and DPHE 2001) arsenic and iron than Pleistocene aquifer sediments, which point to an increased proportion of labile arsenic in Holocene deposits. Nonetheless, some studies have concluded that conditions in the Holocene aquifers are too strongly reducing to be consistent with in-situ Fe(III) reduction to Fe(II) and propose that arsenic is released instead from iron oxides in surficial soils and sediments and transported to depth as a plume by groundwater flow (Polizzotto et al. 2005, 2006).

The nature of organic matter involved in the redox cycling is also a matter of debate. Peat deposits (McArthur et al. 2001) and peaty strata marginal to peat basins (McArthur et al. 2004) have been invoked as important sources of organic matter to drive the redox reactions. Others cite the importance of disseminated organic matter in the sediments and dissolved in water (BGS and DPHE 2001). In these cases, rapid rates of sedimentation accommodate the organic carbon and can prevent its rapid oxidation. Others have concluded that introduction of anthropogenic carbon through drawdown induced by groundwater pumping has been responsible for catalysing the reduction reactions within the aquifers (Harvey et al. 2002, 2006), although this conclusion has not been universally accepted (e.g. Klump et al. 2006; van Geen et al. 2003a). In West Bengal, the lack of detectable thermotolerant coliforms in high-arsenic groundwaters has also been used as an argument against anthropogenic organic matter being the driver for iron reduction and arsenic release at depth (McArthur et al. 2004).

Groundwater flow and flushing has doubtless had an impact on arsenic mobilisation and distributions, both under natural conditions and by modern groundwater pumping (Michael and Voss 2008). Natural hydraulic gradients in the aquifers are low and groundwater flow strongly controlled by local permeability variations. Groundwater arsenic concentrations show a relationship with surface soil and sediment texture, which suggests a larger degree of aquifer recharge, and flushing, in areas with higher surface permeability (van Geen et al. 2006).

It is also likely that longer-term groundwater flow throughout the Quaternary has resulted in flushing of Pleistocene relative to Holocene deposits, especially since palaeo-head gradients would have been greater during pre-Holocene glacial lowstand periods (BGS and DPHE 2001). However, major abstractions of groundwater over the last few decades, particularly that for irrigation use, have changed flow patterns significantly. Many recent studies have shown that shallow groundwaters (<30 m or so) in the Bengal Basin often have short residence times, on the scale of years to decades (e.g. Klump et al. 2006), likely induced by pumping. Pumping-induced flow and mixing have the potential to either increase or decrease solute arsenic concentrations with time.

In recent years, many studies have focussed on the sustainability of the Pleistocene low-arsenic aquifer in respect of potential drawdown of contaminated water following increased abstraction. Modelling has concluded that the groundwater arsenic concentrations in the deep aquifer are likely to remain low provided abstraction is limited to domestic use by hand-pumps, partly because of the comparatively small volumes involved and partly due to the instigation of a hydraulic barrier in the shallow aquifer by irrigation pumping from there (Burgess et al. 2010; Michael and Voss 2008).

12.3.3.2 Nepal

Groundwater is abundant in the Quaternary alluvial sediments of the lowland Terai region of southern Nepal and is an important resource for domestic and agricultural use. Groundwater from tubewells supplies some 11 million people (Chitrakar and Neku 2001). Both shallow and deep aquifers occur throughout most of the Terai region, the deep aquifer being artesian. Quaternary alluvium also infills several intermontaine basins in Nepal, most notably that of the Kathmandu Valley of central Nepal.

A number of surveys of groundwater quality in the Terai region have revealed the presence of arsenic in samples from shallow tubewells (<50 m depth) (Gurung et al. 2005) and some 500,000 people are believed to be exposed to drinking water with arsenic concentrations greater than 50 μg L^{-1} (Pokhrel et al. 2009). Arsenicosis problems have been identified in some areas (Maharjan et al. 2007). By 2004, some 245,000 tubewells had been tested for arsenic, of which 3% were found to exceed 50 μg L^{-1}, the highest observed concentration being 400 μg L^{-1} (Tandukar et al. 2005). From testing in 17 of the 20 Terai districts, the Nepal Red Cross Society (NRCS), also found 3% of groundwater sources sampled having concentrations above 50 μg L^{-1}, the highest observed concentration being 205 μg L^{-1}. The worst-affected districts are Rautahat, Nawalparasi, Parsa and Bara in the central area of the Terai. Other surveys have revealed the occurrence of high-arsenic groundwaters in the River Bagmati area. Concentrations in the Kathmandu Valley appear mostly low (Gurung et al. 2007; Pant 2011).

The high arsenic concentrations in the Terai occur in anaerobic groundwaters with characteristics similar to those found in the Holocene aquifers of the Bengal Basin (high Fe, Mn, NH_4, DOC and low SO_4, NO_3). Dissolved arsenic is dominated by As(III) (Bhattacharya et al. 2003; Tandukar et al. 2005). Deeper tubewells in the Terai appear to have lower arsenic concentrations (generally <10 μg L^{-1}).

12.3.3.3 Taiwan

Health problems due to arsenic in groundwater were first reported in Taiwan between 1910 and 1920 (Chen and Liu 2007) although awareness of the link with drinking water was delayed until the 1960s (Tseng et al. 1968). Blackfoot disease (BFD, a peripheral vascular disorder with similarities to gangrene) is a well-publicised health problem of the region and is most likely linked to past occurrence of arsenic in the drinking water, although the high humic-acid concentration in groundwaters of the region has been cited as an additional causal factor (Lu et al. 1990). A number of other diseases, including internal cancers, have also been described. High groundwater arsenic concentrations have been recognised in both the south-west Chianan Plain (Chen and Liu 2007; Tseng et al. 1968) and north-east Lanyang Plain (Hsu et al. 1997; Lewis et al. 2007), although BFD occurrence appears restricted to the south-west (Lewis et al. 2007).

Groundwater samples from Chianan Plain have arsenic concentrations in the range 10–1,800 μg L^{-1} with a significant number being more than 400 μg L^{-1} (Chen and Liu 2007). These occur in groundwaters from deep artesian wells (mostly 80–250 m) abstracted from late Pleistocene alluvial and deltaic sediments. The groundwaters are strongly reducing with evidence of dissolved sulphide and CH_4 in some samples. Groundwaters from the top 50–80 m are also reducing but are saline with SO_4 concentrations of 2,700–6,000 mg L^{-1}, associated with marine deposition and subsequent evaporation (Chen and Liu 2007). Groundwaters abstracted from Lanyang are also artesian, but of shallower depth (typical range 16–40 m) (Hsu et al. 1997). Here, arsenic concentrations reach up to 600 μg L^{-1} (Hsu et al. 1997). In each area, the groundwaters are strongly reducing, dissolved As being dominated by As(III) (Chen et al. 1994). Groundwater samples taken from shallow open dug wells have low arsenic concentrations (Guo et al. 1994).

12.3.3.4 Cambodia, Laos and Vietnam

The discovery of high concentrations of arsenic in groundwaters from Cambodia, Laos and Vietnam is, perhaps surprisingly, relatively recent. Development of groundwater from tubewells in these countries has increased significantly

since the 1990s (Berg et al. 2006; Chanpiwat et al. 2011; Polya et al. 2005) and this recent change is perhaps responsible for the delayed recognition. High arsenic concentrations are found in groundwater from Holocene sediments. These are deposited by the Mekong River in all three countries, the Red River in Vietnam and the Bassac River (Buschmann et al. 2007) and the Tonle Sap (Berg et al. 2007) in Cambodia.

In Cambodia, 10 affected provinces have been identified but the most severely impacted is Kandal Province where extremes up to 3,500 μg L^{-1} have been reported (Berg et al. 2007; Buschmann et al. 2007; Sampson et al. 2008). Berg et al. (2006) found 20% of samples analysed exceeding 50 μg L^{-1}. The groundwaters occur mainly in groundwater from wells >16 m deep (Polya et al. 2005). In Laos, more limited data are available but concentrations along the Mekong Valley have been reported in the range <0.5–278 μg L^{-1} (Chanpiwat et al. 2011).

In Vietnam, aquifers in both the Mekong delta and Red River delta are exploited for water supply and the total number of tubewells in use could be of the order of one million (Berg et al. 2007). Aquifers exploited are of both Holocene and Pleistocene age. In the Vietnamese Mekong delta, the shallow aquifer is affected by salinity and is not used for drinking water. Groundwater for irrigation and drinking is therefore abstracted from depths of 150–250 m from Neogene deposits. Holocene aquifers in the northern Vietnamese part of the Mekong delta, upstream of the zone affected by saline intrusion, have arsenic concentrations in the range <1–845 μg L^{-1} (Berg et al. 2007).

In the Red River delta, concentrations up to 3,000 μg L^{-1} occur (Berg et al. 2007; Nguyen et al. 2009) in an area which includes the city of Hanoi, particularly the southern part. Holocene sediments up to about 30 m thick form the shallow aquifer, below which, Pleistocene sands and gravels occur. These are separated in places by a clay layer several metres thick, although 'windows' exist which allow hydraulic continuity between the two. Holocene sand deposits are largely capped by surficial confining overbank clays (typically some 2–10 m thick) (Postma et al. 2007). Both Holocene and Pleistocene aquifers are used for water supply. It is estimated that around one million people are drinking water in the delta area with an As concentration greater than 50 μg L^{-1} (Winkel et al. 2011).

Arsenic-affected groundwaters in these countries are reducing (with high concentrations of Fe, Mn and DOC; dissolved As overwhelmingly dominated by As(III)) (Chanpiwat et al. 2011; Rowland et al. 2007). In Vietnam, high concentrations of NH_4 and CH_4 are also found in some areas (Berg et al. 2007; Postma et al. 2007). As elsewhere in Asia, arsenic mobilisation in the Holocene aquifers has been linked to reductive dissolution of iron oxides in the sediments. In the Red River delta, these have been taken to be relatively crystalline oxide phases such as haematite or goethite (Postma et al. 2007). As in Bangladesh however, there are proponents of the origin of the arsenic lying in surface soils and sediments with subsequent transport to the Holocene aquifer (Polizzotto et al. 2008). In an extensive survey of groundwater in the Red River delta, Winkel et al. (2011) have documented the drawdown of As from the Holocene to the Pleistocene aquifer as a result of long-term heavy abstraction of groundwater.

To date, limited numbers of arsenicosis patients have been recognized in Cambodia, probably due to relatively recent introduction of tubewell water as a source of drinking water (Berg et al. 2007). However, the first cases were diagnosed (Kandal Province) in 2006 (Sampson et al. 2008). Few cases have been reported in Vietnam and to the best of our knowledge no arsenic-health problems have been documented in Laos.

12.3.3.5 Northern China

Arsenic (concentrations in excess of the Chinese national standard of 50 μg L^{-1}) has been found in groundwaters from a number of areas in northern China, including Xinjiang, Shanxi, Jilin and Liaoning Provinces as well as Inner Mongolia (Fig. 12.2) (Bian et al. 2012; Deng et al. 2009; Ning et al. 2007; Smedley et al. 2003; Xie et al. 2009). The earliest problems were recognised in Xinjiang Province where concentrations of 40–750 μg L^{-1} were found in deep artesian groundwater (wells to 660 m deep) from the Quaternary Dzungaria (Junggar) Basin (Wang and Huang 1994). Under artesian conditions, As concentrations increased with depth; shallow (non-artesian) groundwaters had concentrations between <10 and 68 μg L^{-1}.

In the Datong and Jinzhong Basins of Shanxi Province, arsenic concentrations up 4,400 μg L^{-1} have been found (Guo et al. 2003; Xie et al. 2008, 2009). Datong Basin groundwaters have highest observed concentrations in the depth range 15–60 m (Xie et al. 2008), present under usually strongly reducing conditions (Xie et al. 2009). Groundwaters from the Hetao Plain of Inner Mongolia also have high concentrations, up to 1,350 μg L^{-1} (Guo et al. 2001), being highest in boreholes with a depth range of 15–30 m (open dug wells of 3–5 m depth had low concentrations).

In the Huhhot Basin, part of the Tumet Plain of Inner Mongolia, arsenic also occurs in strongly reducing groundwaters from mainly Holocene alluvial and lacustrine aquifers. Problems are worst in the lowest-lying parts of the basin (Smedley et al. 2003). These groundwaters also have characteristically high concentrations of dissolved Fe, Mn, DOC, HCO_3 and NH_4 and low concentrations of nitrate and sulphate (Luo et al. 1997; Smedley et al. 2003). The arsenic, at concentrations up to 1,500 μg L^{-1}, is again dominated by As(III). Unlike the Bengal Basin, deep pre-Holocene aquifers also have relatively high, though less extreme,

arsenic concentrations (up to 300 μg L^{-1}). Many of these waters have associated high concentrations of humic acid which may have an impact on arsenic mobilisation, as well as health effects. Unusually, groundwater from several hand-dug wells in the low-lying part of the Huhhot Basin also has relatively high arsenic concentrations, up to 550 μg L^{-1}. This may owe to the maintenance of relatively reducing conditions at shallow depths as a result of high organic content (solid and dissolved) (Smedley et al. 2003).

Documented health problems derive especially from Xinjiang Province and Inner Mongolia. The problems are manifested by skin lesions, including melanosis and hyperkeratosis, vascular disorders and lung, skin and bladder cancer.

12.3.3.6 Hungary and Romania

Concentrations of arsenic above 50 μg L^{-1} have been identified in groundwaters from alluvial sediments in the southern part of the Great Hungarian Plain (part of the Pannonian Basin) of Hungary and neighbouring parts of Romania (Fig. 12.2). Concentrations up to 150 μg L^{-1} (average 32 μg L^{-1}) have been found by Varsányi et al. (1991). The Plain, some 110,000 km^2 in area, consists of a thick sequence of subsiding Quaternary sediments. Groundwaters vary from Ca-Mg-HCO_3-type in the recharge areas of the basin margins to Na-HCO_3-type in the low-lying discharge regions. Groundwaters in deep parts of the basin (80–560 m depth) with high arsenic concentrations are reducing and many have reported high concentrations of dissolved organic matter (humic acid quoted as up to 20 mg L^{-1}) (Varsányi et al. 1991). The groundwaters have largest arsenic concentrations in the lowest parts of the basin, where the sediment is fine-grained. Gurzau and Gurzau (2001) reported concentrations up to 176 μg L^{-1} in the associated aquifers of neighbouring Romania.

12.3.4 Inland Basins in Arid and Semi-arid Areas

12.3.4.1 Mexico

In terms of arsenic occurrence, the Principal aquifer of Comarca Lagunera in the states of Coahuila and Durango is the best-documented region in Mexico. The population exposed to arsenic in drinking water with >50 μg L^{-1} has been estimated at around 400,000 (Del Razo et al. 1990). Health problems related to arsenic were first recognized there in the early 1960s and symptoms include melanosis, keratosis and skin cancer. Groundwater from a large number of wells has been found with arsenic concentrations well in excess of the national standard for drinking water (currently 25 μg L^{-1}). Concentrations have been found in the range 3–740 μg L^{-1}, around half being above 50 μg L^{-1} (Del Razo et al. 1990, 1994; IMTA 1992; Rosas et al. 1999). Highest concentrations are found in the San Pedro, Tlahualilo and Francisco I. Madero areas (Fig. 12.4). The region has been subject to heavy groundwater abstraction and water levels have fallen significantly over the last century. The water table lies more than 100 m below surface in the centre of the depression.

The groundwaters of Comarca Lagunera are predominantly oxidising with neutral to high pH. Del Razo et al. (1990) quoted pH values for groundwaters in the range 6.3–8.9. They also noted that most (>90%) of the groundwater samples investigated had arsenic present predominantly as As(V). The groundwater also has high concentrations of fluoride (up to 3.7 mg L^{-1}) (Cebrián et al. 1994). The high arsenic concentrations occur in the low-lying parts of the basin which act as zones of groundwater discharge. Being an internal drainage basin the discharge area contains groundwater with significant residence times; radiocarbon dating and supporting stable-isotopic measurements reveal groundwater model ages up to 30,000 years in the low-lying areas (Brouste et al. 1997) (Fig. 12.4). Convergent flow within the discharge zones likely facilitates accumulation of solutes such as arsenic.

Arsenic problems have also been identified in groundwaters from the states of Sonora and Chihuahua in northern Mexico. Wyatt et al. (1998) found concentrations in Sonora in the range 2–305 μg L^{-1} and observed similar correlations with fluoride, which reached up to 7.4 mg L^{-1}. In Chihuahua, arsenic concentrations reached 650 μg L^{-1} in saline and F-rich groundwater from wells up to 250 m deep (Junta Municipal 2000).

12.3.4.2 Chile

Health problems related to arsenic in drinking water were first recognised in northern Chile in 1962. Symptoms included skin-pigmentation changes, keratosis, skin cancer, cardiovascular problems and respiratory disease. More recently, chronic arsenic ingestion has been linked to lung and bladder cancer. Arsenic exposure dated back from the period 1955–1970. Both surface waters and groundwaters from Administrative Region II (incorporating the cities of Antofagasta, Calama and Tocopilla) of northern Chile (Cáceres et al. 1992) have high concentrations. The region is arid, being part of the Atacama Desert, and water resources are limited. High As concentrations are accompanied by high salinity and concentrations of boron and lithium. This in part relates to evaporation but is also significantly affected by geothermal inputs from the El Tatio geothermal field. Arsenic concentrations below 100 μg L^{-1} in surface waters and groundwaters are apparently quite rare, and concentrations up to 21,000 μg L^{-1} have been found. Karcher et al. (1999) quoted ranges of 100–1,000 μg L^{-1} in untreated surface waters and groundwaters (average 440 μg L^{-1}).

The affected groundwaters of Chile are predominantly oxidising with As(V) the dominant species. The aquifers are composed of volcanic rocks and volcanogenic sediments, but the arsenic sources are not well-characterised. In Antofagasta, concentrations of arsenic in sediments are ca. 3.2 mg kg^{-1} (Cáceres et al. 1992). Sediments from the Rio Loa and its tributaries have much higher concentrations (26–2,000 mg kg^{-1}) as a result of geothermal inputs from the river system (Romero et al. 2003). Additional exposure to arsenic from the smelting of copper ore has also been noted in northern Chile (Cáceres et al. 1992).

12.3.4.3 Argentina

The Chaco-Pampean Plain of central Argentina constitutes perhaps one of the largest regions of high-arsenic groundwaters known, covering around one million km^2. High concentrations of arsenic have been found in parts of the provinces of Córdoba, La Pampa, Santa Fe, Buenos Aires, Tucumán, San Luis, Chaco and Santiago Del Estero. An estimated two million people live in these As-affected areas (Castro de Esparza 2009). Symptoms typical of chronic arsenic poisoning, including skin lesions and some internal cancers, have been recorded (Bates et al. 2004;

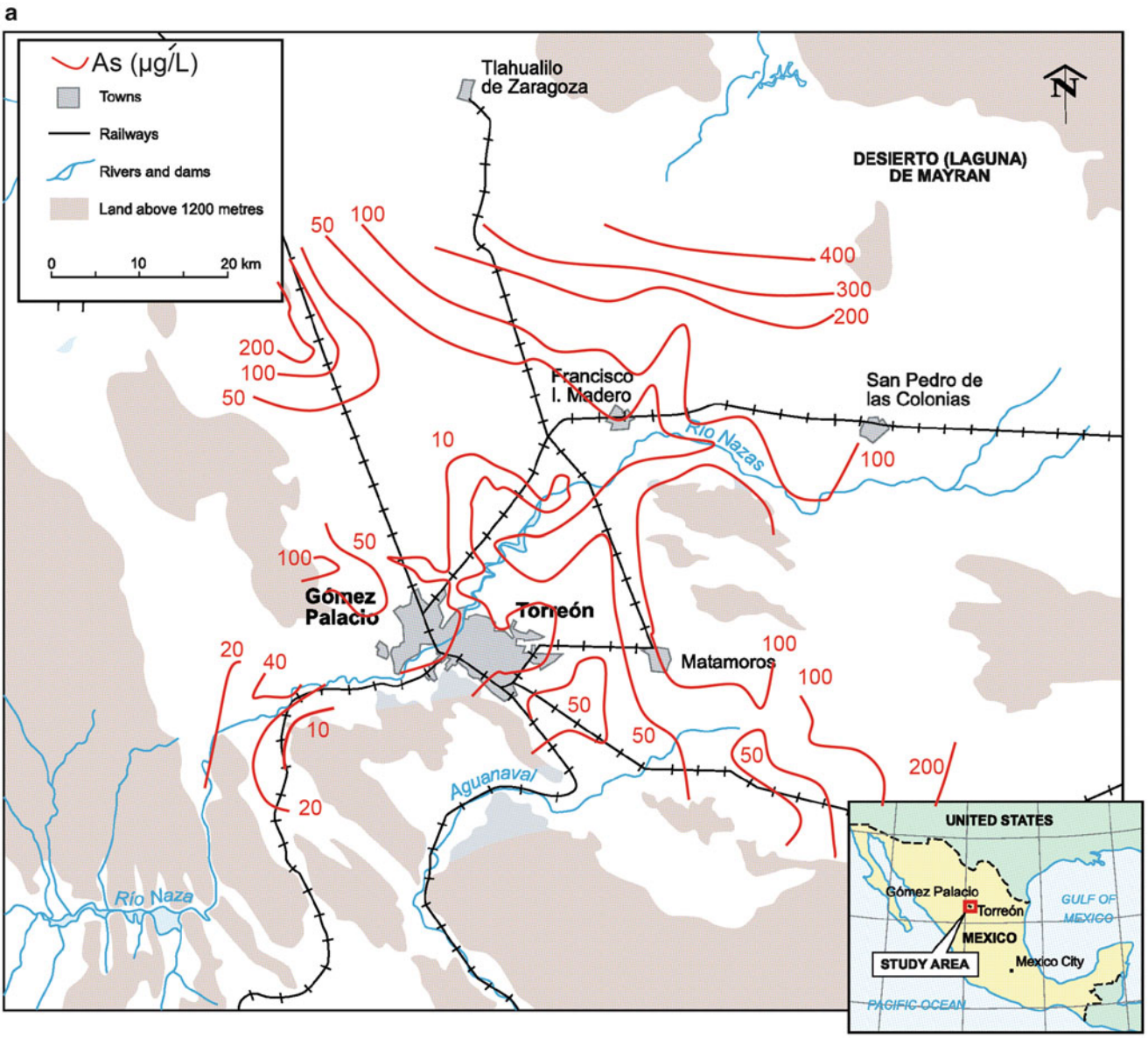


Fig. 12.4 (**a**)Distributions of arsenic in the groundwaters of Comarca Lagunera, Mexico (Data from IMTA 1992). (**b**) Groundwater bodies distinguished by residence time (Data from Brouste et al. 1997)

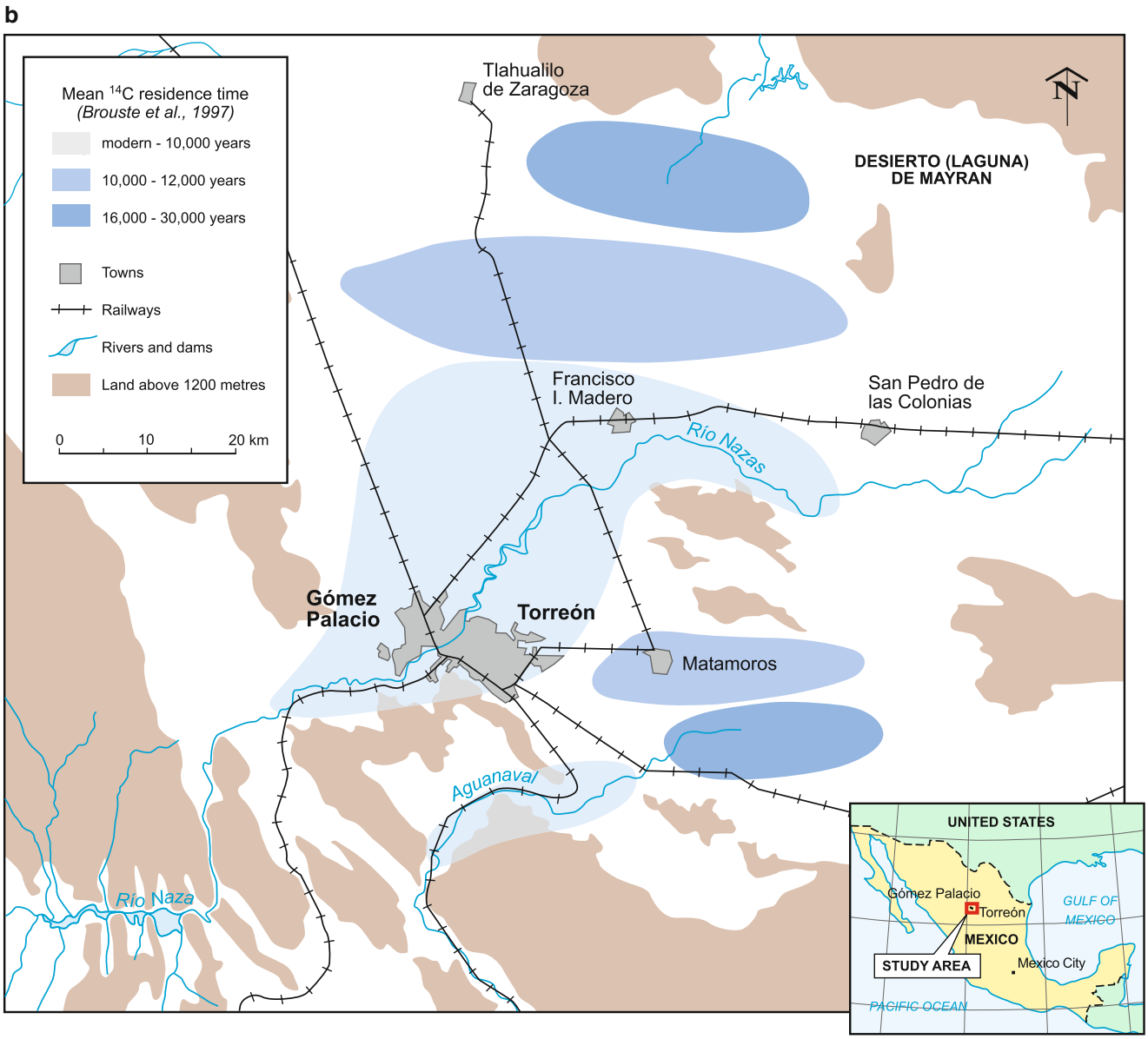


Fig. 12.4 (continued)

Hopenhayn-Rich et al. 1996). The high-arsenic groundwaters are mostly from Quaternary deposits of silt-grade loess, with intermixed rhyolitic or dacitic volcanic ash (Nicolli et al. 1989, 2010; Smedley et al. 2002). The sediments display abundant evidence of post-depositional diagenetic changes under semi-arid climatic conditions, with common occurrences of palaeosols and calcrete.

A large range of As concentrations is observed in the groundwaters, with many samples exceeding 1,000 μg L^{-1}. Nicolli et al. (1989) reported concentrations in groundwaters from Córdoba in the range 6–11,500 μg L^{-1} (median 255 μg L^{-1}). Smedley et al. (2002) found concentrations in La Pampa Province in the range < 4–5,280 μg L^{-1} (median 145 μg L^{-1}). Nicolli et al. (2010) found concentrations in Tucumán Province of 11–1,660 μg L^{-1} (median 46 μg L^{-1}). Blanes et al. (2011) reported concentrations up to 1,070 μg L^{-1} in groundwater from Chaco Province, Bhattacharya et al. (2006) found concentrations of 7–15,000 μg L^{-1} in Santiago Del Estero and Farías et al. (2003) found concentrations of 10–593 μg L^{-1} in groundwater from four Pampean provinces.

The groundwaters from the Chaco-Pampean loess plain have a large salinity range, with often high pH and alkalinity and are predominantly oxidising. Smedley et al. (2002) found pH values of 7.0–8.7 in groundwaters from La Pampa; Nicolli et al. (2010) found values of 6.3–9.2 in Tucumán. The high-As Pampean groundwaters are usually of Na-HCO_3 type. Under the overwhelmingly oxic conditions of the groundwaters, As(V) usually dominates

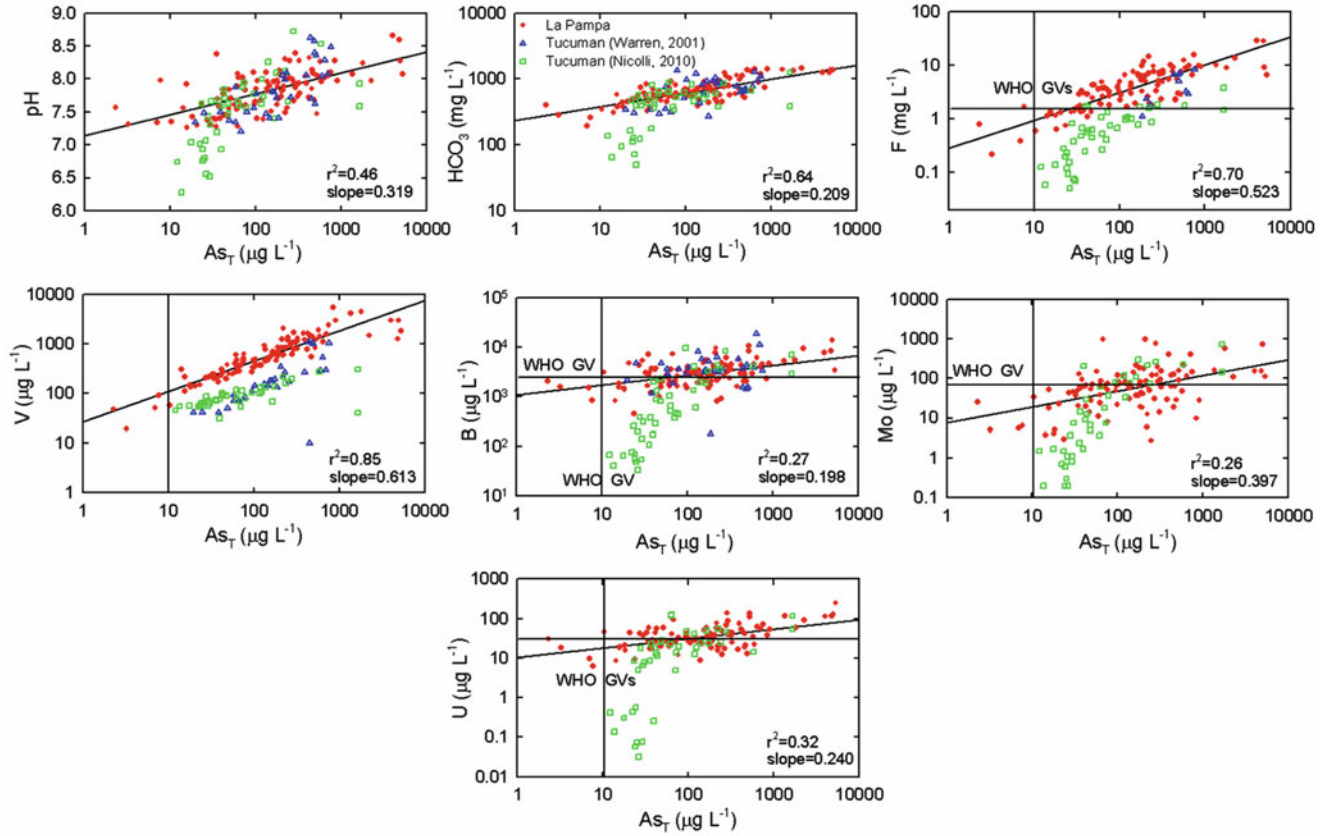


Fig. 12.5 Variations in total arsenic concentration with pH, alkalinity, fluoride, vanadium, boron, molybdenum and uranium in groundwater from La Pampa and Tucumán provinces, Argentina (Data from Nicolli et al. 2010; Smedley et al. 2002; Warren 2001)

(Bhattacharya et al. 2006; Smedley et al. 2002). Generation of the Na-HCO_3 high-pH waters is achieved through silicate hydrolysis (e.g. albite weathering), carbonate reaction and associated ion exchange.

Arsenic mobility in the Chaco-Pampean groundwaters is likely controlled to a large extent by pH-dependent sorption to metal oxides in the sediments. Positive correlations between dissolved As and pH in the groundwaters (Fig. 12.5) are consistent with weaker sorption of As(V) to iron oxide surfaces at high pH (Bhattacharya et al. 2005; Smedley et al. 2002). Presence of Mn(IV) oxides, along with low concentrations of solid organic carbon (Sracek et al. 2009) are likely contributory in maintaining oxic conditions through the aquifer.

A unifying feature of the Chaco-Pampean loess sediments is the presence of rhyolitic volcanic ash either intermixed with the sediments or in thin discrete layers. The components of the ash may constitute the ultimate sources of the dissolved As, though linkages are lost through subsequent weathering and redistribution. Dissolved As is also generally well-correlated with other anion and oxyanion elements (F, V, HCO_3, Mo and B; Fig. 12.5), some of these elements having very high concentrations. Volcanic ash may also be the ultimate source of these species (Nicolli et al. 1989).

12.3.4.4 Nicaragua

Alluvial aquifers of the Sébaco Valley of western Nicaragua contain groundwater with occasional high As concentrations. Altmirano Espinoza and Bundschuh (2009) reported concentrations in the range 10–122 µg L^{-1} and cited an earlier (2005) UNICEF survey in Jinotega municipality in which concentrations up to 1,200 µg L^{-1} were found. In the village of El Zapote, a well drilled in 1994 and tested for As in 1996 had a recorded concentration of 1,320 µg L^{-1}. Residents drinking the water displayed skin and respiratory disorders (Gomez et al. 2009) and the well was closed in 1996. The groundwater in the El Zapote well derived from Tertiary volcanic rocks which also form the precursors of the Quaternary alluvial aquifer in the valley. The groundwaters of the Sébaco Valley have pHs in the range 6–8, are oxic, and arsenic is dominantly As(V) (Altamirano Espinoza and Bundschuh 2009). The higher-As groundwaters appear to be of Na-HCO_3 composition.

Morales et al. (2009) also found high As concentrations in groundwater from neighbouring Esteli Department, with observed concentrations up to 115 μg L^{-1}, an apparent positive correlation between As and pH and similar dominance of As(V).

12.3.4.5 Spain

Cenozoic aquifers of the Duero Basin of north-central Spain have recently been found to contain arsenic with concentrations up to 613 μg L^{-1}. Sediments are mixed, carbonate-rich alluvial and lacustrine deposits and water sources within them include springs, shallow wells and boreholes up to 500 m deep (Gomez et al. 2006). Groundwaters are mainly oxic but with variable redox potentials and often high salinity and SO_4 concentrations. Groundwater pH ranges from 5.9 to 10.6 and the highest arsenic concentrations generally occur in waters with pH >8. Correlations have been found between arsenic and both vanadium and tungsten and, in shallow wells, also with uranium which reaches concentrations up to 216 μg L^{-1}. Concentrations of boron are also high in some (Gomez et al. 2006).

12.3.4.6 South-West USA

Concentrations of arsenic in most groundwaters of the USA are low. From some 20,000 analyses of groundwater samples taken from public-supply wells across the USA, Focazio et al. (2000) found that 11% exceeded 10 μg L^{-1} and 2% exceeded 50 μg L^{-1} (55% contained 1 μg L^{-1} or less). However, high concentrations have been found in several states, as shown for example by the USGS map of groundwater arsenic occurrence (Ryker 2001). Some of the most widespread and best-documented cases occur in the south-western states of Nevada, California and Arizona. Other aquifers in Maine, Michigan, Minnesota, South Dakota, Oregon, Texas, Oklahoma and Wisconsin have groundwater with concentrations of arsenic exceeding 10 μg L^{-1} (e.g. Ayotte et al. 2006; Focazio et al. 2000; Scanlon et al. 2009; Schreiber et al. 2000; Welch et al. 2000). From the high-arsenic areas, various sources and processes have been cited, including natural dissolution/desorption reactions, geothermal water and mining activity. High concentrations occur under both reducing and oxidising conditions in different areas. Evaporative concentration is thought to be an important process in the more arid areas. Human influences including effects of groundwater pumping have also been cited as causal in arsenic mobilisation (Ayotte et al. 2011).

In Nevada, arsenic concentrations over 100 μg/L have been found in a large number of private groundwater supplies (Fontaine 1994). Concentrations exceeding 100 μg L^{-1} with extremes up to 2,600 μg L^{-1} have also been found in shallow groundwaters from the southern Carson Desert (Welch et al. 1988). These are largely present under reducing conditions. The groundwaters also have associated high pH (>8) and high concentrations of phosphorus (some >4 mg L^{-1}) and uranium (>100 μg L^{-1}), concluded to be largely concentrated by evaporation, combined with the influence of redox and desorption processes involving metal oxides.

In California, a large range of arsenic concentrations (<1–2,600 μg L^{-1}) has been found in groundwater from the Tulare Basin of the San Joaquin Valley (Fujii and Swain 1995). Redox conditions are variable and high arsenic concentrations are found in both oxidising and reducing conditions. The proportion of groundwater arsenic present as As(III) increases with increasing well depth. The groundwaters from the Basin are often strongly affected by evaporation with resulting high concentrations of total dissolved solids. Many also have high concentrations of Se (up to 1,000 μg L^{-1}), U (up to 5,400 μg L^{-1}), B (up to 73,000 μg L^{-1}) and Mo (up to 15,000 μg L^{-1}) (Fujii and Swain 1995).

In Arizona, Robertson (1989) also noted the occurrence of high arsenic concentrations in some groundwaters under oxidising conditions (dissolved oxygen 3–7 mg L^{-1}) in alluvial aquifers of the Basin and Range Province. Only limited analysis of arsenic species was carried out, but results obtained suggested that the arsenic was present predominantly as As(V). The dissolved arsenic correlated positively with pH, Mo and V as well as Se and F. Groundwater pH values were in the range 6.9–9.3. Arsenic concentrations in the sediments of the Basin and Range Province were in the range 2–88 mg kg^{-1}. Oxidising conditions persist in the aquifers down to significant depths (600 m) despite the groundwaters having very long residence times (up to 10,000 years). The high arsenic (and other oxyanion) concentrations are a feature of the closed basins of the province.

12.3.5 Sulphide Mineralisation and Mining-Related Arsenic Problems

Arsenic problems have long been recognised in association with zones of ore mineralisation as a result of the high concentrations of arsenic in sulphide minerals. Oxidation of the sulphide minerals, exacerbated by mining activity, potentially leads to the release of arsenic and transition and heavy metals. Oxidation reactions are catalysed by bacteria, notably *Thiobacillus ferrooxidans*, *Thiobacillus thiooxidans* and *Leptospirillum ferrooxidans* (Schrenk et al. 1998). Oxidation releases SO_4 and associated trace elements and generates acidity which maintains high aqueous concentrations of many released trace metals. One of the highest arsenic concentrations ever reported, 850,000 μg L^{-1}, is from a mine seep at Iron Mountain, California (Nordstrom and Alpers 1999). Plumlee et al. (1999) also found arsenic concentrations in USA mine drainage ranging up to 340,000 μg L^{-1}.

Under oxidising conditions, iron released during pyrite oxidation typically precipitates as an iron(III) oxide. Trace elements likewise coprecipitate with, or adsorb onto, these iron oxides. Water is also neutralised in the presence of carbonate minerals. As a result, many mine waters have near-neutral pH values (e.g. Welch et al. 2000) and many have relatively low iron and arsenic concentrations. The considerable capacity of iron oxides to adsorb arsenic means that even in areas where high dissolved arsenic concentrations occur, they are usually restricted to the local area around the zone of oxidation, typically a few kilometres or less. Nonetheless, contamination of soils, sediments and vegetation in mining and mineralised areas can be substantial.

Most mining-impacted water sources are not used for potable supply, although a few notable exceptions exist and have given rise to localised health problems. Arsenic health problems related to mining activity have been well-documented in Ron Phibun District in Nakhon Si Thammarat Province of southern Thailand. The area lies within the South-East Asian Tin Belt and has been mined for several generations. Arsenic-related health problems were first recognised there in 1987. By the late 1990s, around 1,000 people had been diagnosed with skin disorders, particularly in and close to Ron Phibun town (Williams 1997). Arsenic concentrations up to 5,000 μg L^{-1} have been found in shallow groundwaters from Quaternary alluvial sediments that have been extensively dredged during mining operations. Deeper groundwaters from an older limestone aquifer were apparently less contaminated (Williams et al. 1996), although a few high arsenic concentrations occurred, presumably also as a result of contamination from the mine workings. The mobilisation of arsenic is related to the oxidation of arsenopyrite, exacerbated by the former tin-mining activities. The recent appearance in groundwater has occurred during post-mining groundwater rebound (Williams 1997).

A relatively recent discovery has also been made of arsenic-related health problems in northern Burkina Faso. In rural boreholes close to the town of Ouahigouya, concentrations of As in the range <0.5–1,630 μg L^{-1} have been documented following recognition of occurrences of melanosis, keratosis and skin cancer in some village populations. The highest concentrations occurred in boreholes in close proximity to mineralised veins containing arsenopyrite and other sulphide minerals which are hosted within Precambrian (Birimian) metamorphic rocks (Smedley et al. 2007) (Fig. 12.6). The arsenic was overwhelmingly present as As(V), signifying oxidation of the original As-hosting sulphide minerals. Although mining for gold takes place in Burkina Faso, this is chiefly on a small (artisanal) scale and there is no evidence that the mining operations per se have been responsible for the arsenic mobilisation. During the 1970s, health problems of suspected drinking water origin were also identified in a village close to the town of Mogtedo in central Burkina Faso. High concentrations of As (200–1,600 μg L^{-1}) were found in three borehole samples. The residents were relocated and further investigations of As in Birkinabé groundwater ceased until the renewed interest during the mid 2000s.

Environmental arsenic damage has also been documented in geologically analogous areas of neighbouring Ghana, where gold mining occurs on a much larger scale in the Ashanti Region. To date, there is little evidence of detrimental effects on human health through drinking-water exposure. Gold reserves have long been mined in the region where the gold is also associated with sulphide mineralisation, particularly arsenopyrite. Around the town of Obuasi, high arsenic concentrations have been found in soils close to the mines and treatment works. Some high concentrations were also found in river waters close to the mining activity. However, many of the groundwaters had low arsenic concentrations and the highest observed were not proximal to mining activity but appeared to be natural occurrences under mildy reducing aquifer conditions.

Arsenic contamination from mining activities has been identified in numerous areas of the USA (Welch et al. 2000). Groundwater from some areas has very high arsenic concentrations locally (up to 48,000 μg L^{-1}). Well-documented cases include the Fairbanks gold-mining district of Alaska, the Coeur d'Alene Pb-Zn-Ag mining area of Idaho, Bunker Hill mine in Idaho, Leviathan Mine in California, Mother Lode in California, Kelly Creek Valley in Nevada, Clark Fork river area in Montana and Lake Oahe in South Dakota.

Elsewhere, documented cases of mining-related arsenic contamination include the Zimapán Valley (Mexico), Baja California (Mexico), Lavrion (Greece), Minas Gerais (Brazil), Styria (Austria), Zloty Stok (Poland), Oruro Province (Bolivia), west Devon (England), eastern Zimbabwe, Korea, Sarawak (Malaysia) (Smedley and Kinniburgh 2002) and Aznalcollar (Spain) (Fig. 12.2). Doubtless, arsenic problems also exist in many other undocumented mining areas.

Increased concentrations of dissolved arsenic have also been found in parts of the world with local mineralisation which has not been mined. In Wisconsin, USA, arsenic and other trace-element problems in groundwater have arisen as a result of the oxidation of sulphide minerals (pyrite and marcasite) present as a discrete secondary cement horizon in the regional Ordovician sandstone aquifer. Concentrations of arsenic up to 12,000 μg L^{-1} have been found in the well waters (Schreiber et al. 2000). The oxidation appears to have been promoted by groundwater abstraction which has led to the lowering of the piezometric surface at a rate of around 0.6 m $year^{-1}$ since the 1950s, resulting in a partial

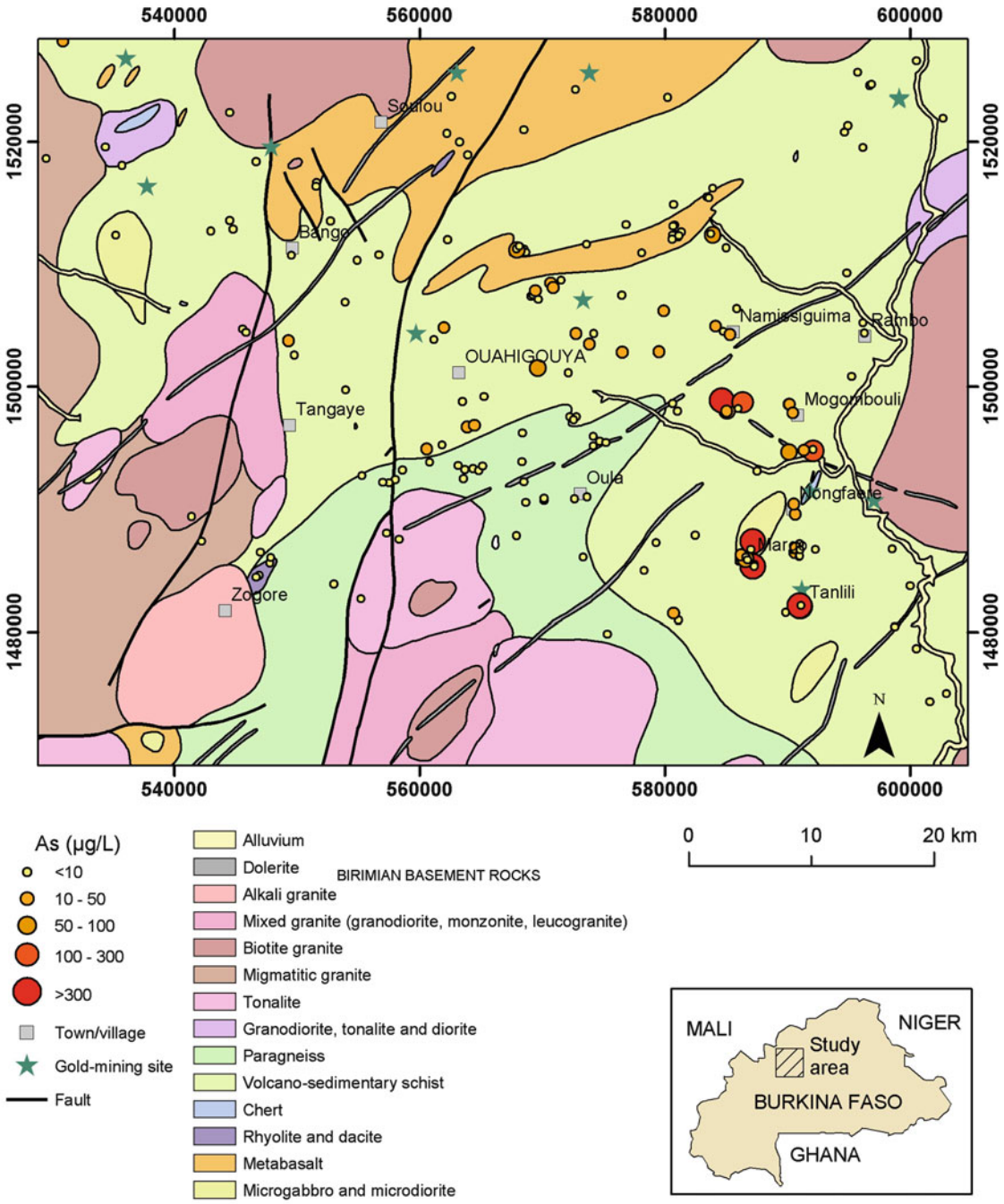


Fig. 12.6 Map of the distribution of arsenic in groundwater from the Ouahigouya area of Burkina Faso (From Smedley et al. 2007). High concentrations are associated closely with zones of sulphide mineralisation, which are also subject to artisanal gold mining

dewatering of the aquifer. The high arsenic concentrations were observed where the piezometric surface intersects, or lies close to, the sulphide cement horizon. Boyle et al. (1998) also recorded naturally high arsenic concentrations (up to 580 µg L^{-1}) in groundwaters from an area of sulphide mineralisation in Bowen Island, British Colombia.

12.3.6 Geothermal Sources

The common occurrence of high concentrations of arsenic in geothermal fluids has been recognised for a long time. Geothermal areas with documented high concentrations include the USA, Japan, New Zealand, Chile, Argentina, Equador,

Kamchatka, France, Italy and Dominica (Fig. 12.2). One of the largest and best-documented geothermal systems is that of Yellowstone National Park, USA where arsenic concentrations up to 7,800 μg L^{-1} have been found in hot springs and geysers (Thompson and Demonge 1996). Geothermal inputs have also given rise to high concentrations of arsenic (up to 370 μg L^{-1}) in waters of the Madison River, USA (Nimick et al. 1998). Others include Honey Lake Basin, California (arsenic up to 2,600 μg L^{-1}), Coso Hot Springs, California (up to 7,500 μg L^{-1}), Imperial Valley, California (up to 15,000 μg L^{-1}), Long Valley, California (up to 2,500 μg L^{-1}), Lassen Volcanic National Park, California (up to 27,000 μg L^{-1}), Steamboat Springs, Nevada (up to 2,700 μg L^{-1}) and Geyser Bight, Umnak Island, Alaska (up to 3,800 μg L^{-1}) (Welch et al. 2000). Geothermal inputs from Long Valley, California are believed responsible for the relatively high concentrations (20 μg L^{-1}) of arsenic in the Los Angeles Aqueduct which supplies water for the city of Los Angeles (Wilkie and Hering 1998). Geothermal inputs also contribute significantly to the high dissolved arsenic concentrations (up to 20 mg L^{-1}) in Mono Lake, California (Maest et al. 1992).

In New Zealand, geothermal waters occur with arsenic concentrations up to 9,000 μg L^{-1} (Webster 2003). River and lake waters receiving inputs from the Wairakei, Ohaaki, Orakei Korako and Atiamuri geothermal fields have concentrations up to 121 μg L^{-1} (Robinson et al. 1995). Concentrations diminish significantly downstream from the geothermal inputs.

Arsenic concentrations in the range 45,000–50,000 μg L^{-1} have been found in geothermal waters from the El Tatio system in the Antofagasta region of Chile (Ellis and Mahon 1977). The geothermal area lies between the volcanoes of the Andes and the Serrania de Tucle. Romero et al. (2003) found concentrations in waters from the Rio Loa and its tributaries in the range 120–10,000 μg L^{-1} as a result of inputs from the El Tatio geothermal system.

In Papallacta Lake, Equador, arsenic of geothermal origin has a range between 1,090 and 7,850 μg L^{-1}. The lake is fed by the Tambo River which in turn receives inflows from hot springs along its route. Speciation of As in water from two hot springs which were Fe-rich and highly reducing showed a dominance of inorganic As(III) (Cumbal et al. 2009). In contrast, two cool springs with higher redox potentials showed a dominance of As(V). The variations indicate oxidation of discharging geothermal fluids over time and mixing with oxygenated surface water.

Arsenic concentrations in the range 100–5,900 μg L^{-1} have been also been found in geothermal waters from Kamchatka (White et al. 1963) and in the range 500–4,600 μg L^{-1} in geothermal fields from Kyushu, Japan (Yokoyama et al. 1993).

White et al. (1963) also reported arsenic concentrations in geothermal waters from Iceland. While concentrations were high compared to most groundwaters (range 50–120 μg L^{-1}), these are much lower than the values found in many other geothermal systems described above. This may be related to the fact that the geothermal fluids in the Icelandic volcanic field are associated with oceanic basaltic magmas, hence having low arsenic concentrations (Webster and Nordstrom 2003). Typical high-arsenic geothermal fluids are associated with acidic volcanic systems in continental settings. In this case, higher arsenic concentrations may be derived from the interaction of geothermal fluids with the continental crust, particularly argillaceous sediments (Nordstrom 2002) in which the element is known to be preferentially partitioned. High arsenic concentrations have not been documented in geothermal systems associated with other volcanic provinces of dominantly basaltic composition, whether oceanic or continental (e.g. Hawaii, East African Rift, Reimann et al. 2003).

Hot springs commonly show a positive correlation between arsenic and chloride (e.g. Welch et al. 1988; Wilkie and Hering 1998). This association was noted by Webster and Nordstrom (2003) to relate to their similar behaviour during sub-surface boiling and phase separation as both elements partition preferentially into the liquid phase.

12.4 Mineral-Water Interactions

12.4.1 Adsorbed Arsenic in Sediments

Of the arsenic problem aquifers identified, it appears to be those hosted in young sediments that are most vulnerable to the development of high-arsenic groundwater on a regional scale. In these, surface reactions are important controls on the trace-element chemistry of the groundwaters. The major minerals which adsorb arsenic (as both arsenate and arsenite) in sediments are the metal oxides, particularly those of iron, aluminium and manganese. Iron oxides are relatively abundant in most sediments and are commonly produced by the weathering of primary minerals. In freshwater sediments and silicate clays, they often constitute about 50% of the iron present (Manning and Goldberg 1997). In sandy aquifers, the iron oxides are probably the most important adsorbents of arsenic because of their relative abundance, strong binding affinity and high specific surface area, especially the freshly-formed amorphous oxides. These are also particularly sensitive to a changing geochemical environment, acid dissolution and reductive dissolution, as well as to changes in mineral structure and crystallinity. Although the affinity for arsenate adsorption is strong, arsenite has been shown to adsorb more strongly to HFO and goethite than arsenate under alkaline conditions (pH > 7) (Dixit and Hering 2003). Aluminium oxides can be expected to play a significant role in arsenic adsorption when present in quantity (Manning and Goldberg 1997). Iron sulphide minerals (FeS, FeS_2) may also adsorb significant quantities of arsenic and it is likely the limited

supply of SO_4 in reducing groundwater that restricts the removal of arsenic from solution in many of the high-arsenic aquifers of south-east Asia (O'Day et al. 2004).

The role of carbonates as arsenic sorbents has also been investigated recently. Experiments by So et al. (2008) and Alexandratos et al. (2007) found that arsenate sorbs readily to calcite surfaces, the affinity depending on alkalinity, pH and ionic strength. Experiments by Goldberg and Glaubig (1988) also observed adsorption of As(V) on calcite, and Romero et al. (2004) speculated that calcite retarded arsenate in the carbonate aquifer of Zimapán, Mexico. By contrast, So et al. (2008) found that arsenite did not sorb significantly to calcite. These observations suggest that calcite would have little effect on arsenic mobility in reducing conditions but could be important in carbonate aquifers under oxic conditions.

12.4.2 Reduced Sediments and the Role of Iron Oxides

A recognised sequence of reduction reactions occurs when sediments are buried and the environment becomes anaerobic (Berner 1981). Such reaction sequences are common in sediments from a wide variety of environments (e.g. Postma et al. 2007). The processes causing changes in iron redox chemistry are particularly important since they can directly affect the mobility of arsenic. The reductive dissolution of hydrous iron oxides and/or the release of adsorbed and coprecipitated arsenic are key to this process. The redox sequence begins with the consumption of oxygen and an increase in dissolved CO_2 from the decomposition of organic matter. NO_3^- decreases by reduction to NO_2^- and ultimately to the gases N_2O and N_2. Insoluble manganese (IV) oxides dissolve by reduction to soluble Mn^{2+} and hydrous ferric oxides are reduced to Fe^{2+}. Subsequent reactions involve SO_4^{2-} reduction to S^{2-}, then CH_4 production from fermentation and methanogenesis, and finally reduction of N_2 to NH_4^+. The position of As(V) reduction in the sequence relative to Fe reduction has been disputed but is expected before SO_4^{2-} reduction. During sulphate reduction, the sulphide produced reacts with any available iron to produce FeS and ultimately pyrite, FeS_2. Iron is often more abundant than sulphur so that excess iron results beyond that which can be converted to pyrite. In the presence of carbonate or phosphate, dissolved Fe(II) may precipitate as siderite (e.g. Postma et al. 2007) or vivianite (Swartz et al. 2004), each of which has potential to remove As from solution (Islam et al. 2005b).

There is also evidence for solid-state transformations of the iron oxides under reducing conditions. This is reflected by a change from red/orange/brown to grey/green/blue coloration. Analysis of the Fe(II) and Fe(III) contents of iron oxides from reduced lake and aquifer sediments often indicate the presence of mixed Fe(II)-Fe(III) oxides (Davison 1993; Fendorf et al. 2010). Green rusts have been identified or suspected in anoxic soils and sediments (Cummings et al. 1999; O'Day et al. 2004; Taylor 1980). These consist of a range of green-coloured mixed Fe(II)-Fe(III) hydroxide minerals. Authigenic magnetite (Fe_3O_4) has also been identified in anaerobic sediments, formed by microbial transformations (Cummings et al. 2000; Fredrickson et al. 1998). Magnetite formation has also been established under reducing conditions in the laboratory (Benner et al. 2002; Dixit and Hering 2003; Guerin and Blakemore 1992; Islam et al. 2005b). Under strongly reducing conditions magnetite is unstable and eventually dissolves. In the presence of high concentrations of H_2S, magnetite converts slowly to pyrite on a scale of centuries or more.

These studies of iron oxides in reducing environments indicate the complexities involved in the sequence of events taking place when Fe(III) oxides are subjected to strongly reducing conditions. The changes are evidently substantial and can result in the partial dissolution of the oxides and their transformation to completely new mineral phases, as well as to changes in their arsenic sorptive capacities. Even quite small changes in arsenic binding could have a large impact on porewater arsenic concentrations because of the large solid/solution ratio found in sediments. Therefore, understanding the changes to the nature of iron oxide minerals in sedimentary environments is an important part of understanding the processes leading to arsenic mobilisation in sedimentary environments.

12.4.3 Role of Microbes

It has become increasingly clear that microbes play an important role in arsenic speciation and mobilisation. They can be significant catalysts in the oxidation of arsenite, respiration of arsenate and methylation and volatilisation reactions (Lloyd and Oremland 2006; Oremland and Stolz 2005; Oremland et al. 2004). The microbiological transformations either provide sources of energy or act as detoxifying mechanisms. A number of chemoautotrophs oxidise As(III) by using oxygen, nitrate or ferric iron as a terminal electron acceptor and CO_2 as their sole carbon source. Some heterotrophs are also capable of As(III) oxidation using organic carbon. Conversely, many prokaryotes (e.g. *Bacillus selentireducens* and *Bacillus arsenicoselantis*) are capable of As(V) respiration or As(III) oxidation (Oremland et al. 2002; Oremland and Stolz 2003; Pederick et al. 2007). Arsenic can also be released indirectly as a

result of other microbially-induced redox reactions. For example, the dissimilatory Fe-reducing bacterium *Shewanella alga* reduces Fe(III) to Fe(II) in scorodite ($FeAsO_4.2H_2O$), releasing As(V) in the process (Cummings et al. 1999). The Fe(III)-reducing bacteria *Geobacter* and *Geothrix* have also been shown to affect arsenic cycling in reduced aquifer sediments (Islam et al. 2005a, b).

Thermophilic bacteria and cyanobacteria have been recognised in geothermal settings and affect arsenic speciation and mineral precipitation in such systems (Webster and Nordstrom 2003). Streams affected by geothermal inputs may also be influenced by microbial activity. In geothermally-fed waters of Hot Creek, California, Wilkie and Hering (1998) concluded that the oxidation of As(III) was bacterially-controlled and that in abiotic systems, this oxidation reaction was much slower. Extremophiles have also adapted to the alkaline and saline conditions in the geothermally-influenced Mono Lake of California (Oremland et al. 2002, 2004).

Organic matter plays a key role in many microbial reactions in sediments and the type of organic matter present affects its reactivity and hence bioavailability. Organic matter has been shown to act both as an energy source for Fe- and As-reducing bacteria and as an electron shuttle enhancing microbial reactions (Islam et al. 2005a; Rowland et al. 2007).

12.5 Common Features of High-Arsenic Groundwater Provinces

12.5.1 The Sources of Arsenic

In the cases where affected groundwaters are found close to obvious geological or industrial sources rich in arsenic (geothermal springs, drainage from mineralised and mining areas, specific contaminant sources), it is clear that the anomalously high arsenic concentrations in the source region are responsible. The extent of this contamination is usually highly localised because the geochemical conditions within most aquifers do not favour arsenic mobilisation on a regional scale. Areas affected by geothermal activity are potentially more widespread since in this case mobilisation of arsenic is not required: arsenic is already present in solution and the size of geothermal reservoirs can be large. This probably accounts for why high-arsenic surface waters are normally located in geothermal areas. Perhaps more remarkable is the way in which very high concentrations of arsenic, up to several mg L^{-1}, are found in groundwaters from areas with unremarkable concentrations of arsenic in the source rocks. In aquifers with extensive areas of high-arsenic groundwater, this appears to be the rule rather than the exception. Most of these cases arise in aquifers derived from relatively young sediments, often consisting of alluvium or loess where the total sediment arsenic concentrations are typically in the range 1–20 mg kg^{-1} (e.g. BGS and DPHE 2001; Postma et al. 2007; Swartz et al. 2004).

Of critical importance is the fact that drinking-water limits for arsenic are very low in relation to the overall abundance of arsenic in the environment. Fortunately, most arsenic is normally retained in the solid phase and does not constitute a problem for potable water supplies. However, it only takes a very small percentage of this 'solid' arsenic to dissolve or desorb to give rise to a serious groundwater problem.

12.5.2 Mobilisation of Arsenic

Two key factors appear to be involved in the formation of high-As groundwaters on a regional scale: firstly, some form of geochemical trigger releases arsenic from the aquifer solid phase into the groundwater. Secondly, the released arsenic remains in the groundwater and must not be flushed away. There are a number of possible geochemical triggers. In mining and mineralised areas, oxidation of sulphide ores is triggered by influxes of oxygen or other oxidising agents. This may follow a lowering of the water table or change in hydrogeological regime or excavation by mining. In most arsenic-affected aquifers, the most important trigger appears to be the desorption or dissolution of arsenic from oxide minerals, particularly iron oxides. An important feature of this process is that the initial adjustment to environmental changes is probably quite rapid since adsorption reactions are surface reactions. The rate-limiting factors are probably those that control the major changes in pH, redox condition and associated water-quality parameters of the aquifer. These are in part related to physical factors such as the rate of diffusion of gases through the sediment, the rate of sedimentation, the extent of microbiological activity and the rates of chemical reactions. Many of these are likely to be rapid on a geological time scale. Dissolution reactions are relatively slow but even oxide dissolution is rapid on a geological time scale and can be observed over the course of weeks or even days in flooded soils (Masscheleyn et al. 1991). A qualification is that if diagenetic changes to the oxide mineral structure take place or if burial of sediment occurs, there could be a slow release of arsenic over a much longer time scale. The rates of release are likely to be most rapid initially and to diminish over time. Once the diagenetic readjustment has taken place and the sediments have equilibrated with their new environment, there should be little further release of arsenic.

The geochemical triggers involved in arsenic mobilisation could arise for a number of reasons. These are discussed further below.

Table 12.3 Summary of documented high-arsenic groundwater provinces and their typical chemical and hydrogeological characteristics

Groundwater environment/ aquifer type	Examples	Typical aquifer conditions	Typical chemical features of high-arsenic groundwaters	Likely mechanisms of arsenic mobilisation
Strongly reducing groundwater	Alluvial/deltaic aquifers of the Bengal Basin, Mekong Valley, northern China, Taiwan, Nepal, Great Hungarian Plain	**Young (Quaternary) sediments, often slow groundwater flow, low-lying parts of aquifers; rapidly accumulated sediments**	High Fe (>1 mg L^{-1}), Mn (>0.5 mg L^{-1}), NH_4-N (>1 mg L^{-1}), high HCO_3 (>500 mg L^{-1}); low NO_3-N (<1 mg L^{-1}), low SO_4 (<5 mg L^{-1}). Often high P (>0.5 mg L^{-1}). Sometimes high concentrations of dissolved organic matter, including humic acid. Arsenic dominated by As(III)	Reductive desorption of As from metal oxides and reductive dissolution of Fe oxides. Competition between arsenic and other anionic species (especially P)
Oxic groundwater, high pH	Inland basins or closed basins (arid and semi-arid areas): Argentina, Mexico, parts of Nicaragua, western USA, Chile, Spain	Young (Quaternary) sediments, slow groundwater flow, low-lying parts of aquifers; often (not always) volcanic components present	pH typically >8, high HCO_3 (>500 mg L^{-1}), low Fe, Mn, often correspondingly high F, U, B, V, Mo, Se. Some groundwaters have high salinity due to evaporation. Arsenic dominated by As(V)	Desorption of As and other oxyanion-forming elements from metal oxides, especially of Fe and Mn; volcanic ash may be ultimate source for some locations
Geothermally-influenced groundwater	**Parts of Kamchatka, Chile, Argentina, Equador, western USA, Japan, New Zealand**	Any aquifers affected by geothermal inputs, especially in rift zones	High Si, B, Li, often high salinity (Na, Cl); high pH (>7); increased groundwater temperature; As speciation variable	Mixing of fresh groundwater with geothermal solutions
Groundwater from sulphide mineralised/ mining areas	Parts of Burkina Faso, Canada, USA, Thailand, Korea, Poland, Greece, Ghana, Zimbabwe, England	Groundwater in fractures in crystalline rocks or alluvial placer deposits	Oxidising or mildly reducing conditions possible; high SO_4 concentrations (typically hundreds of mg L^{-1} or higher); acidic (unless buffered by carbonate minerals); often increased concentrations of other trace metals (Ni, Pb, Zn, Cu, Cd). Arsenic speciation variable	Oxidation of sulphide minerals

12.5.3 Desorption of Arsenic at High pH Under Oxidising Conditions

Under the aerobic and acidic to near-neutral conditions typical of many natural environments, arsenic is strongly adsorbed by oxide minerals as the arsenate ion and the concentrations in solution are low. The non-linear nature of the adsorption isotherm for arsenate (Dzombak and Morel 1990) ensures that the amount of arsenic adsorbed is often relatively large, even when dissolved concentrations of arsenic are low. With an increase in pH, especially above pH 8.5, arsenate tends to desorb from oxide surfaces, thereby increasing the concentration in solution (Dzombak and Morel 1990). The impact of this is magnified by the high solid/solution ratios typical of unconsolidated aquifers.

The presence of other ions in the system can further impact on dissolved arsenic concentrations because other ions such as phosphate compete for sorption sites on the oxide (e.g. HFO, goethite) and will reduce the arsenic loading on the solid surface. At high pH, phosphate is also released (Dzombak and Morel 1990). Other potential competing ions include bicarbonate, silicate and vanadium (Appelo et al. 2002; Smedley et al. 2005; Swedlund and Webster 1999). Bicarbonate is often the dominant anion in high-arsenic groundwaters and concentrations can be high, frequently exceeding 500 mg L^{-1}, and occasionally above 1,000 mg L^{-1} (Table 12.3). Concentrations of silica can also be high in some oxic arid groundwaters; high concentrations of vanadium are more unusual but have been found for example in Argentina (Smedley et al. 2002, 2005).

There are several reasons why the pH might increase but the most important in the present context is the uptake of protons by mineral weathering and ion-exchange reactions, together with the effect of evaporation in arid and semi-arid regions. A pH increase is commonly associated with the development of salinity and the salinisation of soils. Inputs of high-pH geothermal waters may be important in maintaining high arsenic concentrations in some alkaline

lakes. Desorption at high pH is the most likely mechanism for the development of groundwater-arsenic problems under oxidising conditions such as those observed in Argentina, Mexico and parts of the USA for example, and would account for the observed positive correlation of arsenic concentrations with increasing pH (e.g. Fig. 12.5). Increases in pH also induce the desorption of a wide variety of other oxyanions including those of V, U, Mo and B which are observed in many high-pH groundwaters (Fig. 12.5) (Bhattacharya et al. 2005; Smedley et al. 2002, 2005). These specifically-adsorbed anions all interact with adsorption sites on the oxides in a competitive way and so influence the extent of binding of each other.

By contrast, some cations, because of their positive charge, may promote the adsorption of negatively charged arsenate (Wilkie and Hering 1996). Calcium and magnesium are likely to be the most important cations in this respect because of their abundance in most natural waters and their +2 charge. Changes in surface charge and sorption capacity due to Ca/Na ion exchange have been implicated recently in the mobilisation of arsenate in the Southern High Plains aquifer, Texas (Scanlon et al. 2009). Other cations could also have an influence: Fe^{2+} may be important in reduced waters and Al^{3+} in acidic waters.

The aridity described above enables high pH values to be maintained. Aridity also minimises the flushing of released arsenic. It also allows the build-up of high chloride and fluoride concentrations. Other high-pH environments (up to pH 8.3), particularly open-system calcareous environments, are likely to be too well flushed to allow released arsenic to accumulate.

12.5.4 Arsenic Desorption and Dissolution Due to a Change to Reducing Conditions

The onset of strongly reducing conditions, sufficient to enable iron(III) and probably sulphate reduction to take place is another well-documented trigger for the release of arsenic. Arsenic may be released by reductive dissolution, or reductive desorption, or a combination of the two. No clear relationship between dissolved concentrations of Fe(II) or As are expected as a result (Pedersen et al. 2006). Such processes occur in large alluvial systems, especially broad lowland meander belts and braided channels, and in prograding deltas. Microbial reactions are key and the organic carbon content and reactivity in the buried sediment can determine the rate at which reducing conditions are created (Postma et al. 2007; Rowland et al. 2007). Freshly-produced soil organic matter is reactive and readily decomposes; the presence of even small quantities can consume all of the available electron acceptors (e.g. dissolved oxygen, nitrate and sulphate) in the system. Solid-phase Fe (III) in minerals may moderate the rate of reduction. Reducing conditions are maintained as long as the diffusion and convection of dissolved oxygen and other oxidants from the surface is less rapid than their consumption. This is facilitated if there is a confining layer of fine-grained material close to the surface. This commonly occurs in large deltas where fine-grained overbank deposits overlie coarser-grained alluvial deposits.

A change in the redox state of adsorbed arsenic from As (V) to As(III) also occurs during the diagenetic process. Arsenite is found as a dominant redox state in many reducing sediments (e.g. Rowland et al. 2005). Some studies have found As reduction coinciding with Fe(III) reduction (Herbel and Fendorf 2005). Others have concluded that reductive dissolution of iron oxides precedes the reduction of As(V) to As(III) (Islam et al. 2004).

The changing As redox state will also affect a large number of competing reactions. Phosphate-arsenite competition, for example, is likely to be less important than phosphate-arsenate competition. Competition with HCO_3 and organic matter is possible (Appelo et al. 2002; Stachowicz et al. 2007). There is also the potential for arsenite-arsenate competition.

12.5.5 Changes in the Structure of Oxide Minerals

Disordered and fine-grained iron oxides, including HFO, lepidocrocite, schwertmannite and magnetite, are common products of the early stages of weathering. Freshly-precipitated HFO is extremely fine-grained with a specific surface area of 300 $m^2\ g^{-1}$ or more. HFO gradually transforms to more ordered structures such as goethite or haematite with larger crystal sizes and reduced surface areas. Goethite typically has specific surface areas of 150 $m^2\ g^{-1}$ or less and those for haematite are even less (Cornell and Schwertmann 1996). This ageing reaction can take place rapidly in the laboratory but the rate in nature may be inhibited by the presence of other ions, particularly strongly adsorbed ions such as aluminium, phosphate, sulphate, arsenate, bicarbonate and silicate (Cornell and Schwertmann 1996). A reduction in surface area ultimately decreases the amount of As(V) adsorption possible on a weight-for-weight basis. If the site density (site nm^{-2}) and binding affinities of the adsorbed ions remain constant, then as the specific surface area of the oxide mineral is reduced, some of the adsorbed ions may be desorbed (Dixit and Hering 2003).

Under strongly reducing conditions, additional processes could operate which lead to a change in the binding affinity for arsenic. In iron oxides, some of the surface iron could be reduced from Fe(III) to Fe(II) to produce a mixed-valence oxide such as magnetite. This has been shown to have variable effects on arsenic release, especially depending on

microbiological interactions. Several studies have shown that biotransformation of fresh iron oxides can lead to a net reduction in the release of As, at least in the short term until specific surface area is reduced to a point where As binding sites are diminished or reductive dissolution of the oxide occurs (Herbel and Fendorf 2006; Kocar et al. 2006; Pedersen et al. 2006; Tufano and Fendorf 2008). Herbel and Fendorf (2006) observed that in experiments with HFO reduction, *Sulfospirillium barnesii*, a bacterium capable of both Fe(III) and As(V) reduction, high concentrations of Fe(II) were induced which were capable of transforming ferrihydrite to goethite and magnetite, and although As was released initially, release diminished upon mineral transformation. Nonetheless, laboratory and field evidence suggests that at micromolar concentrations of arsenic, freshly-formed HFO binds more arsenic than the more crystalline goethite on a mole of Fe basis (de Vitre et al. 1991) and so a reduction in affinity is more likely overall (Dixit and Hering 2003; Herbel and Fendorf 2005).

In most soils and sediments, total As concentrations tend to correlate with total Fe. In Bangladesh, areas with high-As groundwaters tend to correspond with areas in which the sediments contain relatively large amounts of phosphate- or oxalate-extractable iron (BGS and DPHE 2001; van Geen et al. 2006; Zheng et al. 2005). These observations provide indirect support for the importance of labile iron oxides in releasing arsenic to solution. It is likely that the soils and sediments most sensitive to arsenic release on reduction and ageing are those in which iron oxides are abundant, in which HFO is initially a major fraction of the iron oxides present, and in which other arsenic-sorbing minerals are absent or scarce.

12.5.6 Mineral Dissolution

Mineral dissolution reactions tend to be most rapid under extremes of pH and Eh. For example, iron oxides dissolve under strongly acidic conditions and under strongly reducing conditions. Minor elements, including arsenic, present either as adsorbed (labile) arsenic or as irreversibly-bound (non-labile) arsenic will also tend to be released during the dissolution. This can explain, at least in part, the presence of high arsenic concentrations in acid mine drainage and in strongly reducing groundwaters. Reductive dissolution of Fe(III) oxides accounts for the high Fe(II) content of anaerobic waters.

As described above, one of the most pertinent mineral dissolution processes in respect of arsenic mobilisation is the oxidation of sulphide minerals. Pyrite, the most abundant of these minerals, can be an important source of arsenic, especially where it is freshly exposed by mining or by lowering of the water table. In extreme cases, this can lead to highly acidic groundwaters rich in sulphate, iron and trace metals. As the dissolved iron is neutralised and oxidised, it precipitates as HFO with resultant adsorption and coprecipitation of dissolved As(V). In this sense, pyrite oxidation is not a very efficient mechanism for releasing arsenic into water.

12.5.7 Arsenic Transport Through Aquifers

The geochemical triggers described above are capable of releasing arsenic into groundwater but are not in themselves sufficient to account for the distribution of high-arsenic groundwaters observed in various parts of the world. The released arsenic must necessarily also have not been flushed away or diluted by groundwater flow. The rate of arsenic release must be set against the accumulated flushing of the aquifer during the period of release. The rocks of most aquifers used for drinking water are millions of years old and yet contain groundwater that may be at most a few thousand years old. Hence, a large number of pore volumes of fresh water will have passed through the aquifer over its history. This is also the case in most young aquifers with actively flowing groundwater. By contrast, many alluvial and deltaic aquifers are composed of relatively young sediments. Where groundwater flow is slow, these can contain relatively old groundwater. High groundwater arsenic concentrations can occur on a regional scale when geochemical conditions capable of mobilising arsenic are combined with hydrogeological conditions which prevent its removal.

The Quaternary period has seen considerable changes in climate and global sea level. Variation in groundwater piezometric levels over this period would have induced large variations in base levels of erosion and in groundwater flow regimes and rates. During the last glaciation, some 21,000–13,500 years ago, sea levels would have been up to 130 m below present mean sea level. This was a worldwide phenomenon and would have affected all then existing coastal aquifers. Continental and closed basin aquifers on the other hand would not have been affected. The increased hydraulic gradient in coastal aquifers during the glacial period would have resulted in correspondingly large groundwater flows and extensive flushing. The arsenic in groundwater in these older aquifers would therefore tend to have been flushed away. The deep unsaturated zone would also have led to more extensive oxidation of the shallower horizons with possible increased sorption of arsenic to Fe(III) oxides. Aquifers younger than around 7,000 years old, i.e. of Holocene age, will not have been subjected to this increased flushing. Such conditions are applicable to the southern Asian aquifers where Holocene high-As aquifers overlie older Pleistocene aquifers with low As concentrations. Differing flushing histories are likely contributory, in combination with different aquifer redox characteristics (grey vs. red-brown aquifer sands), to the differing aqueous As distributions observed.

The evidence suggests that in the Bengal Basin, Pleistocene aquifers have thus far been largely protected from contamination by As-rich groundwater from the overlying Holocene aquifers by clay aquitards that separate the two. However, that could change if drawdown of groundwater is induced by heavy pumping of the deep aquifer (Michael and Voss 2008). Such drawdown appears to have already affected the Pleistocene aquifer of the Red River delta of Vietnam, where intervening clay aquitards are thinner or in places absent (Winkel et al. 2011) and the aquifer is contaminated in places as a result of a century of heavy groundwater abstraction. Perhaps even more vulnerable are the currently low-As zones identified in Holocene deposits as lateral flow induced by pumping is more favourable than vertical flow in the stratified sediments (Fendorf et al. 2010).

At a local scale, small variations in relief or in drainage patterns may dictate local flow patterns and hence the distribution of arsenic-rich groundwater. Low-lying areas favour deposition of fine-grained deposits and are prone to flooding, under which reducing conditions can prevail and groundwater flow is restricted. Small-scale topographic depressions can also be the locations of seasonal discharge which restrict arsenic flushing. Such accumulations of dissolved As in topographic depressions have been observed for example in Argentina, China, Bangladesh, Hungary and Mexico (Nath et al. 2010; Smedley et al. 2002, 2003; Varsányi et al. 1991). It is clear that flat low-lying areas, particularly large deltas and inland basins, are prone to potentially high-arsenic groundwaters since they combine many of the risk factors identified above.

12.6 Concluding Remarks

This account has attempted to characterise the distribution of arsenic in the environment, and to describe the main geochemical controls on its speciation and mobilisation. We have also summarised the documented arsenic-rich areas and highlighted the main types of environment under which arsenic concentrations in waters are expected to be problematic. Doubtless there are other areas of the world where arsenic problems are yet to be recognised. However, increased awareness of the need for testing, increased health surveillance and increased recognition of the various risk factors associated with high-arsenic water mean that these are likely to be diminished in number. There are however, other areas where problems have been recognised but the scale is yet to be determined.

Environmental research has progressed significantly since the first aquifers were mapped or cases of arsenicosis recognised, although uncertainties remain in some of the key biogeochemical processes involved. On a simple level, mobilisation of arsenic relates to changes in speciation and mineral solubility as a result of pH and redox changes in aquifers and the environment. But the processes involve a complex interplay of microbial interactions, ageing and diagenesis of metal oxides, ion competitive effects, mineral solubilities, presence of multiple minerals, nature and reactivity of organic matter, and flow.

One of the greatest hindrances to development is the enormous degree of spatial heterogeneity in aqueous As concentrations recognised in many if not all affected aquifers, often seemingly in aquifers with unremarkable concentrations of solid-phase arsenic. Such heterogeneity, on various scales, requires major testing and monitoring operations and hinders prediction of water quality for future drilling programmes. In the efforts to provide alternative sources of low-arsenic water to affected populations, significant challenges also exist to avoid risk substitution. Such risks include less-sustainable supplies, bacterial and other contamination of surface sources or future arsenic contamination of thus-far low-arsenic aquifers.

Groundwater provides an important source of drinking water to many millions of people globally. While this chapter has focused on areas where groundwater arsenic concentrations are often high, it is to be remembered that these are the exceptions. Groundwater more commonly provides a safe and reliable form of drinking water. In reducing exposure to the potentially fatal effects of faecal contamination, the use of groundwater for improving community health in developing countries has been highly successful. With regards to arsenic, most wells in most aquifers are likely to be uncontaminated, even when the groundwaters contain high concentrations of dissolved iron. It is also important to understand why these groundwaters are not affected. It appears that it is only when a number of critical geochemical and hydrogeological factors are combined that high-arsenic groundwaters occur.

See Also the Following Chapters. Chapter 24 (Environmental Medicine), Chapter 25 (Environmental Pathology), Chapter 33 (Modeling Groundwater Flow and Quality)

References

Ahmed KM, Hoque M, Hasan MK, Ravenscroft P, Chowdhury LR (1998) Occurrence and origin of water well methane gas in Bangladesh. J Geol Soc India 51:697–708

Alexandratos VG, Elzinga EJ, Reeder RJ (2007) Arsenate uptake by calcite: macroscopic and spectroscopic characterization of adsorption and incorporation mechanisms. Geochim Cosmochim Acta 71:4172–4187

Altamirano Espinoza M, Bundschuh J (2009) Natural arsenic groundwater contamination of the sedimentary aquifers of southwestern Sebaco valley, Nicaragua. In: Bundschuh J, Armienta MA, Birkle P, Bhattacharya P, Matschullat J, Mukherjee AB (eds) Natural arsenic in groundwaters of Latin America. CRC Press/Balkema, Leiden, pp 109–122

Amirbahman A, Kent DB, Curtis GP, Davis JA (2006) Kinetics of sorption and abiotic oxidation of arsenic(III) by aquifer materials. Geochim Cosmochim Acta 70:533–547

Appelo CAJ, Van der Weiden MJJ, Tournassat C, Charlet L (2002) Surface complexation of ferrous iron and carbonate on ferrihydrite and the mobilization of arsenic. Environ Sci Technol 36:3096–3103

Ayotte JD, Nolan BT, Nucklos JR, Cantor KP, Robinson GR, Baris D, Hayes L, Kargas M, Bress W, Silverman DT, Lubin JH (2006) Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. Environ Sci Technol 40:3578–3585

Ayotte JD, Szabo Z, Focazio MJ, Eberts SM (2011) Effects of human-induced alteration of groundwater flow on concentrations of naturally-occurring trace elements at water-supply wells. Appl Geochem 26:747–762

BAMWSP (2005) Bangladesh Arsenic Mitigation Water Supply Project (BAMWSP). Upazila wise summary results. http://www.bwspp.org/BAMWSPContents/Survey%20Result/Upazila%20Summary.pdf

Barker SLL, Hickey KA, Cline JS, Dipple GM, Kilburn MR, Vaughan JR, Longo AA (2009) Uncloaking invisible gold: use of nanoSIMS to evaluate gold, trace elements and sulfur isotopes in pyrite from Carlin-type gold deposits. Econ Geol 104:897–904

Bates MN, Rey OA, Biggs ML, Hopenhayn C, Moore LE, Kalman D, Steinmaus C, Smith AH (2004) Case-control study of bladder cancer and exposure to arsenic in Argentina. Am J Epidemiol 159:381–389

Bednar AJ, Garbarino JR, Burkhardt MR, Ranville JF, Wildeman TR (2004) Field and laboratory arsenic speciation methods and their application to natural-water analysis. Water Res 38:355–364

Belkin HE, Zheng B, Finkelman RB (2000) Human health effects of domestic combustion of coal in rural China: a causal factor for arsenic and fluorine poisoning. In: The Second World Chinese conference on geological sciences, extended abstracts with programs, Stanford, pp 522–524

Benner SG, Hansel CM, Wielinga BW, Barber TM, Fendorf S (2002) Reductive dissolution and biomineralization of iron hydroxide under dynamic flow conditions. Environ Sci Technol 36:1705–1711

Berg M, Giger W, Tran HC, Pham HV, Trang PTK, Schertenleib R (2006) Extent and severity of arsenic pollution in Vietnam and Cambodia. In: Naidu R, Smith E, Owens G, Bhattacharya P, Nadebaum P (eds) Managing arsenic in the environment – from soil to human health. CSIRO Publishing, Collingwood

Berg M, Stengel C, Trang PTK, Hung Viet P, Sampson ML, Leng M, Samreth S, Fredericks D (2007) Magnitude of arsenic pollution in the Mekong and Red River Deltas – Cambodia and Vietnam. Sci Total Environ 372:413–425

Berner RA (1981) A new geochemical classification of sedimentary environments. J Sedim Petrol 51:0359–0365

BGS, DPHE (2001) Arsenic contamination of groundwater in Bangladesh. In: Kinniburgh DG, Smedley PL (eds) BGS technical report WC/00/19, 4 vols. British Geological Survey, Keyworth (see www.bgs.ac.uk/Arsenic/Bangladesh)

Bhattacharya P, Chatterjee D, Jacks G (1997) Occurrence of arsenic-contaminated groundwater in alluvial aquifers from delta plains, Eastern India: options for safe drinking water supply. Water Resour Dev 13:79–92

Bhattacharya P, Tandukar N, Neku A, Valero AA, Mukherjee AB, Jacks G (2003) Geogenic arsenic in groundwaters from Terai alluvial plain of Nepal. J Phys IV 107:173–176

Bhattacharya P, Claesson M, Fagerberg J, Bundschuh J, Storniolo ADR, Martin RA, Thir JM, Sracek O (2005) Natural arsenic in the groundwater of the alluvial aquifers of Santiago del Estero Province, Argentina. In: Bundschuh J, Bhattacharya P, Chandrasekharam D (eds) Natural arsenic in groundwater: occurrence, remediation and management. Balkema, Leiden, pp 57–65

Bhattacharya P, Claesson M, Bundschuh J, Sracek O, Fagerberg J, Jacks G, Martin RA, Storniolo AD, Thir JM (2006) Distribution and mobility of arsenic in the Rio Dulce alluvial aquifers in Santiago del Estero Province, Argentina. Sci Total Environ 358:97–120

Bian J, Tang J, Zhang L, Ma H, Zhao J (2012) Arsenic distribution and geological factors in the western Jilin province, China. J Geochem Explor 112:347–356

Blanchard M, Alfredsson M, Brodholt J, Wright K, Catlow CRA (2007) Arsenic incorporation into FeS_2 pyrite and its influence on dissolution: a DFT study. Geochim Cosmochim Acta 71:624–630

Blanes PS, Buchhamer EE, Gimenez MC (2011) Natural contamination with arsenic and other trace elements in groundwater of the Central-West region of Chaco, Argentina. J Environ Sci Health A Tox Hazard Subst Environ Eng 46:1197–1206

Bostick BC, Fendorf S (2003) Arsenite sorption on troilite (FeS) and pyrite (FeS_2). Geochim Cosmochim Acta 67:909–921

Boyle DR, Turner RJW, Hall GEM (1998) Anomalous arsenic concentrations in groundwaters of an island community, Bowen Island, British Columbia. Environ Geochem Health 20:199–212

Breit GN, Foster AL, Sanzalone RF, Yount JC, Whitney JW, Welch AH, Islam MK, Islam MN (2001) Arsenic cycling in eastern Bangladesh: the role of phyllosilicates. Geol Soc Am Abstr 32: A192

Brouste L, Marlin C, Dever L (1997) Geochemistry and residence time estimation of groundwater from the upper aquifer of the Chihuahua desert (Comarca Lagunera, Northern Mexico). Appl Geochem 12:775–786

Burgess WG, Hoque MA, Michael HA, Voss CI, Breit GN, Ahmed KM (2010) Vulnerability of deep groundwater in the Bengal Aquifer System to contamination by arsenic. Nat Geosci 3:83–87

Buschmann J, Berg M, Stengel C, Sampson ML (2007) Arsenic and manganese contamination of drinking water resources in Cambodia: coincidence of risk areas with low relief topography. Environ Sci Technol 41:2146–2152

Cáceres L, Gruttner E, Contreras R (1992) Water recycling in arid regions: Chilean case. Ambio 21:138–144

Castro de Esparza ML (2009) The presence of arsenic in drinking water in Latin America and its effect on public health. In: Bundschuh J, Armienta MA, Birkle P, Bhattacharya P, Matschullat J, Mukherjee AB (eds) Natural arsenic in groundwaters of Latin America. CRC Press/Balkema, Leiden, pp 17–29

Cebrián ME, Albores MA, Garci-Vargas G, Del Razo LM, Ostrosky-Wegman P (1994) Chronic arsenic poisoning in humans. In: Nriagu JO (ed) Arsenic in the environment, Part II: Human health and ecosystem effects. Wiley, New York, pp 93–107

Chakraborti D, Basu GK, Biswas BK, Chowdhury UK, Rahman MM, Paul K, Chowdhury TR, Chanda CR, Lodh D, Ray SL (2001) Characterization of arsenic-bearing sediments in the Gangetic delta of West Bengal, India. In: Chappell WR, Abernathy CO, Calderon RL (eds) Arsenic exposure and health effects IV. Elsevier, Amsterdam, pp 27–52

Chakraborti D, Mukherjee SC, Pati S, Sengupta MK, Rahman MM, Chowdhury UK, Lodh D, Chanda CR, Chakraborty AK, Basul GK (2003) Arsenic groundwater contamination in Middle Ganga Plain, Bihar, India: a future danger? Environ Health Perspect 111:1194–1201

Chakraborti D, Sengupta MK, Rahman MM, Ahamed S, Chowdhury UK, Hossain MA, Mukherjee SC, Pati S, Saha KC, Dutta RN, Quamruzzaman Q (2004) Groundwater arsenic contamination and its health effects in the Ganga-Meghna-Brahmaputra plain. J Environ Monit 6:74N–83N

Chanpiwat P, Sthiannopkao S, Cho KH, Kim K-W, San V, Suvanthong B, Vongthavady C (2011) Contamination by arsenic and other trace elements of tube-well water along the Mekong River in Lao PDR. Environ Pollut 159:567–576

Chen KY, Liu TK (2007) Major factors controlling arsenic occurrence in the groundwater and sediments of the Chianan coastal plain, SW Taiwan. Terr Atmos Ocean Sci 18:975–994

Chen SL, Dzeng SR, Yang MH, Chlu KH, Shieh GM, Wal CM (1994) Arsenic species in groundwaters of the Blackfoot disease areas, Taiwan. Environ Sci Technol 28:877–881

Chen SL, Yeh SJ, Yang MH, Lin TH (1995) Trace element concentration and arsenic speciation in the well water of a Taiwan area with endemic Blackfoot disease. Biol Trace Elem Res 48:263–274

Chitrakar RL, Neku A (2001) The scenario of arsenic in drinking water in Nepal. Department of Water Supply & Sewerage, Kathmandu, p 11

Chouinard A, Paquette J, Williams-Jones AE (2005) Crystallographic controls on trace-element incorporation in auriferous pyrite from the Pascua epithermal high-sulfidation deposit, Chile-Argentina. Can Mineral 43:951–963

Chowdhury TR, Manal BK, Samanta G, Basu GK, Chowdhury PP, Chanda CR, Karan NK, Lodh D, Dhar RK, Das D, Saha KC, Chakraborti D (1997) Arsenic in groundwater in six districts of West Bengal, India: the biggest arsenic calamity in the world: the status report up to August 1995. In: Abernathy CO, Calderon RL, Chappell WR (eds) Arsenic: exposure and health effects. Chapman and Hall, London, pp 93–111

Cook SJ, Levson VM, Giles TR, Jackaman W (1995) A comparison of regional lake sediment and till geochemistry surveys – a case-study from the Fawnie Creek area, Central British Columbia. Explor Min Geol 4:93–110

Cornell RM, Schwertmann U (1996) The iron oxides: structure, properties, reactions, occurrence and uses. VCH, Weinheim

Cumbal L, Bundschuh J, Aguirre ME, Murgueitio E, Tipan I, Chavez C (2009) The origin of arsenic in waters and sediments from Papallacta lake area in Equador. In: Bundschuh J, Armienta MA, Birkle P, Bhattacharya P, Matschullat J, Mukherjee AB (eds) Natural arsenic in groundwaters of Latin America. CRC Press/Balkema, Boca Raton

Cummings DE, Caccavo F, Fendorf S, Rosenzweig RF (1999) Arsenic mobilization by the dissimilatory Fe(III)-reducing bacterium Shewanella alga BrY. Environ Sci Technol 33:723–729

Cummings DE, March AW, Bostick B, Spring S, Caccavo F Jr, Fendorf S, Rosenzweig RF (2000) Evidence for microbial Fe(III) reduction in anoxic, mining-impacted lake sediments (Lake Coeur d'Alene, Idaho). Appl Environ Microbiol 66:154–162

Das D, Chatterjee A, Samanta G, Mandal B, Chowdhury TR, Chowdhury PP, Chanda C, Basu G, Lodh D, Nandi S, Chakraborty T, Mandal S, Bhattacharya SM, Chakraborti D (1994) Arsenic contamination in groundwater in 6 districts of West Bengal, India – the biggest arsenic calamity in the world. Analyst 119:N168–N170

Das D, Chatterjee A, Mandal BK, Samanta G, Chakraborti D, Chanda B (1995) Arsenic in ground-water in 6 districts of West Bengal, India – the biggest arsenic calamity in the world. 2. Arsenic concentration in drinking-water, hair, nails, urine, skin-scale and liver-tissue (biopsy) of the affected people. Analyst 120:917–924

Datta DK, Subramanian V (1997) Texture and mineralogy of sediments from the Ganges-Brahmaputra-Meghna river system in the Bengal basin, Bangladesh and their environmental implications. Environ Geol 30:181–188

Davison W (1993) Iron and manganese in lakes. Earth Sci Rev 34:119–163

de Vitre R, Belzile N, Tessier A (1991) Speciation and adsorption of arsenic on diagenetic iron oxyhydroxides. Limnol Oceanogr 36:1480–1485

Deditius AP, Utsunomiya S, Renock D, Ewing RC, Ramana CV, Becker U, Kesler SE (2008) A proposed new type of arsenian pyrite: composition, nanostructure and geological significance. Geochim Cosmochim Acta 72:2919–2933

Del Razo LM, Arellano MA, Cebrian ME (1990) The oxidisation states of arsenic in well-water from a chronic arsenicism area of Northern Mexico. Environ Pollut 64:143–153

Del Razo LM, Hernandez JL, Garcia-Vargas GG, Ostrosky-Wegman P, Cortinas de Nava C, Cebrian ME (1994) Urinary excretion of arsenic species in a human population chronically exposed to arsenic via drinking water. A pilot study. In: Chappell WR, Abernathy CO, Cothern CR (eds) Arsenic exposure and health. Science and Technology Letters, Northwood, pp 91–100

Deng Y, Wang Y, Ma T (2009) Isotope and minor element geochemistry of high arsenic groundwater from Hangjinhouqi, the Hetao Plain, Inner Mongolia. Appl Geochem 24:587–599

Di Benedetto F, Costagliola P, Benvenuti M, Lattanzi P, Romanelli M, Tanelli G (2006) Arsenic incorporation in natural calcite lattice: evidence from electron spin echo spectroscopy. Earth Planet Sci Lett 246:458–465

Dixit S, Hering JG (2003) Comparison of arsenic(V) and arsenic(III) sorption onto iron oxide minerals: implications for arsenic mobility. Environ Sci Technol 37:4182–4189

Dudas MJ (1984) Enriched levels of arsenic in post-active acid sulfate soils in Alberta. Soil Sci Soc Am J 48:1451–1452

Dzombak DA, Morel FMM (1990) Surface complexation modelling – hydrous ferric oxide. Wiley, New York

Ellis AJ, Mahon WAJ (1977) Chemistry and geothermal systems. Academic, New York

Farias SS, Casa VA, Vazquez C, Ferpozzi L, Pucci GN, Cohen IM (2003) Natural contamination with arsenic and other trace elements in ground waters of Argentine Pampean Plain. Sci Total Environ 309:187–199

Fendorf S, Michael HA, van Geen A (2010) Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. Science 328:1123–1127

Focazio MJ, Welch AH, Watkins SA, Helsel DR, Horng MA (2000) A retrospective analysis of the occurrence of arsenic in groundwater resources of the United States and limitations in drinking water supply characterizations. U.S. Geological Survey, Denver, p 11

Fontaine JA (1994) Regulating arsenic in Nevada drinking water supplies: past problems, future challenges. In: Chappell WR, Abernathy CO, Cothern CR (eds) Arsenic exposure and health. Science and Technology Letters, Northwood, pp 285–288

Fredrickson JK, Zachara JM, Kennedy DW, Dong H, Onstott TC, Hinman NW, Li S-M (1998) Biogenic iron mineralization accompanying the dissimilatory reduction of hydrous ferric oxide by a groundwater bacterium. Geochim Cosmochim Acta 62: 3239–3257

Fujii R, Swain WC (1995) Areal distribution of selected trace elements, salinity, and major ions in shallow ground water, Tulare Basin, Southern San Joaquin Valley, California. US geological survey water-resources investigations report 95-4048. USGS

Garai R, Chakraborty A, Dey S, Saha K (1984) Chronic arsenic poisoning from tube-well water. J Indian Med Assoc 82:34–35

Goldberg S (1986) Chemical modeling of arsenate adsorption on aluminum and iron oxide minerals. Soil Sci Soc Am J 50:1154–1157

Goldberg S, Glaubig RA (1988) Anion sorption on a calcareous, montmorillonitic soil – arsenic. Soil Sci Soc Am J 52:1297–1300

Gomez JJ, Lillo J, Sahun B (2006) Naturally occurring arsenic in groundwater and identification of the geochemical sources in the Duero Cenozoic Basin, Spain. Environ Geol 50:1151–1170

Gomez ML, Blarasin MT, Martinez DE (2009) Arsenic and fluoride in a loess aquifer in the central area of Argentina. Environ Geol 57:143–155

Guerin WF, Blakemore RP (1992) Redox cycling of iron supports growth and magnetite synthesis by *Aquaspirillum magnetotacticum*. Appl Environ Microbiol 58:1102–1109

Guo HR, Chen CJ, Greene HL (1994) Arsenic in drinking water and cancers: a brief descriptive view of Taiwan studies. In: Chappell WR, Abernathy CO, Cothern CR (eds) Arsenic exposure and health. Science and Technology Letters, Northwood, pp 129–138

Guo XJ, Fujino Y, Kaneko S, Wu KG, Xia YJ, Yoshimura T (2001) Arsenic contamination of groundwater and prevalence of arsenical dermatosis in the Hetao plain area, Inner Mongolia, China. Mol Cell Biochem 222:137–140

Guo HM, Wang YX, Shpeizer GM, Yan SL (2003) Natural occurrence of arsenic in shallow groundwater, Shanyin, Datong Basin, China. J Environ Sci Health A Toxic/Hazard Subst Environ Eng 38:2565–2580

Gurung J, Ishiga H, Khadka MS (2005) Geological and geochemical examination of arsenic contamination in groundwater in the Holocene Terai Basin, Nepal. Environ Geol 49:98–113

Gurung JK, Ishiga H, Khadka MS, Shrestha NR (2007) The geochemical study of fluvio-lacustrine aquifers in the Kathmandu Basin (Nepal) and the implications for the mobilization of arsenic. Environ Geol 52:503–517

Gurzau ES, Gurzau AE (2001) Arsenic exposure from drinking groundwater in Transylvania, Romania: an overview. In: Chappell WR, Abernathy CO, Calderon RL (eds) Arsenic exposure and health effects IV. Elsevier, Amsterdam, pp 181–184

Gustafsson JP, Tin NT (1994) Arsenic and selenium in some Vietnamese acid sulfate soils. Sci Total Environ 151:153–158

Harvey CF, Swartz CH, Badruzzaman ABM, Keon-Blute N, Yu W, Ali MA, Jay J, Beckie R, Niedan V, Brabander D, Oates PM, Ashfaque KN, Islam S, Hemond HF, Ahmed MF (2002) Arsenic mobility and groundwater extraction in Bangladesh. Science 298:1602–1606

Harvey CF, Ashfaque KN, Yu W, Badruzzaman ABM, Ali MA, Oates PM, Michael HA, Neumann RB, Beckie R, Islam S, Ahmed MF (2006) Groundwater dynamics and arsenic contamination in Bangladesh. Chem Geol 228:112–136

Herbel M, Fendorf S (2005) Transformation and transport of arsenic within ferric hydroxide coated sands upon dissimilatory reducing bacterial activity. In: O'Day PA, Vlassopoulos D, Meng X, Benning LG (eds) Advances in arsenic research: integration of experimental and observational studies and implications for mitigation. American Chemical Society, Washington, DC, pp 77–90

Herbel M, Fendorf S (2006) Biogeochemical processes controlling the speciation and transport of arsenic within iron coated sands. Chem Geol 228:16–32

Hering J, Kneebone PE (2002) Biogeochemical controls on arsenic occurrence and mobility in water supplies. In: Frankenberger W (ed) Environmental chemistry of arsenic. Marcel Dekker, New York, pp 155–181

Hopenhayn-Rich C, Biggs ML, Fuchs A, Bergoglio R, Tello EE, Nicolli H, Smith AH (1996) Bladder-cancer mortality associated with arsenic in drinking water in Argentina. Epidemiology 7:117–124

Horneman A, van Geen A, Kent D, Mathe PE, Zheng Y, Dhar RK, O'Connell S, Hoque M, Aziz Z, Shamsudduha M, Seddique AA, Ahmad KM (2004) Decoupling of As and Fe release to Bangladesh groundwater under reducing conditions. Part 1: Evidence from sediment profiles. Geochim Cosmochim Acta 68:3459–3473

Hsu KH, Froines JR, Chen CJ (1997) Studies of arsenic ingestion from drinking water in northeastern Taiwan: chemical speciation and urinary metabolites. In: Abernathy CO, Calderon RL, Chappell WR (eds) Arsenic exposure and health effects. Chapman and Hall, London, pp 190–209

Huang JH, Matzner E (2006) Dynamics of organic and inorganic arsenic in the solution phase of an acidic fen in Germany. Geochim Cosmochim Acta 70:2023–2033

IMTA (1992) Estudio hidroquímico y isotópico del acuífero granular de la Comarca Lagunera. Instituto Mexicano de Technologia del Agua, Cuernavaca, México

Islam FS, Gault AG, Boothman C, Polya DA, Charnock JM, Chatterjee D, Lloyd JR (2004) Role of metal-reducing bacteria in arsenic release from Bengal delta sediments. Nature 430:68–71

Islam FS, Boothman C, Gault AG, Polya DA, Lloyd JR (2005a) Potential role of the Fe(III)-reducing bacteria Geobacter and Geothrix in controlling arsenic solubility in Bengal delta sediments. Mineral Mag 69:865–875

Islam FS, Pederick RL, Gault AG, Adams LK, Polya DA, Charnock JM, Lloyd JR (2005b) Interactions between the Fe(III)-reducing bacterium *Geobacter sulfurreducens* and arsenate, and capture of the metalloid by biogenic Fe(II). Appl Environ Microbiol 71:8642–8648

Junta Municipal (2000) Prospección hidrogeologica e hidrogeoquímica en el bolsón de la Mesilla, Chihuahua (Zona de Conejos-Medanos), Report Técnico No DS GEO-011/00. Junta Municipal de Agua y Saneamiento de Juárez

Karcher S, Caceres L, Jekel M, Contreras R (1999) Arsenic removal from water supplies in Northern Chile using ferric chloride coagulation. J Chart Inst Water Environ Manage 13:164–169

Klump S, Kipfer R, Cirpka OA, Harvey CF, Brennwald MS, Ashfaque KN, Badruzzaman ABM, Hug SJ, Imboden DM (2006) Groundwater dynamics and arsenic mobilization in Bangladesh assessed using noble gases and tritium. Environ Sci Technol 40:243–250

Kocar BD, Herbel MJ, Tufano KJ, Fendorf S (2006) Contrasting effects of dissimilatory iron(III) and arsenic(V) reduction on arsenic retention and transport. Environ Sci Technol 40:6715–6721

Kondo H, Ishiguro Y, Ohno K, Nagase M, Toba M, Takagi M (1999) Naturally occurring arsenic in the groundwaters in the southern region of Fukuoka Prefecture, Japan. Water Res 33:1967–1972

Lewis C, Ray DT, Chiu KK (2007) Primary geologic sources of arsenic in the Chianan Plain (Blackfoot disease area) and the Lanyang Plain of Taiwan. Int Geol Rev 49:947–961

Lin TH, Huang YL, Wang MY (1998) Arsenic species in drinking water, hair, fingernails, and urine of patients with blackfoot disease. J Toxicol Environ Health A 53:85–93

Lloyd JR, Oremland RS (2006) Microbial transformations of arsenic in the environment: from soda lakes to aquifers. Elements 2:85–90

Lu FJ, Shih SR, Liu TM, Shown SH (1990) The effect of fluorescent humic substances existing in the well water of Blackfoot disease endemic areas in Taiwan on prothrombin time and activated partial thromboplastin time in vitro. Thromb Res 57:747–753

Luo ZD, Zhang YM, Ma L, Zhang GY, He X, Wilson R, Byrd DM, Griffiths JG, Lai S, He L, Grumski K, Lamm SH (1997) Chronic arsenicism and cancer in Inner Mongolia – consequences of well-water arsenic levels greater than 50 $\mu g\ l^{-1}$. In: Abernathy CO, Calderon RL, Chappell WR (eds) Arsenic exposure and health effects. Chapman and Hall, London, pp 55–68

Maest AS, Pasilis SP, Miller LG, Nordstrom DK (1992) Redox geochemistry of arsenic and iron in Mono Lake, California, USA. In: Kharaka Y, Maest A (eds) Proceedings of the seventh international symposium on water-rock interaction, Park City, Utah, USA, 13–18 July 1992. A.A. Balkema, Rotterdam, pp 507–511

Maharjan M, Watanabe C, Ahmad SA, Umezaki M, Ohtsuka R (2007) Mutual interaction between nutritional status and chronic arsenic toxicity due to groundwater contamination in an area of Terai, lowland Nepal. J Epidemiol Community Health 61:389–394

Manning BA, Goldberg S (1997) Adsorption and stability of arsenic (III) at the clay mineral-water interface. Environ Sci Technol 31:2005–2011

Masscheleyn PH, DeLaune RD, Patrick WH (1991) Effect of redox potential and pH on arsenic speciation and solubility in a contaminated soil. Environ Sci Technol 25:1414–1419

McArthur JM, Ravenscroft P, Safiulla S, Thirwall MF (2001) Arsenic in groundwater: testing pollution mechanisms for sedimentary aquifers in Bangladesh. Water Resour Res 37:109–117

McArthur JM, Banerjee DM, Hudson-Edwards KA, Mishra R, Purohit R, Ravenscroft P, Cronin A, Howarth RJ, Chatterjee A, Talukder T, Lowry D, Houghton S, Chadha DK (2004) Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water: the example of West Bengal and its worldwide implications. Appl Geochem 19:1255–1293

Michael HA, Voss CI (2008) Evaluation of the sustainability of deep groundwater as an arsenic-safe resource in the Bengal Basin. Proc Natl Acad Sci U S A 105:8531–8536

Morales L, Puigdomenech C, Punti A, Torres E, Canyellas C, Cortina JL, Sancha AM (2009) Arsenic and water quality of rural community wells in San Juan de Limay, Nicaragua. In: Bundschuh J, Armienta MA, Birkle P, Bhattacharya P, Matschullat J, Mukherjee AB (eds) Natural arsenic in groundwaters of Latin America. CRC Press/Balkema, Leiden, pp 123–128

Mukherjee AB, Bhattacharya P, Jacks G, Banerjee DM, Ramanathan CM, Chandrashekharam D, Chatterjee D, Naidu R (2006) Groundwater arsenic contamination in India. In: Naidu R, Smith E, Owens G, Bhattacharya P, Nadebaum P (eds) Managing arsenic in the environment – from soil to human health. CSIRO Publishing, Collingwood

Nath B, Mallik SB, Stueben D, Chatterjee D, Charlet L (2010) Electrical resistivity investigation of the arsenic affected alluvial aquifers in West Bengal, India: usefulness in identifying the areas of low and high groundwater arsenic. Environ Earth Sci 60:873–884

Nguyen VA, Bang S, Pham Hung V, Kim K-W (2009) Contamination of groundwater and risk assessment for arsenic exposure in Ha Nam province, Vietnam. Environ Int 35:466–472

Nickson RT, McArthur JM, Ravenscroft P, Burgess WG, Ahmed KM (2000) Mechanism of arsenic release to groundwater, Bangladesh and West Bengal. Appl Geochem 15:403–413

Nicolli HB, Suriano JM, Peral MAG, Ferpozzi LH, Baleani OA (1989) Groundwater contamination with arsenic and other trace-elements in an area of the Pampa, province of Cordoba, Argentina. Environ Geol Water Sci 14:3–16

Nicolli HB, Bundschuh J, Garcia JW, Falcon CM, Jean J-S (2010) Sources and controls for the mobility of arsenic in oxidizing groundwaters from loess-type sediments in arid/semi-arid dry climates Evidence from the Chaco-Pampean plain (Argentina). Water Res 44:5589–5604

Nimick DA, Moore JN, Dalby CE, Savka MW (1998) The fate of geothermal arsenic in the Madison and Missouri Rivers, Montana and Wyoming. Water Resour Res 34:3051–3067

Ning Z, Lobdell DT, Kwok RK, Liu Z, Zhang S, Ma C, Riediker M, Mumford JL (2007) Residential exposure to drinking water arsenic in Inner Mongolia, China. Toxicol Appl Pharmacol 222: 351–356

Nordstrom DK (2002) Public health – worldwide occurrences of arsenic in ground water. Science 296:2143–2144

Nordstrom DK, Alpers CN (1999) Negative pH, efflorescent mineralogy, and consequences for environmental restoration at the Iron Mountain superfund site, California. Proc Natl Acad Sci U S A 96:3455–3462

O'Day PA, Vlassopoulos D, Root R, Rivera N (2004) The influence of sulfur and iron on dissolved arsenic concentrations in the shallow subsurface under changing redox conditions. Proc Natl Acad Sci U S A 101:13703–13708

Oremland RS, Stolz JF (2003) The ecology of arsenic. Science 300:93–94

Oremland RS, Stolz JF (2005) Arsenic, microbes and contaminated aquifers. Trends Microbiol 13:45–49

Oremland R, Newman D, Kail B, Stolz J (2002) Bacterial respiration of arsenate and its significance in the environment. In: Frankenberger W (ed) Environmental chemistry of arsenic. Marcel Dekker, New York, pp 273–295

Oremland RS, Stolz JF, Hollibaugh JT (2004) The microbial arsenic cycle in Mono Lake, California. FEMS Microbiol Ecol 48:15–27

Oscarson DW, Huang PM, Defosse C, Herbillon A (1981a) Oxidative power of Mn(IV) and Fe(III) oxides with respect to As(III) in terrestrial and aquatic environments. Nature 291:50–51

Oscarson DW, Huang PM, Liaw WK (1981b) Role of manganese in the oxidation of arsenite by freshwater lake sediments. Clays Clay Miner 29:219–225

Pant BR (2011) Ground water quality in the Kathmandu valley of Nepal. Environ Monit Assess 178:477–485

Pederick RL, Gault AG, Charnock JM, Polya DA, Lloyd JR (2007) Probing the biogeochemistry of arsenic: response of two contrasting aquifer sediments from Cambodia to stimulation by arsenate and ferric iron. J Environ Sci Health A Tox Hazard Subst Environ Eng 42:1763–1774

Pedersen HD, Postma D, Jakobsen R (2006) Release of arsenic associated with the reduction and transformation of iron oxides. Geochim Cosmochim Acta 70:4116–4129

Plumlee GS, Smith KS, Montour MR, Ficklin WH, Mosier EL (1999) Geologic controls on the composition of natural waters and mine waters draining diverse mineral-deposit types. In: Filipek LH, Plumlee GS (eds) Reviews in economic geology, vol 6B. Society of Economic Geologists, Littleton, pp 373–432

Pokhrel D, Bhandari BS, Viraraghavan T (2009) Arsenic contamination of groundwater in the Terai region of Nepal: an overview of health concerns and treatment options. Environ Int 35:157–161

Polizzotto ML, Harvey CF, Sutton SR, Fendorf S (2005) Processes conducive to the release and transport of arsenic into aquifers of Bangladesh. Proc Natl Acad Sci U S A 102:18819–18823

Polizzotto ML, Harvey CF, Li GC, Badruzzman B, Ali A, Newville M, Sutton S, Fendorf S (2006) Solid-phases and desorption processes of arsenic within Bangladesh sediments. Chem Geol 228:97–111

Polizzotto ML, Kocar BD, Benner SG, Sampson M, Fendorf S (2008) Near-surface wetland sediments as a source of arsenic release to ground water in Asia. Nature 454:505–508

Polya DA, Gault AG, Diebe N, Feldman P, Rosenboom JW, Gilligan E, Fredericks D, Milton AH, Sampson M, Rowland HAL, Lythgoe PR, Jones JC, Middleton C, Cooke DA (2005) Arsenic hazard in shallow Cambodian groundwaters. Mineral Mag 69:807–823

Postma D, Larsen F, Hue NTM, Duc MT, Viet PH, Nhan PQ, Jessen S (2007) Arsenic in groundwater of the Red River floodplain, Vietnam: controlling geochemical processes and reactive transport modeling. Geochim Cosmochim Acta 71:5054–5071

Rahman MM, Chowdhury UK, Mukherjee SC, Mondal BK, Paul K, Lodh D, Biswas BK, Chanda CR, Basu GK, Saha KC, Roy S, Das R, Palit SK, Quamruzzaman Q, Chakraborti D (2001) Chronic arsenic toxicity in Bangladesh and West Bengal, India – a review and commentary. J Toxicol Clin Toxicol 39:683–700

Reich M, Becker U (2006) First-principles calculations of the thermodynamic mixing properties of arsenic incorporation into pyrite and marcasite. Chem Geol 225:278–290

Reimann C, Bjorvatn K, Frengstad B, Melaku Z, Tekle-Haimanot R, Siewers U (2003) Drinking water quality in the Ethiopian section of the East African Rift Valley I – data and health aspects. Sci Total Environ 311:65–80

Riedel FN, Eikmann T (1986) Natural occurrence of arsenic and its compounds in soils and rocks. Wiss Umwelt 3–4:108–117

Robertson FN (1989) Arsenic in groundwater under oxidizing conditions, south-west United States. Environ Geochem Health 11:171–185

Robinson B, Outred H, Brooks R, Kirkman J (1995) The distribution and fate of arsenic in the Waikato River System, North Island, New Zealand. Chem Spec Bioavailab 7:89–96

Romero L, Alonso H, Campano P, Fanfani L, Cidu R, Dadea C, Keegan T, Thornton I, Farago M (2003) Arsenic enrichment in waters and sediments of the Rio Loa (Second Region, Chile). Appl Geochem 18:1399–1416

Romero FM, Armienta MA, Carrillo-Chavez A (2004) Arsenic sorption by carbonate-rich aquifer material, a control on arsenic mobility at Zimapan, Mexico. Arch Environ Contam Toxicol 47:1–13

Rosas I, Belmont R, Armienta A, Baez A (1999) Arsenic concentrations in water, soil, milk and forage in Comarca Lagunera, Mexico. Water Air Soil Pollut 112:133–149

Rosenboom JW (2004) Not just red or green. An analysis of arsenic data from 15 upazilas in Bangladesh. APSU, Arsenic Policy Support Unit, Bangladesh

Rowland HAL, Gault AG, Charnock JM, Polya DA (2005) Preservation and XANES determination of the oxidation state of solid-phase arsenic in shallow sedimentary aquifers in Bengal and Cambodia. Mineral Mag 69:825–839

Rowland HAL, Pederick RL, Polya DA, Pancost RD, Van Dongen BE, Gault AG, Vaughan DJ, Bryant C, Anderson B, Lloyd JR (2007) The control of organic matter on microbially mediated iron reduction and arsenic release in shallow alluvial aquifers, Cambodia. Geobiology 5:281–292

Ryker SJ (2001) Mapping arsenic in groundwater: a real need, but a hard problem – why was the map created? Geotimes 46:34–36

Sampson ML, Bostick B, Chiew H, Hagan JM, Shantz A (2008) Arsenicosis in Cambodia: case studies and policy response. Appl Geochem 23:2977–2986

Savage KS, Tingle TN, O'Day PA, Waychunas GA, Bird DK (2000) Arsenic speciation in pyrite and secondary weathering phases, Mother Lode Gold District, Tuolumne County, California. Appl Geochem 15:1219–1244

Savage KS, Bird DK, O'Day PA (2005) Arsenic speciation in synthetic jarosite. Chem Geol 215:473–498

Scanlon BR, Nicot JP, Reedy RC, Kurtzman D, Mukherjee A, Nordstrom DK (2009) Elevated naturally occurring arsenic in a semiarid oxidizing system, Southern High Plains aquifer, Texas, USA. Appl Geochem 24:2061–2071

Schreiber ME, Simo JA, Freiberg PG (2000) Stratigraphic and geochemical controls on naturally occurring arsenic in groundwater, eastern Wisconsin, USA. Hydrogeol J 8:161–176

Schrenk MO, Edwards KJ, Goodman RM, Hamers RJ, Banfield JF (1998) Distribution of *Thiobacillus ferrooxidans* and *Leptospirillum ferrooxidans*: implications for generation of acid mine drainage. Science 279:1519–1522

Shotyk W (1996) Natural and anthropogenic enrichments of As, Cu, Pb, Sb, and Zn in ombrotrophic versus minerotrophic peat bog profiles, Jura Mountains, Switzerland. Water Air Soil Pollut 90:375–405

Smedley PL, Kinniburgh DG (2002) A review of the source, behaviour and distribution of arsenic in natural waters. Appl Geochem 17:517–568

Smedley PL, Nicolli HB, Macdonald DMJ, Barros AJ, Tullio JO (2002) Hydrogeochemistry of arsenic and other inorganic constituents in groundwaters from La Pampa, Argentina. Appl Geochem 17:259–284

Smedley PL, Zhang M-Y, Zhang G-Y, Luo Z-D (2003) Mobilisation of arsenic and other trace elements in fluviolacustrine aquifers of the Huhhot Basin, Inner Mongolia. Appl Geochem 18: 1453–1477

Smedley PL, Kinniburgh DG, Macdonald DMJ, Nicolli HB, Barros AJ, Tullio JO, Pearce JM, Alonso MS (2005) Arsenic associations in sediments from the loess aquifer of La Pampa, Argentina. Appl Geochem 20:989–1016

Smedley PL, Knudsen J, Maiga D (2007) Arsenic in groundwater from mineralised Proterozoic basement rocks of Burkina Faso. Appl Geochem 22:1074–1092

So HU, Postma D, Jakobsen R, Larsen F (2008) Sorption and desorption of arsenate and arsenite on calcite. Geochim Cosmochim Acta 72:5871–5884

Sracek A, Novak M, Sulovsky P, Martin R, Bundschuh J, Bhattacharya P (2009) Mineralogical study of arsenic-enriched aquifer sediments at Santiago del Estero, Northwest Argentina. In: Bundschuh J, Armienta MA, Birkle P, Bhattacharya P, Matschullat J, Mukherjee AB (eds) Natural arsenic in groundwaters of Latin America. CRC Press/Balkema, Leiden

Stachowicz M, Hiemstra T, Van Riemsdijk WH (2007) Arsenic-bicarbonate interaction on goethite particles. Environ Sci Technol 41:5620–5625

Stollenwerk KG (2003) Geochemical processes controlling transport of arsenic in groundwater: a review of adsorption. In: Stollenwerk KG (ed) Arsenic in ground water: geochemistry and occurrence. Kluwer, Boston, pp 67–100

Stollenwerk KG, Breit GN, Welch AH, Yount JC, Whitney JW, Foster AL, Uddin MN, Majumder RK, Ahmed N (2007) Arsenic attenuation by oxidized aquifer sediments in Bangladesh. Sci Total Environ 379:133–150

Swartz CH, Blute NK, Badruzzman B, Ali A, Brabander D, Jay J, Besancon J, Islam S, Hemond HF, Harvey CF (2004) Mobility of arsenic in a Bangladesh aquifer: inferences from geochemical profiles, leaching data, and mineralogical characterization. Geochim Cosmochim Acta 68:4539–4557

Swedlund PJ, Webster JG (1999) Adsorption and polymerisation of silicic acid on ferrihydrite, and its effect on arsenic adsorption. Water Res 33:3413–3422

Tandukar N, Bhattacharya P, Jacks G, Valero AA (2005) Naturally occurring arsenic in groundwater of Terai region in Nepal and mitigation options. In: Bundschuh J, Bhattacharya P, Chandrasekharam D (eds) Natural arsenic in groundwater: occurrence, remediation and management. A.A. Balkema, Leiden, pp 41–48

Taylor RM (1980) Formation and properties of the Fe(II)Fe(III) hydroxy-carbonate and its possible significance in soil formation. Clay Miner 15:369–382

Thompson JM, Demonge JM (1996) Chemical analyses of hot springs, pools, and geysers from Yellowstone National Park, Wyoming, and vicinity, 1980–1993. USGS Open-File Report

Tseng WP, Chu HM, How SW, Fong JM, Lin CS, Yeh S (1968) Prevalence of skin cancer in an endemic area of chronic arsenicism in Taiwan. J Natl Cancer Inst 40:453–463

Tufano KJ, Fendorf S (2008) Confounding impacts of iron reduction on arsenic retention. Environ Sci Technol 42:4777–4783

van Geen A, Zheng Y, Stute M, Ahmed KM (2003a) Comment on "Arsenic mobility and groundwater extraction in Bangladesh" (II). Science 300:584C

van Geen A, Zheng Y, Versteeg R, Stute M, Horneman A, Dhar R, Steckler M, Gelman A, Small C, Ahsan H, Graziano JH, Hussain I, Ahmed KM (2003b) Spatial variability of arsenic in 6000 tube wells in a 25 km^2 area of Bangladesh. Water Resour Res 39, art. no.-1140

van Geen A, Zheng Y, Cheng Z, Aziz Z, Horneman A, Dhar RK, Mailloux B, Stute M, Weinman B, Goodbred S, Seddique AA, Hope MA, Ahmed KM (2006) A transect of groundwater and sediment properties in Araihazar, Bangladesh: further evidence of decoupling between As and Fe mobilization. Chem Geol 228:85–96

van Geen A, Zheng Y, Goodbred S, Horneman A, Aziz Z, Cheng Z, Stute M, Mailloux B, Weinman B, Hoque MA, Seddique AA, Hossain MS, Chowdhury SH, Ahmed KM (2008) Flushing history as a hydrogeological control on the regional distribution of arsenic in shallow groundwater of the Bengal Basin. Environ Sci Technol 42:2283–2288

Varsányi I, Fodré Z, Bartha A (1991) Arsenic in drinking water and mortality in the southern Great Plain, Hungary. Environ Geochem Health 13:14–22

Wang L, Huang J (1994) Chronic arsenism from drinking water in some areas of Xinjiang, China. In: Nriagu JO (ed) Arsenic in the environment, Part II: Human health and ecosystem effects. Wiley, New York, pp 159–172

Warren C (2001) Hydrogeology and groundwater quality of Los Pereyras, Tucuman, Argentina, University College. University College London, London

Webster JG (1999) Arsenic. In: Marshall CP, Fairbridge RW (eds) Encyclopaedia of geochemistry. Chapman and Hall, London, pp 21–22

Webster JG (2003) Arsenic adsorption on natural riverine suspended particulate material (SPM). Abstr Pap Am Chem Soc 226:U583

Webster JG, Nordstrom DK (2003) Geothermal arsenic. In: Stollenwerk KG (ed) Arsenic in ground water: geochemistry and occurrence. Kluwer, Boston, pp 101–125

Welch AH, Lico MS, Hughes JL (1988) Arsenic in ground-water of the Western United States. Ground Water 26:333–347

Welch AH, Westjohn DB, Helsel DR, Wanty RB (2000) Arsenic in ground water of the United States: occurrence and geochemistry. Groundwater 38:589–604

White DE, Hem JD, Waring GA (1963) Chemical composition of sub-surface waters. In: Fleischer M (ed) Data of geochemistry, 6th edn. Geological Survey, Washington, DC, pp 1–67

Wilkie JA, Hering JG (1996) Adsorption of arsenic onto hydrous ferric oxide: effects of adsorbate/adsorbent ratios and co-occurring solutes. Colloids Surf A Physicochem Eng Asp 107:97–110

Wilkie JA, Hering JG (1998) Rapid oxidation of geothermal arsenic (III) in streamwaters of the eastern Sierra Nevada. Environ Sci Technol 32:657–662

Williams M (1997) Mining-related arsenic hazards: Thailand case-study. British Geological Survey technical report WC/97/49, 36 pp

Williams M, Fordyce F, Paijitprapapon A, Charoenchaisri P (1996) Arsenic contamination in surface drainage and groundwater in part of the southeast Asian tin belt, Nakhon Si Thammarat Province, southern Thailand. Environ Geol 27:16–33

Winkel LHE, Pham Thi Kim T, Vi Mai L, Stengel C, Amini M, Nguyen Thi H, Pham Hung V, Berg M (2011) Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century. Proc Natl Acad Sci U S A 108:1246–1251

Wyatt CJ, Fimbres C, Romo L, Mendez RO, Grijalva M (1998) Incidence of heavy metal contamination in water supplies in Northern Mexico. Environ Res 76:114–119

Xie X, Wang Y, Su C, Liu H, Duan M, Xie Z (2008) Arsenic mobilization in shallow aquifers of Datong Basin: hydrochemical and mineralogical evidences. J Geochem Explor 98:107–115

Xie X, Ellis A, Wang Y, Xie Z, Duan M, So C (2009) Geochemistry of redox-sensitive elements and sulfur isotopes in the high arsenic groundwater system of Datong Basin, China. Sci Total Environ 407:3823–3835

Ying SC, Kocar BD, Griffis SD, Fendorf S (2011) Competitive microbially and Mn oxide mediated redox processes controlling arsenic speciation and partitioning. Environ Sci Technol 45: 5572–5579

Yokoyama T, Takahashi Y, Tarutani T (1993) Simultaneous determination of arsenic and arsenious acids in geothermal water. Chem Geol 103:103–111

Zheng Y, van Geen A, Stute M, Dhar R, Mo Z, Cheng Z, Horneman A, Gavrieli I, Simpson HJ, Versteeg R, Steckler M, Grazioli-Venier A, Goodbred S, Shahnewaz M, Shamsudduha M, Hoque MA, Ahmed KM (2005) Geochemical and hydrogeological contrasts between shallow and deeper aquifers in two villages of Araihazar, Bangladesh: implications for deeper aquifers as drinking water sources. Geochim Cosmochim Acta 69:5203–5218

Fluoride in Natural Waters

13

W. Mike Edmunds and Pauline L. Smedley

Contents

W.M. Edmunds (✉)
School of Geography and the Environment, Oxford University, Oxford, UK
e-mail: wme@btopenworld.com

P.L. Smedley
British Geological Survey, Keyworth, Nottinghamshire, UK

13.1 Introduction

The element fluorine has long been recognised to have benefits for dental health: low-fluoride intake has been linked to development of dental caries and the use of fluoride toothpastes and mouthwashes is widely advocated in mitigating dental health problems. Fluoridation of water supplies to augment naturally low fluoride concentrations is also undertaken in some countries. However, despite the benefits, optimal doses of fluoride appear to fall within a narrow range. The detrimental effects of ingestion of excessive doses of fluoride are also well documented. Chronic ingestion of high doses has been linked to the development of dental fluorosis, and in extreme cases, skeletal fluorosis. High doses have also been linked to cancer (Marshall 1990), although the association is not well-established (Hamilton 1992).

Drinking water is particularly sensitive in this respect because large variations in fluoride concentration exist in water supplies in different areas. Concentrations in natural waters span more than four orders of magnitude, although values typically lie in the 0.1–10 mg L^{-1} range. Where concentrations are high, drinking water can constitute the dominant source of fluorine in the human diet. Concentrations in drinking water of around 1 mg L^{-1} are often taken to be optimal. However, chronic use of drinking water with concentrations above about 1.5 mg L^{-1} is considered to be detrimental to health. The WHO (2011) guideline value for fluoride in drinking water remains at 1.5 mg L^{-1}. Many countries also use this value as a national standard for drinking water, although the standard in China is 1.0 mg L^{-1} (Table 13.1). The US Environmental Protection Agency (EPA) has set the primary standard (enforceable limit) at 4 mg L^{-1} for fluoride in drinking water, although the secondary standard (non-enforceable) for United States drinking water is 2 mg L^{-1}. In Tanzania, the national standard is as high as 8 mg L^{-1}, reflecting the difficulties with compliance in a country with regionally high fluoride concentrations and problems with water scarcity.

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_13, British Geological Survey © NERC 2013

Table 13.1 Regulations and recommendations for fluoride in drinking water from a number of organisations or countries

Institution/nation	Limit/guideline	Value (mg L^{-1})	Comment
WHO	Guideline value (GV)	1.5	2011 guidelines, as previous
U.S. EPA	Maximum contaminant level (MCLG)	4	Enforceable regulation
U.S. EPA	Secondary standard	2	Guideline intended to protect against dental fluorosis; not enforceable
EC	Maximum permissible value	1.5	1998 regulations
Canada	National standard	1.5	
India	National standard	1.5	'Acceptable' limit 1.0 mg L^{-1}
China	National standard	1	
Tanzania	National standard	8	Interim standard

High fluoride concentrations are most often associated with groundwaters as these accumulate fluoride from rock dissolution as well as geothermal sources. Many high-fluoride groundwater provinces have been recognised in various parts of the world, particularly northern China, India, Sri Lanka, Mexico, western USA, Argentina and many countries in Africa. Fluoride removal by water treatment is carried out in some countries. However, as many of the high-groundwater provinces occur in developing countries, fluoride removal practices vary widely and many high-fluoride water sources are used without treatment. As a result, large populations throughout parts of the developing world suffer the effects of chronic endemic fluorosis. Estimates are not well-established, but more than 200 million people worldwide are thought to be drinking water with fluoride in excess of the WHO guideline value. This includes around 66 million in India (Majumdar 2011), 45 million people in China (Wuyi et al. 2002) and some 5 million in Mexico (Diaz-Barriga et al. 1997). The population at risk in Africa is unknown but is also likely to be tens of millions.

Despite the clear evidence for health problems related to fluoride in drinking water and the links between fluoride occurrence and geology, there have been few reviews on the hydrogeochemistry of fluoride. This chapter addresses the hydrogeochemical aspects of fluoride in water, particularly groundwater, and outlines the links with health impacts. The chapter characterises the typical ranges of concentrations of fluoride found in water bodies, along with their distribution, speciation and mechanisms of mobilisation. The principles of fluoride behaviour are also illustrated with case studies from Canada, East Africa, Ghana, India, Sri Lanka and the UK.

13.2 History of Fluoride Research and Links with Health

Chemists first intimated the potential for an association between fluorine and health in the nineteenth century from its variable presence in bones and teeth. In the latter part of that century, fluorine was recommended for administration to pregnant women in the interests of dental health. In 1892, it was suggested that the high incidence of dental caries in England was due to deficiency of fluoride in the diet. However it was not until the first quarter of the twentieth century that a clear association between fluorine in water supplies and dental health was established by medical scientists. The earliest studies in Europe and North America were aided by the fact that populations were less mobile than today and water supplies in rural areas were generally from wells or springs.

The first studies on fluorosis in the United States were carried out in the 1930s (Dean and Elvove 1937). These established high fluoride concentrations in drinking water as a likely factor in fluorosis disease and led to the adoption (US Public Health Service 1943, 1946) of an initial upper limit for fluoride in US drinking-water supplies of 1.0 mg L^{-1}, which was later revised to 1.5 mg L^{-1}. In Britain, the earliest studies (1922) were conducted on school children in Maldon, Essex, where an association was established between mottled teeth and fluoride in Chalk groundwater at concentrations of 4.5–5.5 mg L^{-1} (Ainsworth 1933; Hoather 1953) (see also Chap. 30, this volume).

The recognition early in the twentieth century that certain areas of the USA, UK, and elsewhere had clearly defined patterns of dental caries and dental fluorosis prompted closer examination of the association with the geology and the underlying groundwater supplies. In the early studies, a strong link between caries and low-fluoride water emerged and this led, as early as 1945, to the fluoridation of fluoride-deficient water supplies in Grand Rapids, Michigan (Maier 1950). Fluoridation of some British water supplies was also initiated in the 1950s although the policy has not been adopted widely: today around 10% of the population of England & Wales is supplied with fluoridated water. The requirement for water companies to fluoridate is stipulated by the regional health authority.

There has been much debate over the alleged benefits of fluoridation of drinking-water supplies (Hamilton 1992) and the issue is still strongly contentious (Connett 2007). One of

Table 13.2 Health effects of fluoride concentrations in drinking water (From Dissanayake 1991)

Fluoride concentration range (mg L^{-1})	Chronic health effects
Nil	Limited growth and fertility
0.0–0.5 mg L^{-1}	Dental caries
0.5–1.5 mg L^{-1}	Promotes dental health, prevents tooth decay
1.5–4.0 mg L^{-1}	Dental fluorosis (mottled teeth)
4–10 mg L^{-1}	Dental fluorosis, skeletal fluorosis
>10 mg L^{-1}	Crippling fluorosis

the reasons is that many of the scientific studies are considered to be incomplete or flawed (Marshall 1990). Studies of water fluoridation also yield conflicting results (Hamilton 1992). Some conclude that the evidence for caries prevention favours fluoridation as an effective policy (Yeung 2008). On the other hand, large reductions in the incidence of tooth decay over the last few decades have occurred in many developed countries (Diesendorf 1986; Pizzo et al. 2007), whether or not water fluoridation has been adopted. The reductions have been linked to the large increases in use of fluoride toothpastes and mouthwashes which have increased fluoride intake, together with educational programmes and health promotion campaigns. This undermines the case for fluoridation, at least in industrialised nations (Pizzo et al. 2007). Some studies have even suggested that fluoridation has been detrimental to health (Zan-dao and Yan 2002).

The significance of water-derived fluoride in developed nations has become further obscured since the Second World War as a result of greater variety in diets and the increased mobility of populations. A number of other foodstuffs, including tea, have also been established as key sources of additional fluoride. Today, these sources are augmented to a small extent by exposure from anthropogenic sources such as industrial emissions. Nevertheless, it is clear that exposure to fluoride from drinking water and food remains an important factor.

Despite the uncertain consequences of fluoride in drinking water at low concentrations (0.7 mg L^{-1} or less), the chronic effects of exposure to excessive fluoride in drinking water are well-established. The most common symptom is dental fluorosis ('mottled enamel'), a condition involving interaction of fluoride with tooth enamel, which involves staining or blackening, weakening and possible eventual loss of teeth. With higher exposure to fluoride, skeletal fluorosis can result. This manifests in the early stages as osteosclerosis, involving hardening and calcifying of bones and causes pain, stiffness and irregular bone growth. At its worst, the condition results in severe bone deformation and debilitation. Long-term exposure to fluoride in drinking water at concentrations above about 1.5 mg L^{-1} can result in dental fluorosis (Table 13.2), while values above 4 mg L^{-1} can result in skeletal fluorosis and above about 10 mg L^{-1}, crippling fluorosis can result (Dissanayake 1991). Recent studies have suggested that in arid areas with consumption of large quantities of water, lower concentrations should be appropriate (<1 mg L^{-1}) (Viswanathan et al. 2009). Young children are at greatest risk as the fluoride affects the development of growing teeth and bones. Once developed, the symptoms of fluorosis are irreversible. However, nutrition is also an important factor in the onset of fluorosis disease. Dietary deficiencies in calcium and vitamin C are recognised as important exacerbating factors. Links between high fluoride and other health problems, including birth defects (Hamilton 1992) and cancer (Eyre et al. 2009) are less clearly defined. See also Chap. 30 this volume.

13.3 The Hydrogeochemical Cycle of Fluorine

13.3.1 Atmospheric and Surface Water Inputs

The hydrogeochemical cycle of fluorine is illustrated in Fig. 13.1. The cycle involves transfer of fluorine to the atmosphere by volcanic emissions, evaporation, marine aerosols, and industrial pollution. Wet and dry deposition transfer fluorine to the biosphere and geosphere. In the geosphere, uptake and release of fluorine are controlled by various water-rock interactions and by inputs from anthropogenic sources.

Rainfall constitutes an important component of the cycle. Fluorine sources in rainfall include marine aerosols, volcanic emissions, and anthropogenic introduction of chlorofluorocarbons (CFCs) and industrial emissions. Industrial aerosols are especially produced from coal burning, brick making and aluminium smelting (Fuge and Andrews 1988) and fluorine exposure is also linked to domestic coal burning (Finkelman et al. 1999). Fluoride can substitute to a minor extent for OH^- in minerals, including clay minerals, and this is released (as is the OH^- by dehydroxylation) upon heating in industrial processes. Fluoride concentrations in rainfall are low and accurate data are sparse as a result of analytical difficulties. If marine aerosols contribute to rainfall compositions in their seawater proportions, rainfall in coastal areas with 10 mg L^{-1} Cl should have 0.68 μg L^{-1} F. In many continental areas where rainfall Cl concentrations are near or below 1 mg L^{-1}, fluoride inputs to streams and groundwaters from rainfall should be at or below 0.1 μg L^{-1}. In fact, rainfall fluoride concentrations tend to be higher than these estimates, implying that some fractionation favours greater uptake of the more volatile fluoride at the sea surface or that the ratio is increased by atmospheric fluoride inputs.

The concentrations of fluoride in pristine rainfall are difficult to assess since human impacts are likely to

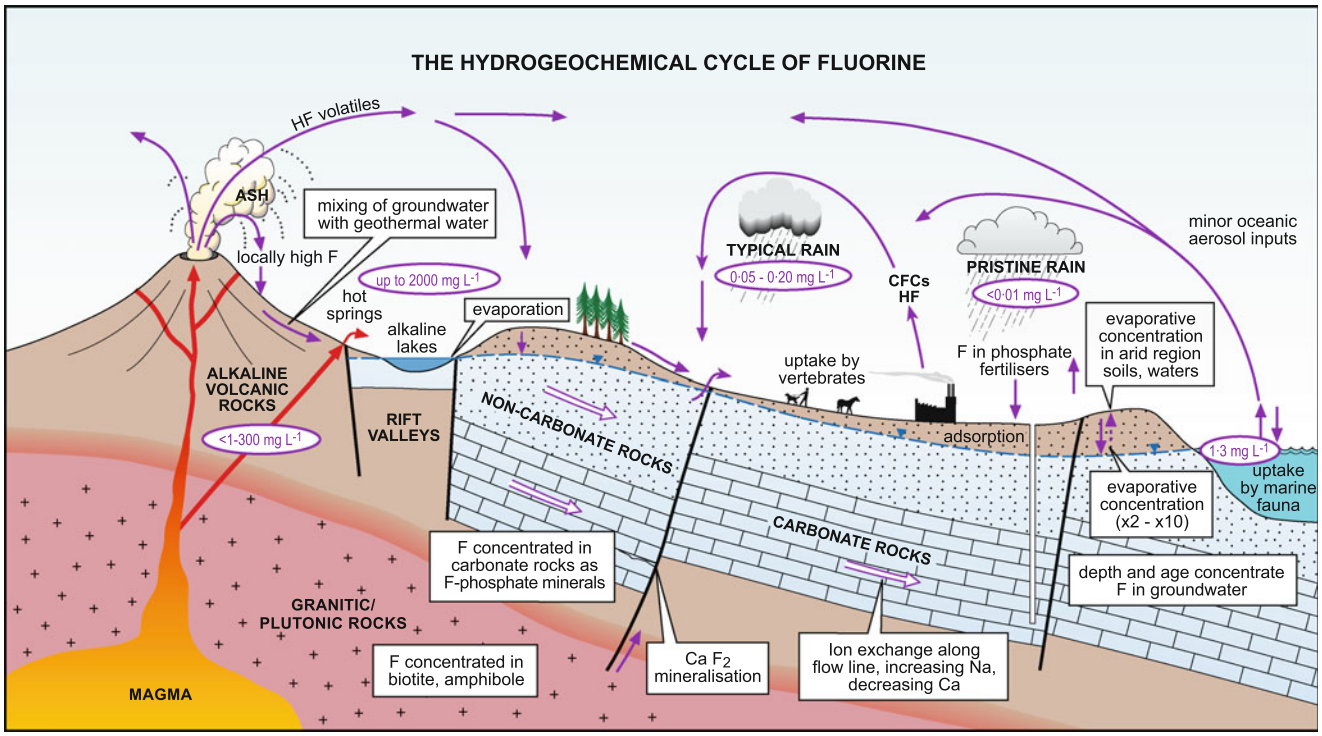


Fig. 13.1 Schematic diagram showing the fluorine hydrogeochemical cycle

dominate in most areas. Comparative studies at inland and coastal sites in Virginia, USA, found median values of 4 and 9 μg L^{-1} respectively (Table 13.3) and marine aerosol inputs were considered to be small (Barnard and Nordstrom 1982). Higher concentrations, in the range <20–80 μg L^{-1} (average 30 μg l^{-1}), have been found in rainfall from Anuradhapura, central Sri Lanka. Such values are likely to reflect a large contribution from marine aerosols. Neal (1989) reported typical fluoride concentrations in rainfall from Wales of 20–70 μg l^{-1}, reflecting in part marine influences. Occasional higher values up to 220 μg l^{-1} were interpreted as increased atmospheric inputs, presumably of anthropogenic origins. Saether and Andreassen (1989) found concentrations in Norwegian precipitation up to 253 μg l^{-1}, with values typically of 13–25 μg l^{-1}. They used the correlation between sulphate and fluoride to suggest that anthropogenic contributions amount to as much as 98% of the total fluoride, derived principally from industrial aerosols.

Concentrations of fluoride in surface waters are generally much higher than in rainfall, though still typically in the μg L^{-1} range. Stream flow in upland areas of Wales has concentrations in the range 30–70 μg L^{-1} (Neal et al. 1997). In the Bengal Basin, the Ganges-Brahmaputra river system has mean fluoride concentrations in the range 66–154 μg L^{-1} (Datta et al. 2000) and rivers from southern Ghana have concentrations of around 30–140 μg l^{-1} (Table 13.3, Botchway et al. 1996). In most cases, concentrations in surface waters are less than 300 μg L^{-1}.

Despite these generally low concentrations, fluoride in surface waters can be much higher in geothermal areas. Many alkaline lakes in the East African Rift Valley for example have concentrations of the order of tens to hundreds of mg L^{-1} (up to 1,980 mg L^{-1}; Table 13.4).

13.3.2 Mineral Sources of Fluoride

Fluorine is more abundant in the earth's crust (625 mg kg^{-1}) than its sister halogen element chlorine (130 mg kg^{-1}). Chloride is highly mobile in the aqueous environment and most is found in the oceans. By contrast, fluorine is mostly retained in minerals. Fluorine is the lightest of the halogen elements and is also the most electronegative. It has an ionic radius very similar to that of OH^- and substitutes readily in hydroxyl positions in late-formed minerals in igneous rocks. It is mobile under high-temperature conditions and being a light, volatile element it is found along with boron and to a lesser extent chloride in hydrothermal solutions. Thus, concentrations of fluoride are generally localised in their geological occurrence. Most is found in acidic igneous rocks, mineralised veins and sedimentary formations where biogeochemical reactions have taken place.

Table 13.3 Ranges of fluoride concentrations in various natural waters

Country	Aquifer	Range of F (mg L^{-1})	Average (mg L^{-1})	No. analyses	Reference
Rainfall					
Norway		0.0–0.253	0.013–0.025		Saether and Andreassen (1989)
USA	Virginia – coastal Virginia – inland	0.002–0.02	0.009 0.004		Barnard and Nordstrom (1982)
Sri Lanka	Anuradhapura	<0.02–0.08	0.03	6	BGS (unpublished data)
UK	Chilton, south-east England		0.096	24	Edmunds et al. (1987)
UK	Mid Wales	0.02–0.22	0.02		Neal (1989)
UK	Loch Fleet, Scotland	<0.05	<0.05	10	Cook et al. (1991)
Surface waters					
India	Ganges River		0.154		Datta et al. (2000)
India	Meghna River		0.066		Datta et al. (2000)
India	Brahmaputra River		0.120		Datta et al. (2000)
UK	Hafren River, Wales	0.03–0.07	0.05		Neal (1989)
Ghana	Rivers, Accra region	0.03–0.14	0.09	5	Botchway et al. (1996)
Ghana	Ponds, Accra region	0.07–0.30	0.18	5	Botchway et al. (1996)
Surface waters in high-fluoride regions					
Tanzania, Kenya, Uganda	Lakes on alkaline volcanic rocks and others	0.2–1,627		200	Kilham and Hecky (1973)
Ethiopia	River Awash	0.9–1.3	1.1	2	Ashley and Burley (1994)
Ethiopia	Rift Valley lakes	1.7–264	54.3	10	Tekle-Haimanot et al. (2006)
Kenya	Streams bordering Lake Magadi	0.1–1.9	0.6	14	Jones et al. (1977)
Kenya	Lake Naivasha	1.7–1.8	1.75	2	Jones et al. (1977)
Tanzania	Lake Magadi	759–1,980	1,281	13	Jones et al. (1977)
Tanzania	Little Magadi Lake	668–754	711	2	Jones et al. (1977)
Tanzania	Lakes, alkaline volcanic area	60–690		9	Nanyaro et al. (1984)
Tanzania	Rivers, alkaline volcanic area	12–26		9	Nanyaro et al. (1984)
Soil water					
UK	Upland Wales	0.02–0.30	0.05		Neal (1989)
Geothermal springs					
Tanzania	Mbulu springs	up to 99			Bugaisa (1971)
Tanzania	Lake Natron thermal springs	330			Bugaisa (1971)
Kenya	Hot springs (>50°C) bordering Lake Magadi	141–166	7	155	Jones et al. (1977)
Kenya	Warm springs (<50°C) bordering Lake Magadi	50–146	14	82	Jones et al. (1977)
Ethiopia	Rift Valley hot springs	to over 13		20	Tekle-Haimanot et al. (2006)
Tunisia	Hot springs	up to 4			Travi (1993)
USA	Western states	0.8–30.8			Nordstrom and Jenne (1977)
USA	Wister mudpots, Salton Sea	14–15			Ellis and Mahon (1977)
USA	Yellowstone National Park	up to 50		140	Deng et al. (2011)
Iceland	Spring, Reykjavik	1.0			Ellis and Mahon (1977)
Iceland	Spring, Hveragerdi	1.1–1.9			Ellis and Mahon (1977)
Kamchatka	Spring	0.8			Ellis and Mahon (1977)
Former USSR	Spring, Makhachkala, Dagestan	0.4			Ellis and Mahon (1977)
New Zealand	Springs	0.3–8.4			Ellis and Mahon (1977)
Taiwan	Spring, Tatun	7.3			Ellis and Mahon (1977)
France	Vichy water	3.8–8.0			Goni et al. (1973)
France	Mont Dore springs	0.07–3.6			Jacob (1975)
France	Plombières springs	1.9–7.0			Fritz (1981)
China	Springs, Shixingsian, Guangdong	up to 45			Fuhong and Shuquin (1988)

(continued)

Table 13.3 (continued)

Country	Aquifer	Range of F (mg L^{-1})	Average (mg L^{-1})	No. analyses	Reference
Tibet, China	Yangi geothermal field	13.0–22.7	19.2	8	Guo et al. (2009)
Portugal	Rio Vouga hot springs	0.04–20.5	11.0	6	Ten Haven et al. (1985)
Groundwater: crystalline basement rocks					
Norway	Caledonian basic and ultrabasic rocks		1.89		Banks et al. (1998)
	Precambrian granite		1.69		Banks et al. (1998)
	Precambrian anorthosite, charnockite		<0.05		Banks et al. (1998)
Norway	Igneous and metamorphic rocks, Hordaland	0.02–9.48	0.30	1,063	Bårdsen et al. (1996)
Cameroon	Granite	0.2–15.2			Fantong et al. (2010)
Ghana	Crystalline basement, including granite and metasediment	0.09–3.8	1.07	118	Smedley et al. (1995)
Senegal	Granitoids, pelites, schists, amphibolites	0.1–3.5			Travi (1993)
India	Crystalline basement, Andhra Pradesh	up to 20			Rao et al. (1974)
India	Crystalline basement, Nalgonda, Andra Pradesh	0.5–7.6	3.6	433	Reddy et al. (2010)
India	Archaean granites and gneisses, east and south-east Karnataka	0.80–7.4	3.5	25	Suma Latha et al. (1999)
India	Archaean banded gneiss, south-east Rajasthan	<0.1–16.2	1.28	2,649	Gupta et al. (1993)
India	Archaean granite	0.3–6.9		188	Vijaya Kumar et al. (1991)
Pakistan	Nagar Parkar, Sindh Province	1.1–7.9	3.33	32	Rafique et al. (2009)
Sri Lanka	Crystalline basement, including granite and charnockite	<0.02–10			Dissanayake (1991)
South Africa	Western Bushveld Complex	0.1–10	143	3,485	McCaffrey (1998)
Groundwater: volcanic rocks					
Ethiopia	Volcanic bedrock (Wonji/Shoa area)	6.1–20.0	12.9	14	Ashley and Burley (1994)
Ethiopia	Pleistocene sediment above volcanic bedrock (Wonji/Shoa)	2.1–4.6	3.4	6	Ashley and Burley (1994)
Ethiopia	Pleistocene sediment above volcanic bedrock (Metahara)	2.7–15.3	5.9	15	Ashley and Burley (1994)
Tanzania	Ngorongoro Crater & Lemagrut volcanic cone	40–140			Bugaisa (1971)
Tanzania	Kimberlites, Shinyanga	110–250			Bugaisa (1971)
Groundwater: sediments and sedimentary basins					
China	Quaternary sands, Hunchun Basin, north-east China	1.0–7.8		19	Woo et al. (2000)
Argentina	Quaternary loess, La Pampa	0.03–29	5.2	108	Smedley et al. (2002)
Argentina	Quaternary loess, Quequen Grande River Basin	0–5.7	1.84	135	Martinez et al. (2012)
India	Quaternary alluvium Agra, Uttar Pradesh	0.11–12.8	2.1	658	Gupta et al. (1999)
UK	Cretaceous Chalk, Berkshire	<0.1–2.4	0.74	22	Edmunds et al. (1989)
	Cretaceous Chalk, London Basin	0.11–5.8	1.44	21	Edmunds et al. (1989)
	Cretaceous Lower Greensand, London	<0.1–0.35	0.17	26	Edmunds et al. (1989)
	Triassic Sandstone, Shropshire	<0.1–0.17	<0.1	40	Edmunds et al. (1989)
	Triassic Sandstone, Lancashire	<0.1–0.14	0.1	16	Edmunds et al. (1989)
	Triassic Sandstone, Cumbria	<0.1–0.26	<0.1	23	Edmunds et al. (1989)
Canada	Carboniferous clastic rocks, Gaspé,	0.02–28	10.9	20	Boyle and Chagnon (1995)
	alluvial/glacial (<30 m depth)		0.09	39	Boyle and Chagnon (1995)
Canada	Non-marine Upper Cretaceous sediments, Alberta Basin	0.01–22.0	1.83	469	Hitchon (1995)
Germany	Cretaceous Chalk Marls	<0.01–8.9	1.28	179	Queste et al. (2001)
Libya	Miocene, Upper Sirte Basin	0.63–3.6	1.4	11	Edmunds (1994)
Sudan	Cretaceous, Nubian Sandstone (Butana area)	0.29–6.2	1.8	9	Edmunds (1994)
Senegal	Palaeocene sediments	1.5–12.5		26	Travi (1993)

(continued)

Table 13.3 (continued)

Country	Aquifer	Range of F (mg L^{-1})	Average (mg L^{-1})	No. analyses	Reference
Senegal	Maastrichtian sediments	1.1–5.0		32	Travi (1993)
Tunisia	Cretaceous to quaternary sediments	0.1–2.3		59	Travi (1993)
USA	Carboniferous sediments, Ohio	0.05–5.9		255	Corbett and Manner (1984)

Table 13.4 Chemical compositions of a range of groundwaters with naturally high fluoride concentrations. The table also gives the modelled dominant species in the waters and fluorite saturation indices, calculated using PHREEQC

Country	Units	Ghana	United Kingdom	Sri Lanka	Argentina	Canada	Tanzania
Location		Yorogu Abagabisi, Bolgatanga	Spalding, Lincolnshire	Paniyankadawala, Anuradhapura	El Cruce, La Pampa	Maria, Gaspé Peninsula	Imalanguzu, Singida
Water type		Unconfined	Confined	Unconfined	Unconfined		Unconfined
Well depth	m	26	99	Nd	16	>30	110
pH		6.63	8.32	6.96	8.1	9.2	8.86
Temperature	°C	31.2	11	29.5			27.7
SEC		348		3,840	2,610	598	1,360
Ca	mg L^{-1}	27.6	2.0	173	24.1	4.3	17.6
Mg	mg L^{-1}	13.7	4.1	179	22.9	0.7	1.37
Na	mg L^{-1}	18.8	540	355	616	137	332
K	mg L^{-1}	1.56	4.0	3.02	12.6	0.4	2.07
HCO_3	mg L^{-1}	146	506	516	1,180	232	845
SO_4	mg L^{-1}	1.65	33	15.5	190	13.6	20.8
Cl	mg L^{-1}	5.89	490	1,050	113	56.7	13.6
NO_3-N	mg L^{-1}	4.92	<0.1	9.00	45.9	nd	7.81
F	mg L^{-1}	3.6	5.6	4.4	15.8	10.9	17.5
Si	mg L^{-1}	34.4	6.0	45.0	28.1	3.5	54.2
Fe	mg L^{-1}	0.37	0.48	0.007	0.055	0.087	0.053
Al	µg L^{-1}	30	14	nd	40	67	81
Be	µg L^{-1}	1	nd	nd	83	nd	2
U	µg L^{-1}	1.3	nd	nd	71	nd	nd
B	mg L^{-1}			0.15	4.58		0.46
Charge imbalance	%	+3.8	+2.01	−0.51	−0.65	−0.06	−2.13
F^-	%	95.98	98.56	79.21	96.54	99.58	99.04
NaF	%	0.04	0.98	0.46	1.04	0.29	0.63
MgF^+	%	2.91	0.44	18.86	2.22	0.09	0.17
CaF^+	%	0.49	0.02	1.46	0.18	0.04	0.16
AlF_3	%	0.35					
AlF_2^+	%	0.18					
HF°	%	0.04					
AlF_4^-	%	0.02					
$BF(OH)_3^-$	%				0.01		
$SI_{fluorite}$	log	−0.32	−1.01	+0.18	+0.75	+0.05	+0.62

Data source for Gaspé Peninsula: Boyle and Chagnon (1995)
SEC specific electrical conductance, *nd* not determined

Fluorine occurs in primary minerals, especially biotites and amphiboles (Fig. 13.2), where it substitutes for hydroxyl positions in the mineral structure. An example is biotite:

$$K_2(Mg,Fe)_4(Fe,Al)_2[Si_6Al_2O_{20}](OH)_2(\mathbf{F},Cl)_2$$

On weathering, the fluorine tends to be released preferentially from these minerals. Where biotites and amphiboles are abundant, as for example in granite, these form a major source of fluoride in water bodies.

Other high-temperature fluorine minerals such as topaz are less soluble. Apatite ($Ca_5(Cl,\mathbf{F},OH)(PO_4)_3$), which may form at both high and low temperatures, is another important source of fluorine (Fig. 13.2). Substituted apatites with high fluorine are more soluble than purer (high-temperature) apatites. Fluorite ($Ca\mathbf{F}_2$) is the main fluorine mineral, which occurs in localised secondary hydrothermal vein deposits and as a relatively rare authigenic mineral in sediments.

In marine sediments, fluorine is concentrated both by adsorption onto clays and also by biogeochemical processes involving the removal of phosphorus. Limestones may contain localised concentrations of fluorapatite, especially francolite. Most sandstones contain very low concentrations of fluorine and hence the fluoride in resident groundwaters may also be low (see also Chap. 2, this volume).

13.3.3 Sources and Reactions in Soils

Typical concentrations of fluoride in soils are around 20–500 mg kg^{-1} (Kabata-Pendias and Pendias 1984). Similar values of 200–400 mg kg^{-1} were given by Fuge and Andrews (1988), although higher concentrations were found in soils from mineralised areas of the UK. Concentrations up to 3,700 mg kg^{-1} were reported in Welsh soils, up to 20,000 mg kg^{-1} in Pennine soils of the English Midlands and up to 3,300 mg kg^{-1} in soils from Cornwall, all associated with hydrothermal mineralisation, including the presence of fluorite. Some plant species in the areas of mineralisation have also increased concentrations of fluoride which are potentially detrimental to humans and grazing animals (Fuge and Andrews 1988).

Typically, only a small amount of the element present naturally in soils (e.g. <10 mg kg^{-1}) is easily soluble, the remainder residing within a variety of minerals (Pickering 1985). Fluoride concentrations have been found in the range 24–1,220 mg kg^{-1} (26 samples) in Argentine soils, the water-soluble fraction being in the range 0.53–8.33 mg kg^{-1}. The soluble fraction in these may be higher than typical because of the fine grain-size of the loess-derived soils (Lavado and Reinaudi 1979). Easily soluble fluoride has also been reported in sodic loess-derived soils from Israel and China (Fuhong and Shuquin 1988; Kafri et al. 1989).

Of the anthropogenic inputs to soils, high fluoride concentrations are found in phosphate fertilisers (8,500–38,000 mg kg^{-1}) (Kabata-Pendias and Pendias 1984) and sewage sludge (80–1,950 mg kg^{-1}) (Rea 1979). Contributions from these can increase the concentrations in agricultural soils considerably.

The amount of fluoride adsorbed by soils varies with soil type, soil pH, salinity and fluoride concentration (Fuhong and Shuquin 1988; Lavado and Reinaudi 1979). Adsorption is favoured in slightly acidic conditions and uptake by acid soils can be up to ten times that of alkaline soils. Fluoride adsorption releases OH^-, though the release is non-stoichiometric. Wenzel and Blum (1992) noted that, while minimum mobility occurs at pH 6.0–6.5, it is increased at pH <6 as a result of the formation of $[AlF]^{2+}$ and $[AlF_2]^+$ complexes in solution. Fine-grained soils also generally retain fluoride better than sandy types. Adsorption is favoured strongly by the presence of freshly precipitated $Fe(OH)_3$ or $Al(OH)_3$. Clay minerals are also effective adsorbents (Wuyi et al. 2002) as is soil organic matter (Fuge and Andrews 1988). In soil profiles containing mainly sand and with little clay, iron or aluminium, up to half of the infiltrating fluoride in water may pass through the soil profile (Pickering 1985). Where continued loading of fluoride occurs from anthropogenic sources, it is likely that the soil-retention capacities may be exceeded and fluoride will percolate to the water table. In most cases however, soils act as a sink rather than a source of fluoride and water reaching the water table is likely to have low fluoride concentrations and be dominated by atmospheric concentrations. Nevertheless, low fluoride concentrations afforded by retention in soils may be offset by evapotranspiration at the soil surface. This may increase fluoride concentrations reaching the water table by up to five times in temperate climates and 10–100 times under semi-arid conditions.

13.3.4 Fluoride in Solution

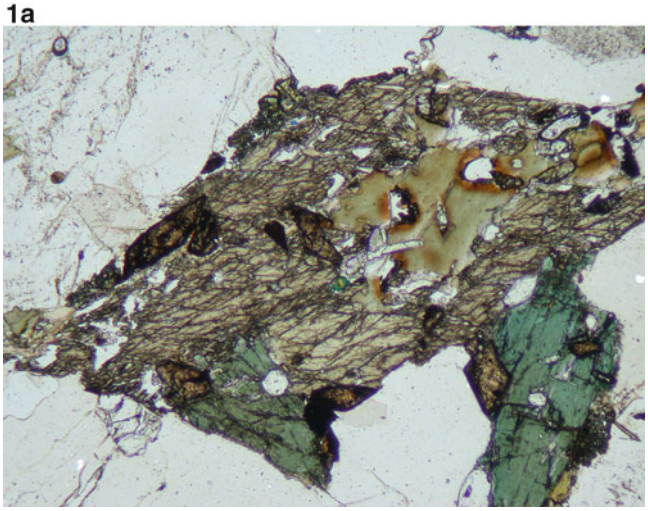
Fluorine occurrence in natural waters is closely related to its abundance in the local minerals and rocks. It is also strongly associated with mineral solubility and in this regard the mineral fluorite, being least soluble and with favourable dissolution kinetics at low temperature, exerts the main control on aqueous concentrations in the natural environment. An upper limit on fluoride activities in aqueous solution is controlled by the solubility product, $K_{fluorite}$:

$$CaF_2 = Ca^{2+} + 2F^- \quad (13.1)$$

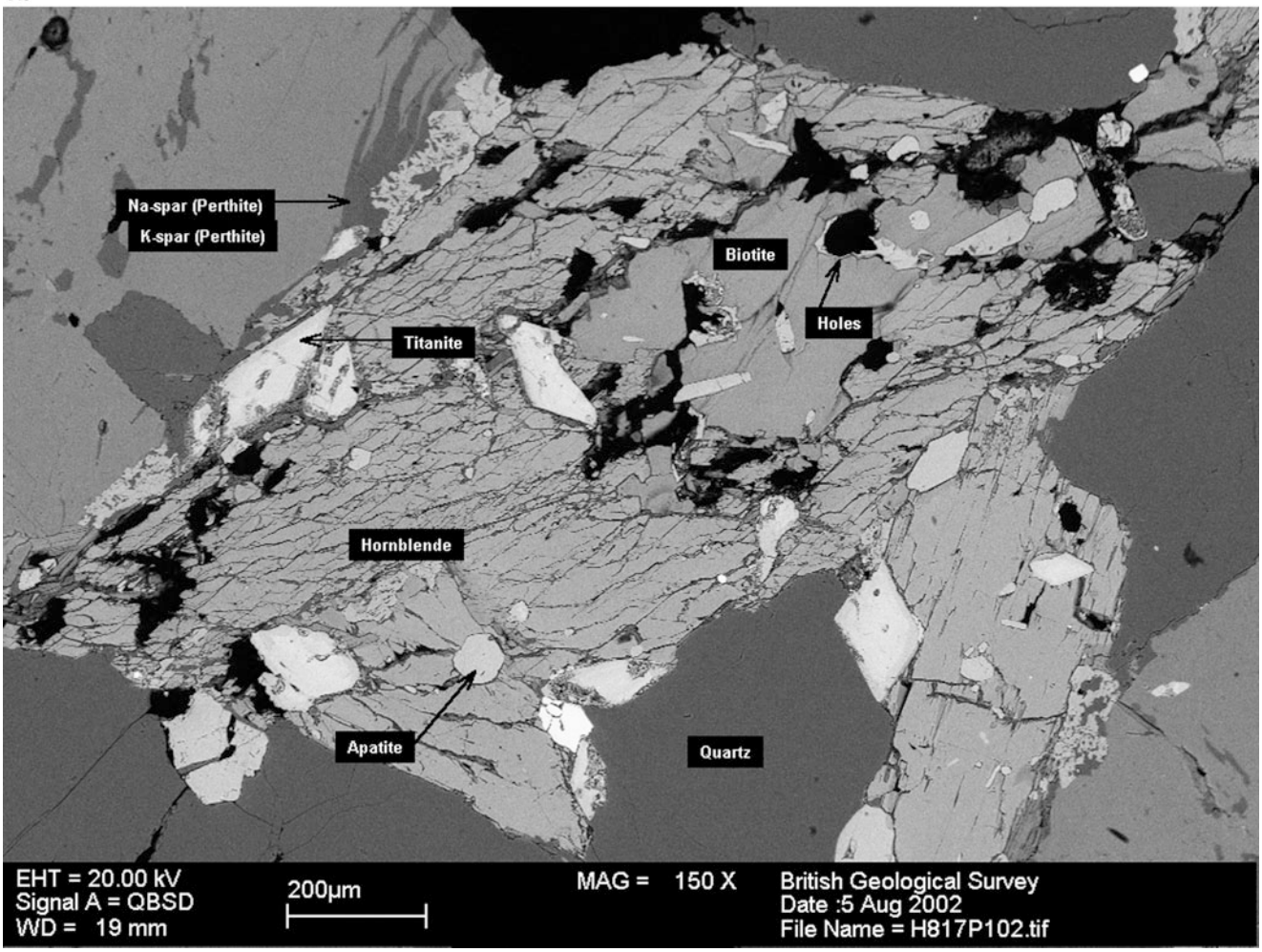
$$K_{fluorite} = (Ca^{2+}) \cdot (F^-)^2 = 10^{-10.57} \text{at } 25^\circ C \quad (13.2)$$

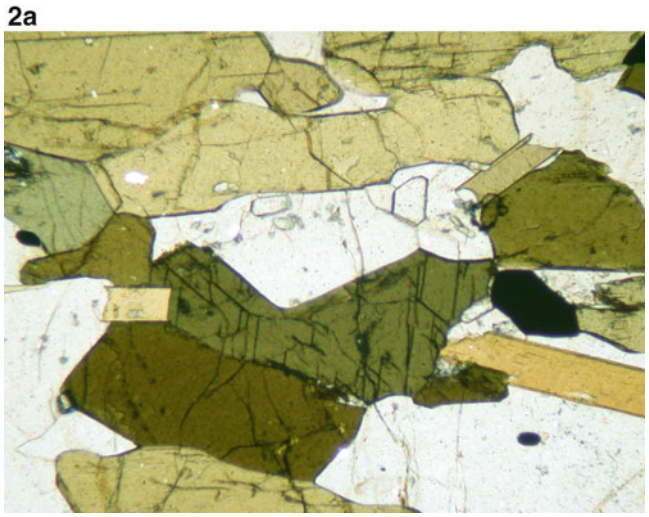
$$\text{or, } \log K_{fluorite} = (Ca^{2+}) + 2\log(F^-) = -10.57 \quad (13.3)$$

1a



1b


Na-spar (Perthite)
K-spar (Perthite)
Biotite
Holes
Titanite
Hornblende
Apatite
Quartz
EHT = 20.00 kV
Signal A = QBSD
WD = 19 mm
200µm
MAG = 150 X
British Geological Survey
Date :5 Aug 2002
File Name = H817P102.tif



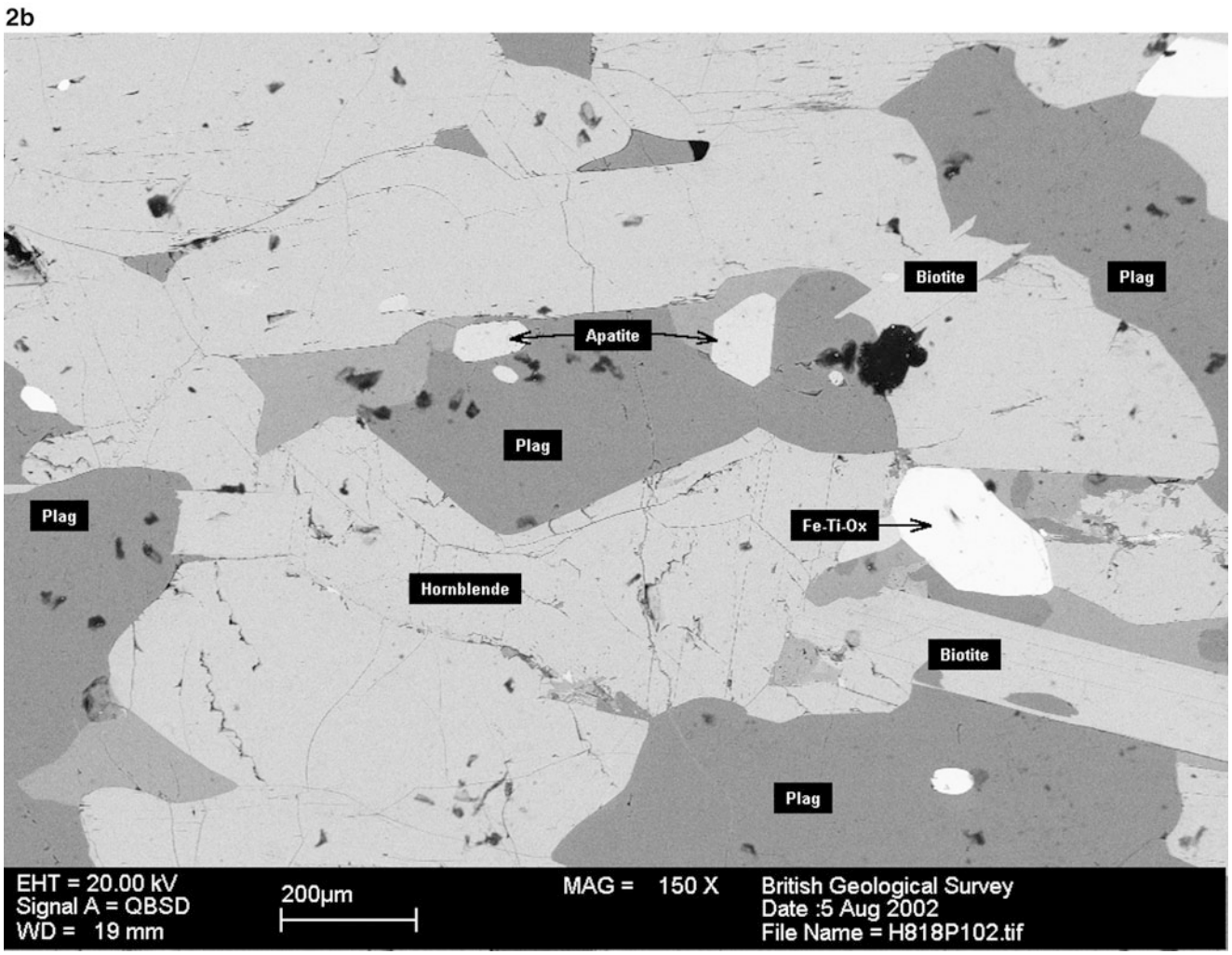


Fig. 13.2 Photomicrographs of typical crystalline basement rocks showing high-fluorine minerals, on the left by plane-polarised light, polars rotated 90°; on the right by back-scattered SEM. *1*: view of *hornblende engulfing biotite, sphene* (*titanite*) and *euhedral apatite; granite*, Bongo, Upper East Region, Ghana. *2*: View of *hornblende, biotite, quartz and apatite; granite,* Anuradhapura, Sri Lanka

This is an important relationship since it shows that in the presence of fluorite, concentrations of fluoride are directly proportional to Ca^{2+} concentrations. For example, in the presence of 10^{-3} M Ca at 25°C, dissolved F^- should be limited to 3.1 mg L^{-1}. It is therefore likely to be the absence of Ca in solution that allows higher concentrations of fluoride to be stable in solution, although precipitation of carbonates has also been linked to loss of fluoride from solution (Reddy et al. 2010; Turner et al. 2005). Low-Ca groundwater conditions arise in volcanic regions dominated by alkaline volcanic rocks (e.g. Ashley and Burley 1994; Kilham and Hecky 1973) and also in conditions where cation exchange occurs naturally (e.g. Handa 1975). Here, removal of Ca^{2+} is achieved by exchange with Na^+ from clay minerals. In both cases, the waters are typically of Na-HCO_3 type. Cation exchange occurs in soils and in groundwaters along flow gradients in response to changing chemistry (Edmunds and Walton 1983) and in zones of saline intrusion.

The solubility of fluorite (Eq. 13.2) is temperature-dependent. Fluorite solubility decreases with temperature and this is reflected with a change in the equilibrium constant. For example the value of $K_{fluorite}$ changes from $10^{-10.57}$ at 25°C to $10^{-10.8}$ at 10°C. Further, the equilibrium constant is affected by the salt content of the waters (the ionic strength) with the solubility increasing with salinity, providing the waters are below saturation with respect to calcite.

Reaction times with aquifer minerals are also important in controlling fluoride concentration. High fluoride concentrations can occur in groundwaters which have long residence times in the host aquifers, probably as a result of diagenetic reactions. Surface waters usually have low concentrations, as do most shallow groundwaters from open wells as they represent young, recently infiltrated, rainwater. Deeper (older) groundwaters from boreholes are therefore most likely to contain high concentrations of fluoride. Exceptions occur in active volcanic areas where surface water and shallow groundwaters are affected by hydrothermal inputs.

High fluoride concentrations are also a feature of arid climatic conditions (Fuhong and Shuquin 1988; Handa 1975; Smedley et al. 2002). Here, groundwater infiltration and flow rates are slow, allowing prolonged reaction times between water and rocks. Fluoride concentrations are lower in the humid tropics because of high rainfall inputs and their diluting effect on groundwater chemical composition.

Table 13.4 shows the chemical compositions of a range of typical high-fluoride groundwaters from various parts of the world. The data illustrate the dominance of Na over Ca in most, though not all, waters and the high concentrations of HCO_3 that typify high-fluoride waters. The table also gives the likely dominant species, calculated using PHREEQC (Parkhurst and Appelo 1999). Free $F^-$is overwhelmingly the dominant form in most natural waters with minor additional amounts of complexes with major cations (Ca, Na, Mg). However, the speciation is strongly pH-dependent. In acidic conditions, strong complexing with H^+ and Al^{3+} can occur (Deng et al. 2011). At pH 3.5, HF° is likely to be the dominant species (Hem 1985). Fluorine also readily forms complexes with B, Be, V, U, Fe^{3+} and Si and these may enhance fluorine mobilisation if present in solution in significant quantity. The formation of complexes may have ramifications for human health for example in aluminium-bearing waters, where the total fluorine may be much higher than the measured ionic fluoride. Thus aluminium fluoride may stabilise the fluoride as a complex ion, but if these complexes are broken down during metabolism they could release both $F^-$and potentially toxic Al.

Table 13.4 gives the saturation indices for fluorite for each of the waters. Fluorite solubility is likely to control the upper limits of fluoride concentrations in most natural waters. The table shows that most are at or near saturation with respect to fluorite, although a couple of samples are significantly oversaturated ($SI_{fluorite}$ >0.5). This presumably reflects the slower kinetics of precipitation of fluorite compared with its dissolution or discrepancies related to mineral purity or grain size.

13.3.5 Fluoride Distributions in Groundwater

One of the best ways to understand the controls on the concentrations of fluoride in groundwaters is to examine its distribution according to geology, depth, salinity and other controlling parameters. Several studies have been made which take large populations of data and examine them statistically, as in Alberta, Canada (Hitchon 1995), India (Gupta et al. 1999), Germany (Queste et al. 2001) and the UK (Edmunds et al. 1989). A statistical summary of fluorine distributions in sedimentary basins, and other aquifers in various parts of the world, is included in Table 13.3.

In Britain, an inter-comparison of fluoride in 13 representative major aquifers was made alongside other trace elements by Edmunds et al. (1989). These results are summarised in Fig. 13.3 as a cumulative-frequency plot. A clear distinction is seen between groundwaters in sandstones and those in carbonate aquifers. A very narrow range of concentrations is found in all Triassic sandstones and median values do not exceed 150 μg L^{-1}. For most of the sandstone groundwaters, the fluoride concentrations do not significantly exceed the rainfall input concentrations after allowing for evapotranspiration. The only aquifer to deviate from the general trend is the Moray Sandstone in Scotland in which fluorite is recorded as a cementing phase associated

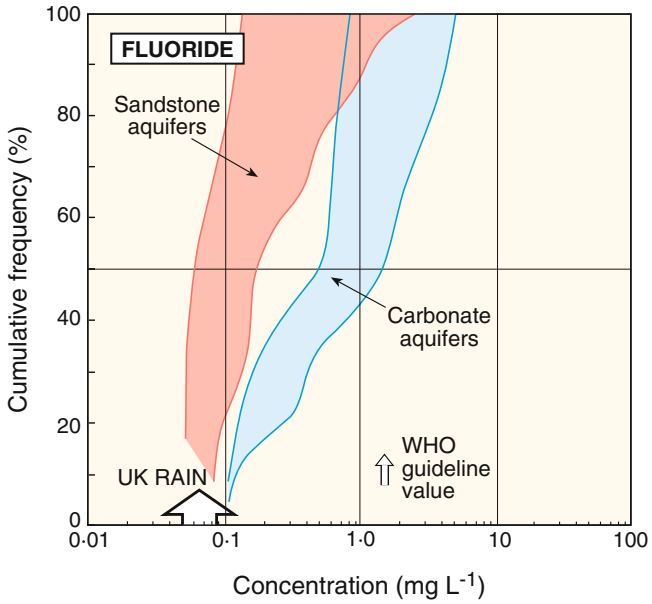


Fig. 13.3 Cumulative-frequency diagram showing the distribution of fluoride in groundwater from *carbonate* and *non-carbonate aquifers* in Britain (Modified from Edmunds et al. 1989)

with vein mineralization and may also be associated with occurrence of phosphate minerals (Edmunds et al. 1989). Median concentrations in limestones lie between 0.6 and 1.5 μg L^{-1}, the higher values found in fine-grained chalk. Although rainfall concentrations provide a contribution, the main source of the fluoride here is biogenic and closely related to phosphate-enriched horizons. The highest fluoride concentrations in British groundwaters are found in confined carbonate aquifers where the combination of a geological source of fluorine (CaF_2), residence time and some calcium depletion due to Na-Ca exchange leads to fluoride concentrations controlled by fluorite mineral solubility. In the majority of British groundwaters, fluorite solubility is not achieved.

13.4 High-Fluoride Groundwater Provinces

Although the concentrations of fluoride are likely to be higher in groundwater than in other water bodies as a result of water-rock interaction, the concentrations in most groundwaters are below the upper concentrations considered detrimental to health, as shown by the British studies in Sect. 13.5. However, aquifers with high groundwater fluoride concentrations have been recognised in a number of regions across the world. The distribution of documented cases where fluoride occurs substantially in wells at concentrations >1.5 mg L^{-1} is shown in Fig. 13.4. Endemic fluorosis is a problem in many, though not all, of these regions. High-fluoride groundwaters are found in many parts of the developing world in particular, and many millions of people rely on groundwater with concentrations above the WHO guideline value for their normal drinking-water supply. Worst-affected areas are arid parts of China, India, Sri Lanka, West Africa (Ghana, Ivory Coast, Senegal), North Africa (Libya, Sudan, Tunisia), South Africa, the East African Rift Valley (Kenya, Uganda, Tanzania, Ethiopia, Rwanda), northern Mexico and central Argentina (Fig. 13.4). Problems have also been reported in parts of Pakistan. In the early 1980s, it was estimated that around 260 million people in 30 countries worldwide were drinking water with more than 1 mg L^{-1} of fluoride (Smet 1990). In India alone, endemic fluorosis is a major problem in 17 out of the country's 22 states.

High-fluoride groundwaters are typically of Na-HCO_3 type with relatively low Ca concentrations (<20 mg L^{-1} or so), and with neutral to alkaline pH values (around 7–9). Fluoride problems are largely found in groundwater from basement aquifers, particularly granites, where fluorine-rich minerals are abundant; in active volcanic zones where fluorine is derived from the volcanic rocks and geothermal sources, and groundwaters in some sediments, particularly in arid areas.

13.4.1 Basement Aquifers

Fluoride problems have been found in crystalline basement rocks, particularly those of granitic composition from several areas across the world (Table 13.3). Granitic rocks contain a relatively large proportion of high-fluorine minerals such as micas, apatite and amphiboles. Fluorite is also an occasional accessory mineral in these rocks.

Basement aquifers in large parts of southern Asia are known to suffer from severe fluoride and fluorosis problems. In India, the worst-affected states are Rajasthan, Andhra Pradesh, Uttar Pradesh, Tamil Nadu and Karnataka (Handa 1975; Maithani et al. 1998; Rao 2002; Reddy et al. 2010; Suma Latha et al. 1999). Fluorosis has also been reported in Assam (Chakraborti et al. 2000; Kotoky et al. 2008). In Pakistan, high fluoride concentrations have also been reported in Sindh Province (Rafique et al. 2009) (Table 13.3). In Sri Lanka, the Dry Zone in the eastern and north-central parts of the country has groundwater with high concentrations up to 10 mg L^{-1}, and associated dental and possibly skeletal fluorosis. In the Wet Zone in the west, groundwater fluoride concentrations are low, probably as a result of intensive rainfall and long-term leaching of fluorine from rocks (Dissanayake 1991). Here, the incidence of dental caries is reported to be high.

Several countries in Africa also have high groundwater fluoride concentrations in basements areas, including parts

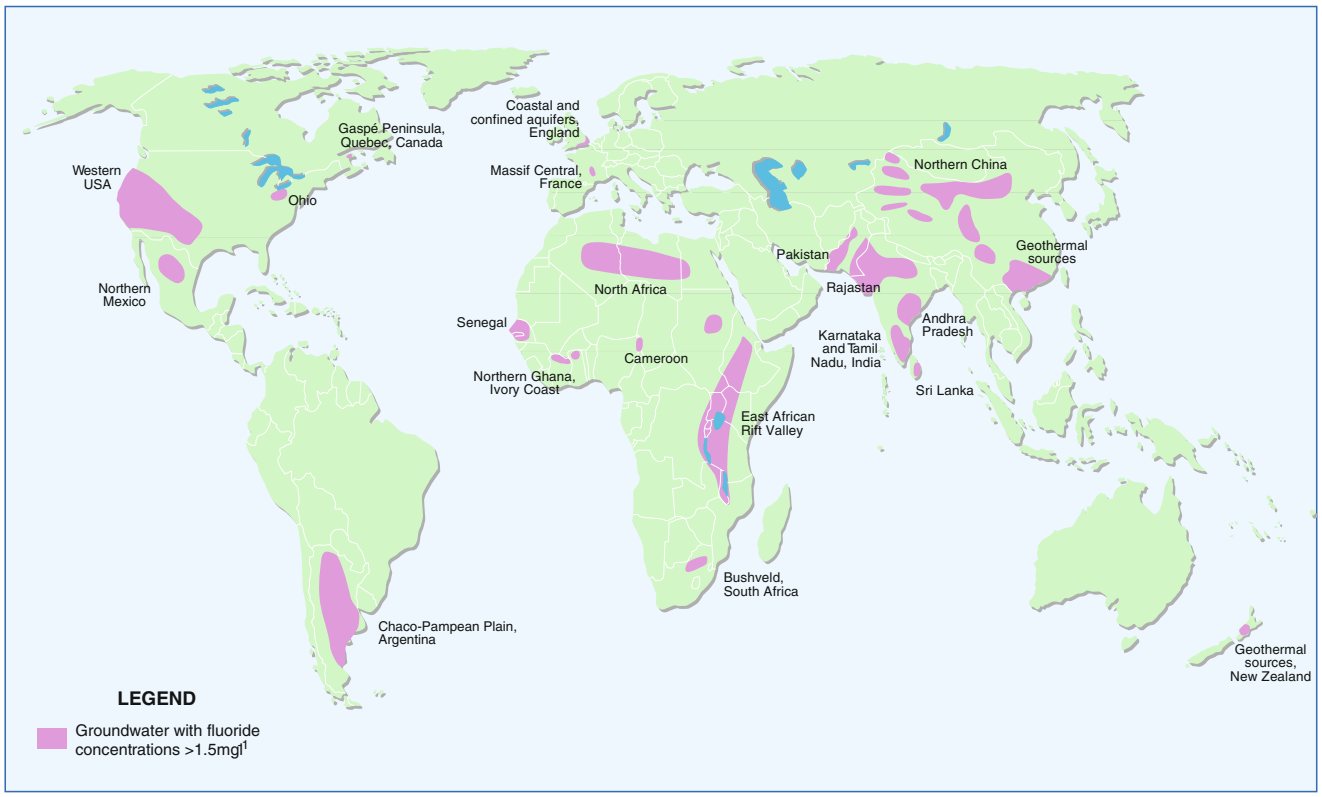


Fig. 13.4 World map of documented high-fluoride (>1.5 mg L^{-1}) groundwaters

of Cameroon, Ghana, Ethiopia, Malawi, Senegal, Tanzania and South Africa (e.g. Fantong et al. 2010; McCaffrey 1998; Tekle-Haimanot et al. 2006; Travi 1993).

13.4.2 Volcanic Areas and Geothermal Sources

High fluoride concentrations have been reported in geothermal sources and active volcanic belts in the western USA (Nordstrom and Jenne 1977), Iceland, Taiwan, Tibet, New Zealand, the former Soviet Union (Ellis and Mahon 1977), France, Algeria and Tunisia (Travi 1993) and the East African Rift Valley (Ayenew 2008). The most common types of geothermal water are alkali-chloride solutions with near-neutral pH values. In these, fluoride concentrations are typically in the range 1–10 mg L^{-1} (Ellis and Mahon 1977) and the waters also commonly have increased concentrations of Si and B, as well as often high As, NH_3 and H_2S. Under acidic conditions, concentrations in geothermal sources can reach more than 1,000 mg L^{-1} (Ellis 1973), with dissolved fluoride in the form of HF, $HF_2^-$and SiF_6^{2-}.

In New Zealand, fluoride concentrations in deep waters have been reported in the range 1–12 mg L^{-1} with maximum concentrations being influenced by subsurface temperature and the solubility of fluorite (Mahon 1964). The main volcanic province with high fluoride waters, the East African Rift Valley, is outlined as a case study below.

13.4.3 Sedimentary Aquifers

Some sedimentary aquifers also have groundwaters with high fluoride concentrations, particularly in arid and semi-arid regions. In La Pampa Province of central Argentina, high-pH (>8), HCO_3-rich groundwaters have concentrations of dissolved fluoride up to 29 mg L^{-1} (Smedley et al. 2002). The concentrations of fluoride are positively correlated with pH and HCO_3 and suggest that the fluoride was mobilised at high pH from F-bearing minerals present in the sediments. These contain rhyolitic volcanic ash. Fluoride concentrations were highly variable, but highest concentrations were found in groundwater from shallow depths (top 20 m). High-fluoride groundwaters are also observed in Quaternary sediment aquifers from arid regions of China. Here, soil salinity and pH have been highlighted as key factors in fluoride accumulation in groundwater (Fuhong and Shuquin 1988).

A number of large sedimentary basins have notable increases in fluoride concentration with increasing groundwater residence time. As groundwaters flow downgradient in a typical sedimentary aquifer, Na for Ca exchange takes

place and groundwaters evolve from Ca-HCO_3 to Na-HCO_3 types. Under low-Ca conditions in old groundwaters, fluoride can be maintained at relatively high concentrations unlimited by the precipitation of CaF_2. Fluoride concentrations therefore often show an increase in groundwaters down the flow line and with increasing depth as a result of increasing residence time (Edmunds and Walton 1983; Travi 1993). A case study of the English Lincolnshire Limestone is given in Sect. 13.5.

Edmunds (1994) reported fluoride concentrations up to 3.6 mg L^{-1} in old groundwaters from the Upper Sirte Basin of Libya and up to 6.2 mg L^{-1} in groundwaters from the Cretaceous Nubian Sandstone of Sudan (Table 13.3). The concentrations were variable in each aquifer but were found to be higher in marine sediments than in freshwater sediments, reflecting the overall higher concentrations of fluorine in the sediments themselves. Despite the high dissolved concentrations, most were undersaturated with respect to fluorite as a result of low dissolved Ca concentrations.

Groundwater from Cretaceous to Tertiary aquifers of western Senegal has fluoride concentrations ranging between <0.01–13 mg L^{-1} (Travi 1993). The sediments form part of an 8,000 m thick sequence forming a sedimentary basin, which overlies Precambrian basement. In the aquifers, fluoride concentrations generally increase in the direction of groundwater flow. The source of the dissolved fluoride is taken to be phosphatic horizons (containing fluorapatite), particularly in argillaceous deposits. In the Maastrichtian aquifer, fluoride concentrations range between 1.1 and 5.0 mg L^{-1} with increases corresponding to a facies change from sandy to more argillaceous deposits (Travi 1993). The high concentrations are present in groundwater up to 30,000 years old.

In the sequence of Cretaceous to Quaternary sediments of western Tunisia, Travi (1993) recorded fluoride concentrations between 0.1 and 2.3 mg L^{-1}. Highest concentrations were found in groundwaters from the Cretaceous 'Complexe Terminal' aquifer (average 1.7 mg L^{-1}). The Tunisian sediments are also phosphatic, although the relationships between phosphate occurrence and dissolved fluoride concentrations were found to be less clear than observed in Senegal.

13.5 Case Studies to Illustrate High-Fluoride Groundwater Systems

A number of aquifers have been selected as case studies to illustrate the different geochemical controls on fluoride mobilisation most clearly and where possible associated health problems may also be explained.

13.5.1 Rajasthan, India – Pre-Cambrian Basement Aquifer

The distribution of fluoride in India has been reviewed by Handa (1975). He recognised Rajasthan as the state most seriously affected by high fluoride, although other high-fluoride areas were also recognised in the country. The distribution of fluoride in Rajasthan (11 districts) was also reviewed by Gupta et al. (1993). Although there are many accounts of the incidence of dental and skeletal fluorosis in the province, it is difficult to find studies linking to regional geology. However, studies in Sirohi district (one of the administrative districts of Rajasthan) seem typical of the rather complex geology of the region (Maithani et al. 1998) and illustrate the nature of the fluoride occurrence.

Rajasthan has an arid climate with low but highly variable annual rainfall. In Sirohi district, fluoride concentrations up to 16 mg L^{-1} have been found in groundwater from dug wells and boreholes at depths between 25 and 75 m during geochemical exploration for uranium (Maithani et al. 1998). The aquifer comprises Proterozoic metasediments with intrusions of granite and rhyolite. Significant fluorite mineralisation associated with the granites and volcanic rocks is also reported from the adjacent Jalore district (Fig. 13.5). This geological setting seems typical of much of Rajasthan where a mixed assemblage of basement rocks outcrops as islands between a highly weathered series of younger Quaternary sediments. The association of fluoride-endemic areas with the bedrock geology is often obscure and it is unclear whether anomalies are related to primary bedrock or secondary enrichment in the sediments.

In the 150 km^2 studied by Maithani et al. (1998), 117 samples were collected and over 75% of these contained groundwater with a fluoride concentration in excess of 1.5 mg L^{-1}. A good correlation with the bedrock geology is observed, with highest concentrations being found in association with the granites, acid volcanic rocks and basic dykes (Fig. 13.5). These dykes act as barriers, which slow down groundwater flow and permit prolonged contact times to raise concentrations of groundwater fluoride. Low-fluoride areas in the east (<1.5 mg L^{-1}) are associated with carbonate rocks and higher calcium may inhibit an excess of fluoride. It is not clear from the study to what extent the alluvium acts as a source (or a conduit) for the fluoride-bearing water and no groundwater flow data are shown.

Handa (1975) recognised a negative relationship between calcium and fluoride in the groundwaters of the Sirohi region (Fig. 13.6). Groundwaters from both bedrock and alluvium showed the relationship. The findings demonstrated the solubility control of fluorite on the fluoride concentrations.

The ill-effects of high fluoride on human health, including dental fluorosis and a few cases of skeletal fluorosis,

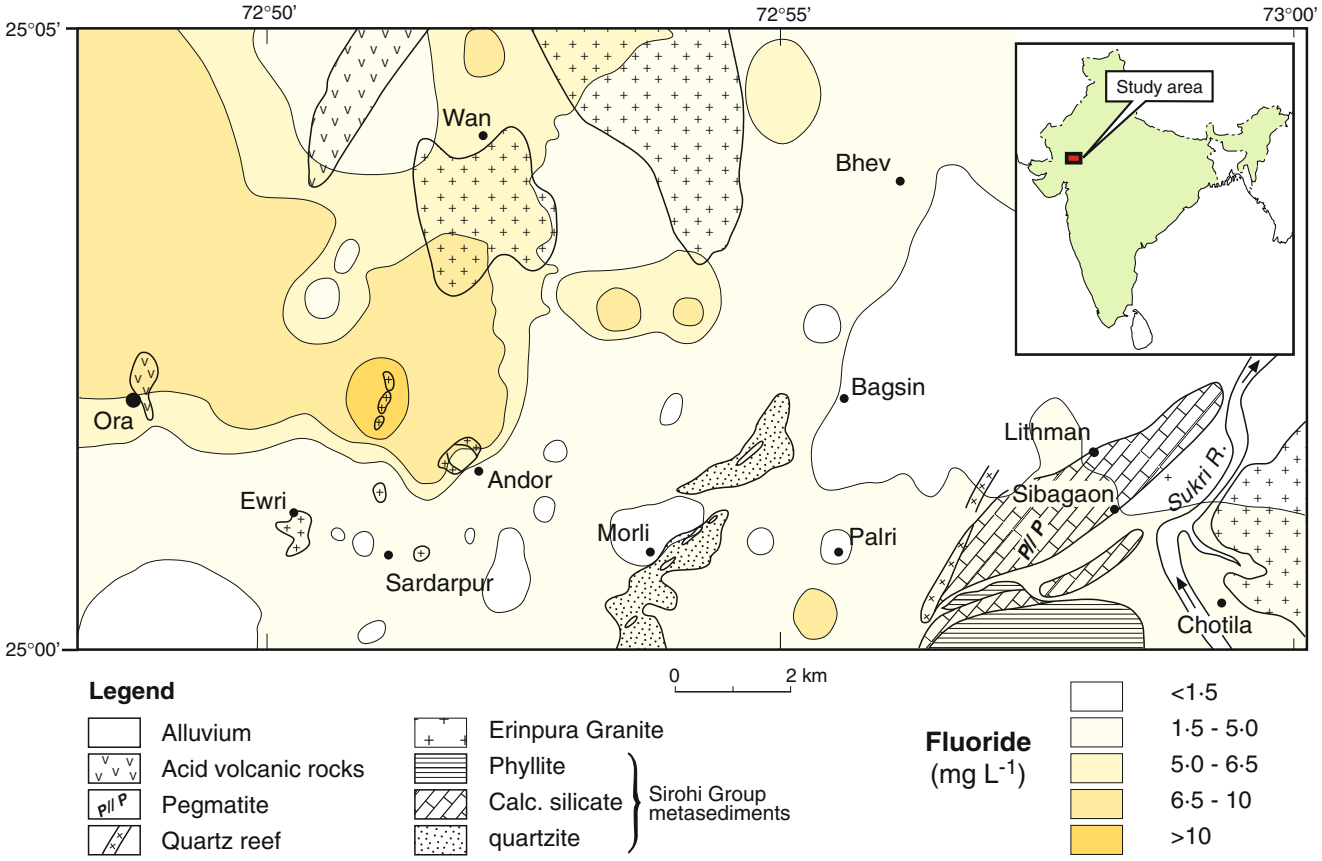


Fig. 13.5 Geological map of parts of Sirohi district, Rajasthan, India with contoured groundwater-fluoride concentrations (Reproduced with permission from Maithani et al. 1998)

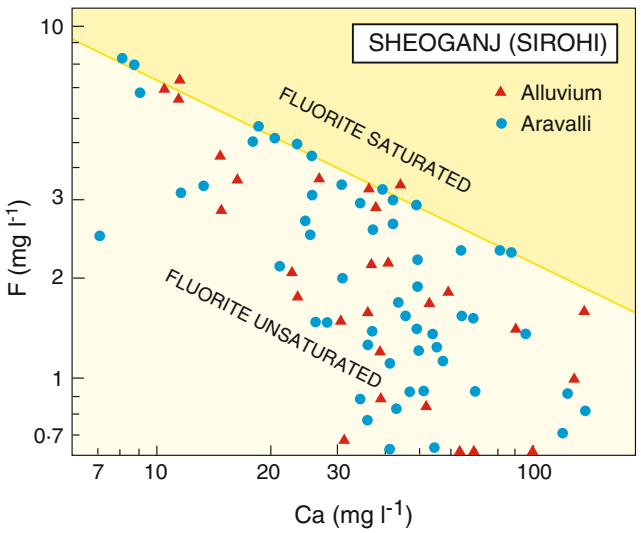


Fig. 13.6 Variation of calcium with fluoride in groundwaters from the Sirohi District of Rajasthan (From data provided by Handa 1975), along with the saturation curve for fluorite. Groundwaters from alluvium and Aravalli Hills (*bedrock*) are distinguished

were observed in villages in the high-fluoride region of Sirohi, although complete medical statistics have not been reported (Handa 1975; Maithani et al. 1998). Although the Maithani et al. (1998) study indicates a strong association between geology and groundwater-fluoride occurrence, a full appraisal of the situation is limited by the absence of hydrogeological information, especially water levels and flow data. Additional hydrochemical data would also assist identification of the causes of high and low fluoride.

13.5.2 Northern Ghana – Alkaline Granitic Basement

The Bolgatanga area in the Upper East Region of northern Ghana demonstrates the often-strong influence of geology, specifically certain types of granite, on fluoride concentrations in groundwater. The region experiences maximum temperatures in the range 20–40°C and average annual rainfall around 1,000 mm (Murray 1960). The climate is semi-arid, largely as a result of northeasterly harmattan (dry, desert) winds, which affect the region from December to March. Geology comprises crystalline basement rocks, including Upper Birimian (Precambrian) rocks of mixed meta-igneous and meta-sedimentary origin.

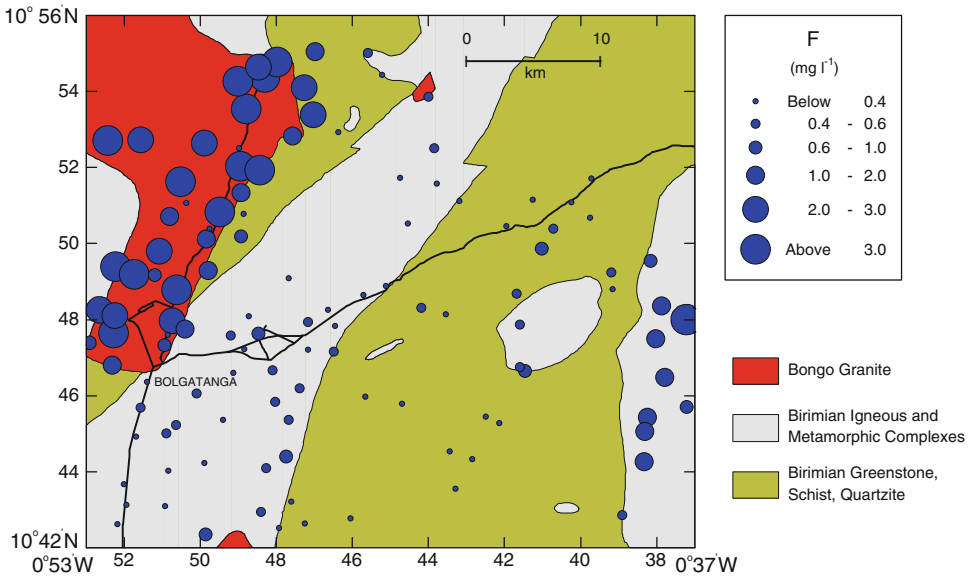


Fig. 13.7 Geological map of the Bolgatanga area of Upper East Region, Ghana, showing the distribution of fluoride concentrations in groundwater. The concentrations are highest where associated with outcrops of biotite- and hornblende-bearing granite and granodiorite (From Smedley et al. 1995)

Intruded into these are a suite of coarse granites, the Bongo Granite suite, and associated minor intrusions, the ages of which are unknown. The Bongo Granite comprises typically a pink microcline-hornblende-granite with interstitial quartz, plagioclase and biotite (Fig. 13.2). The granite is slightly alkaline, with the amphibole appearing to be in the hornblende-arfvedsonite range (Murray 1960). Accessory mineral phases include abundant sphene as well as magnetite, apatite, zircon, rutile and more rarely fluorite. The Bongo Granite has been found to contain up to 0.2% fluorine (Smedley et al. 1995). The Tongo Granite just to the south of the study area (Fig. 13.7) is similar in composition (Murray 1960). To the east of the study area (Sekoti district), a north–south tract of granodiorite has been classed together with other hornblende-biotite granodiorites (Fig. 13.7 and Murray 1960). However, in outcrop the rock type forms rounded tors resembling those of the Bongo Granite and some workers have grouped the Sekoti Granodiorite with the Bongo suite. The Sekoti Granodiorite comprises mainly plagioclase and quartz with biotite, often replacing amphibole, and accessory apatite and sphene (Murray 1960).

Groundwater is abstracted for drinking water mostly from shallow hand-pumped boreholes (typically 30–50 m deep), although some traditional shallow hand-dug wells also occur with depths of less than 10 m. Surveys of groundwater quality (Apambire et al. 1997; Smedley et al. 1995) indicate clearly that highest fluoride concentrations occur in groundwater from the Bongo area in the west, with values reaching up to 3.8 mg L^{-1}, and from Sekoti district in the east, with values reaching up to 3.2 mg L^{-1} (Fig. 13.7). The higher concentrations from any given rock type are generally found in borehole waters. Water from shallow hand-dug wells typically (but not always) has very low fluoride concentrations (<0.4 mg L^{-1}) as these have had much shorter reaction times with the host rocks. They also circulate within the superficial weathered overburden layer rather than the fractured granite at greater depth. From this association, it is possible that shallower groundwaters in such fluoride-vulnerable zones are safer sources of drinking water with respect to fluoride than deeper sources. However, the potential for bacterial contamination is increased in such open, shallow wells.

Villagers in some parts of the Bolgatanga area suffer from dental fluorosis as a result of the high concentrations of dissolved fluoride. The problem is particularly prevalent in Bongo and Sekoti districts, although the numbers of people affected and the regional extent is not known in detail. The fluorosis is particularly prevalent among children. There is to date no evidence of skeletal fluorosis in the region.

13.5.3 The English Lincolnshire Limestone – Downgradient Evolution

A common situation in many aquifers is to encounter increasing fluoride concentrations in the groundwaters down the flow gradient. This situation arises initially as a result of continuous dissolution of fluoride from minerals in the carbonate aquifers, up to the limit of fluorite solubility, followed by ion-exchange reactions involving removal of calcium. The phenomenon is particularly well-illustrated in the Jurassic Lincolnshire Limestone aquifer of eastern England (Edmunds 1973; Edmunds and Walton 1983).

The gently-dipping oolitic Lincolnshire Limestone is typically 30 m thick and downgradient becomes confined beneath clays and marls. Groundwater flow is predominantly through fractures, although there is also an intergranular porosity (13–18%). The unaltered limestone at depth is grey-green in colour due to the presence of Fe^{2+}. However, oxidation of the limestone on a geological timescale has produced an orange-brown oxidised zone through groundwater action. This is also reflected in an oxidation-reduction 'barrier' in the groundwaters, with aerobic groundwaters near to outcrop and reducing waters (with some dissolved iron) at depth.

Fluoride concentrations in the limestone at outcrop are quite low (up to 0.2 mg L^{-1}). The relationship between fluoride and other major ions and iodide is shown in Fig. 13.8. The groundwaters are essentially of calcium-bicarbonate type and are saturated with calcite, buffering the concentrations of Ca and HCO_3 over the first 12 km of the flow line and producing 'hard' water. Thereafter, the profile shows distinctive ion exchange, with Ca^{2+} being replaced by $2Na^{+}$. This produces a 'soft', sodium- bicarbonate groundwater (with a high pH, up to 8.5) and increased bicarbonate concentrations. In these waters, the Ca^{2+} falls to as low as 1 mg L^{-1}, yet the groundwater adjusts under the new physicochemical conditions to maintain saturation with calcite.

It can be seen that the fluoride concentration increases progressively across the aquifer and this increase starts in the calcite-buffered part of the aquifer. The progressive decrease in concentration of Ca^{2+} allows the continuing rise in fluoride concentrations to a maximum of 5.6 mg L^{-1}. The fluoride increase is not controlled by the increase in salinity down-gradient in the aquifer; an increase in chloride concentration only begins some 17 km from outcrop. It is interesting to see that the other halogen measured in this study, iodide, follows the chloride behaviour and not the water-rock interaction pathway shown clearly by fluoride. The source of fluoride in this case is thought to be traces of phosphate minerals in the impure limestone.

This situation was utilised by the water engineers in the 1950s, noting "by a fortunate geological accident, it is possible to obtain in the geographical centre of the district a pure unaggressive water which has in effect been softened, fluoridated and transported free of charge by nature" (Lamont 1958). Today, most of the public water supply in the region is derived from the unconfined and semi-confined sections of the aquifer with lower natural concentrations of fluoride. There are additional water-supply problems in the Lincolnshire Limestone with the need to manage the high nitrate concentrations from agricultural chemicals although nature again assists the management of this problem in the confined aquifer as nitrate is reduced naturally through redox reactions.

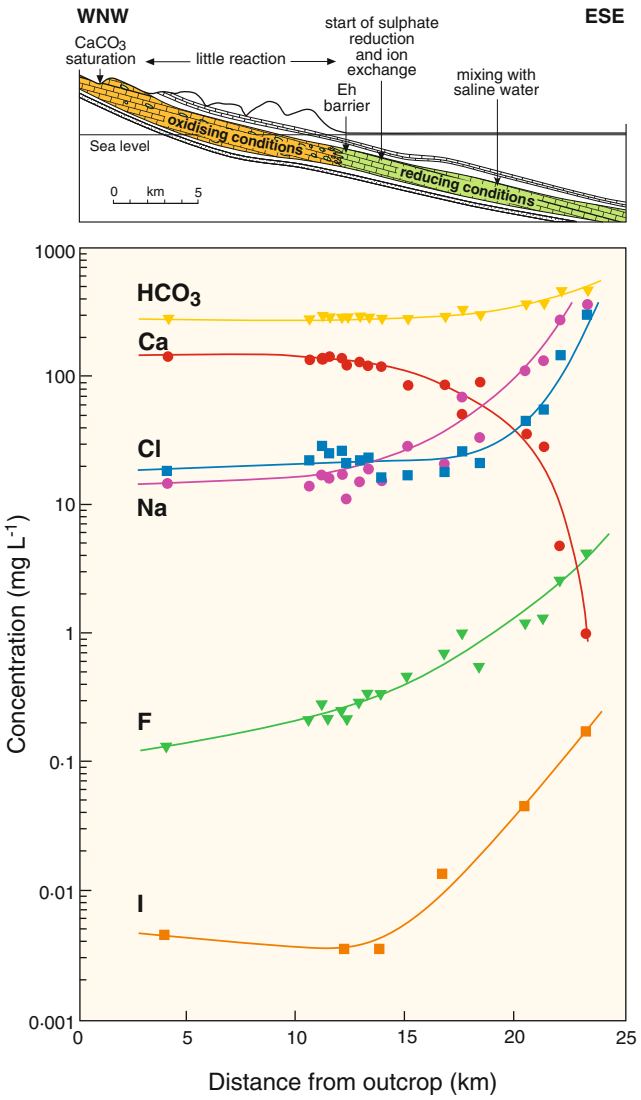


Fig. 13.8 Profiles of changing groundwater chemistry with increasing distance from the outcrop zone of the Lincolnshire Limestone, eastern England (After Edmunds and Walton 1983). A geological cross section across the aquifer is also shown

13.5.4 Sri Lanka – Climatic and Geological Controls

Sri Lanka consists of over 90% metamorphic rocks of Precambrian age. These include metasedimentary and metavolcanic rocks, gneisses and granitoids. Within these are abundant fluorine-bearing minerals such as micas, hornblende and apatite, as well as less common fluorite, tourmaline, sphene and topaz. As a result, large parts of Sri Lanka can be considered a fluorine province. Concentrations of fluoride up to 10 mg L^{-1} have been recorded in the groundwater and the incidence of dental fluorosis among children is high (Dissanayake 1991). Groundwater compositions across the island indicate that

fluoride problems have a strong geographical control linked to both geology and climate (Fig. 13.9).

Dissanayake (1991) compiled a hydrogeochemical atlas for Sri Lanka, based on the results of analyses of 1,970 water samples from shallow wells. It can be seen from the fluoride map produced (Fig. 13.9) that high concentrations lie in the eastern and north central regions of the country. The central hill country and the south-west coastal regions have relatively low fluoride concentrations. Although rocks containing fluoride-rich minerals underlie almost the whole country (except the Jaffna Peninsula), areas with mean annual rainfall above 500 mm experience few fluoride problems. In these wetter areas, there is a tendency for the soluble ions, including F^-, to be leached out under the effects of high rainfall. Indeed, the Wet Zone of Sri Lanka is found to have quite high incidence of dental caries, presumably related to fluoride deficiency. In the Dry Zone, evaporation brings the soluble ions to the surface by capillary action. Despite these generalities, the precise composition of the rock types needs to be considered since even in the Dry Zone, fluoride concentrations vary considerably. In particular, within the Dry Zone the high concentrations coincide with rocks rich in mafic minerals such as hornblende- and biotite-gneisses, charnockites, marble and calc-gneisses (Dissanayake 1991).

The high-fluoride groundwaters of Sri Lanka have many characteristics in common with high-fluoride aquifers elsewhere. In an investigation of groundwater chemistry in Anuradhapura, north central Sri Lanka, groundwaters were found to have typically near-neutral pH values (6.0–7.9), with high alkalinity (alkalinity as HCO_3 up to 713 mg L^{-1}; BGS, unpublished data; 123 samples). Groundwater from many wells also has relatively high salinity (SEC up to 3,850 µS cm^{-1}, average 1,260 µS cm^{-1}), the high salinity values reflecting the importance of water-rock reactions and evapotranspiration in the semi-arid climatic conditions.

The excess or deficiency of fluoride in groundwater is of special concern in Sri Lanka since 75–80% of the population depend on wells without treatment. Unlike high iron for example, fluoride has no taste and a high percentage of the population who suffer from dental fluorosis were unaware of the problem until the later stages of the disease. The use of fluoride toothpaste has alleviated the condition for those suffering from dental caries, but by 1991 at least, no

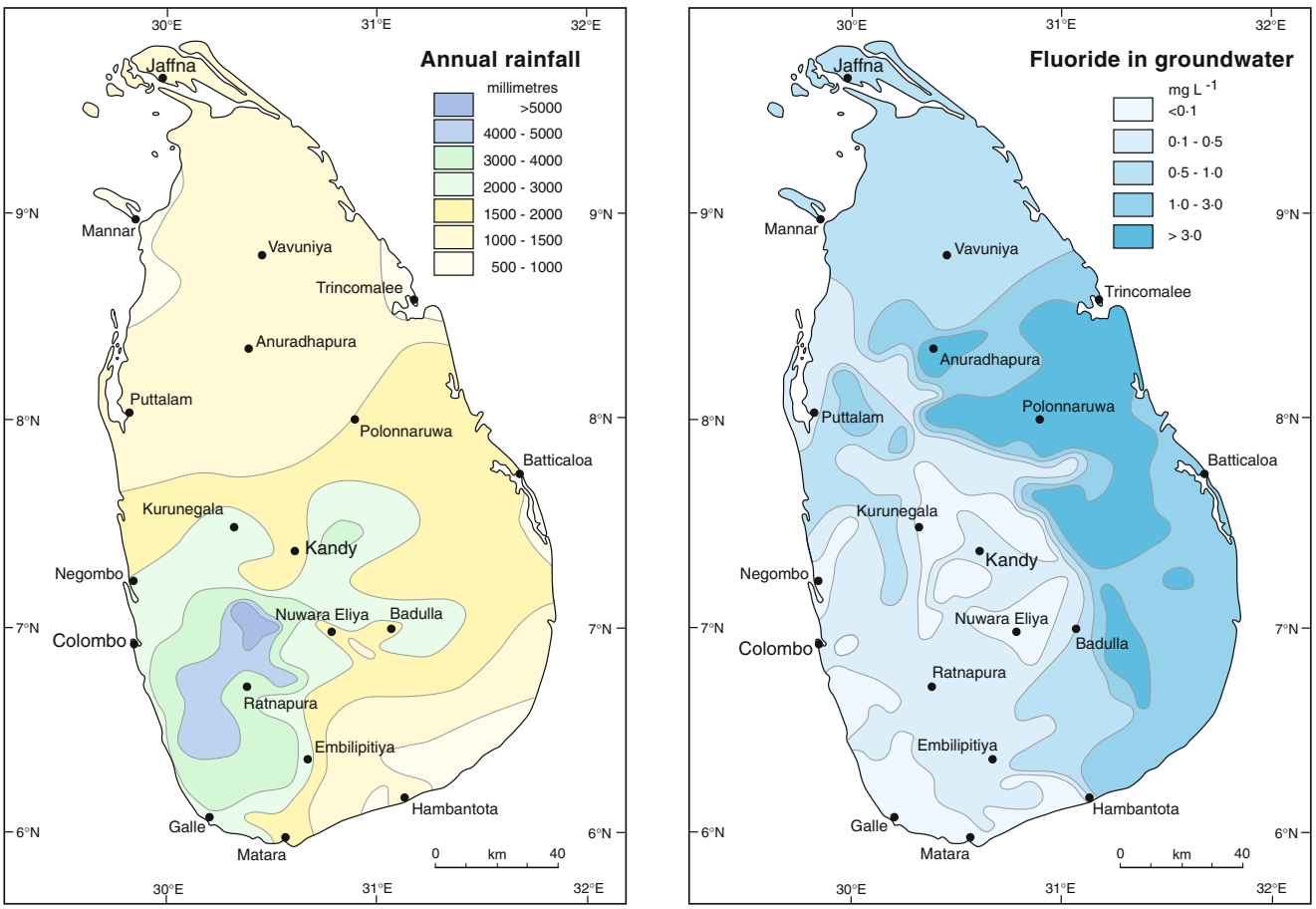


Fig. 13.9 Maps of the regional distribution of annual rainfall and fluoride concentrations in groundwater in Sri Lanka (After Dissanayake 1991; Zubair 2003)

preventative measures were taken against fluorosis. As a result of the studies (Dissanayake 1991), it was recommended that very detailed maps showing the concentrations of fluoride in groundwater should be made, followed up by rural-water-supply schemes to avoid the use of high-fluoride groundwater. The development of relatively inexpensive defluoridation techniques was also recommended, coupled with rural education programmes.

13.5.5 The East African Rift Valley – The Alkaline Volcanic Province

Some of the highest concentrations of fluoride ever recorded have been found in water from the East African Rift Valley. The Rift extends through Eritrea, Djibouti, Ethiopia, Kenya, Tanzania, Uganda, Rwanda, Burundi and Malawi and excessive fluoride concentrations have been found in groundwaters, hot springs, alkaline lakes and some river systems in all these countries (Table 13.3). The region has well-documented cases of severe dental and skeletal fluorosis (Ayenew 2008; Bugaisa 1971; Gaciri and Davies 1993; Nanyaro et al. 1984).

The anomalous fluoride concentrations have been linked to a number of sources and processes. However, they are in large part related to the development of hyper-alkaline volcanic rocks in the rift zone, including nepheline and carbonatite magmas and associated ash deposits. These are capable of accumulating large concentrations of fluoride in melts and volatile fractions. Hence, water bodies in the rift zone can accumulate fluoride directly as a result of weathering of these rocks, as well as from high-fluoride geothermal solutions. The fine-grained and friable ashes are likely to be particularly reactive. Weathering of the silicate minerals in the lavas and ashes by silicate hydrolysis reactions produces Na-HCO_3-rich groundwaters (Ayenew 2008; Jones et al. 1977), which are relatively depleted in Ca and Mg. Hence, high concentrations of fluoride can occur as the solubility of fluorite (CaF_2) is not a limiting factor. Ultimately, villiaumite (NaF) should limit the dissolved concentrations, although as this mineral is very soluble, fluoride can reach very high concentrations before this limit is achieved.

In the area of Mount Meru, a Neogene volcano in northern Tanzania, Nanyaro et al. (1984) reported fluoride concentrations of 12–76 mg L^{-1} in rivers draining the volcano's slopes (Table 13.3) and 15–63 mg L^{-1} in associated springs. They attributed the high concentrations to weathering of fluorine-rich alkaline igneous rocks and to contributions from fumaroles and gases as well as to the redissolution of fluorine-rich trona (Na_2CO_3. $NaHCO_3.2H_2O$), which occurs as a seasonal encrustation in low-lying river valleys and lake margins as a result of extreme evaporation. The river water chemistry varies significantly seasonally as a result of dilution following periods of heavy rainfall.

In perhaps the most extensive survey of water sources in the East African Rift, Tekle-Haimanot et al. (2006) outlined the distributions of fluoride in surface water and groundwater sampled nationally in Ethiopia. From 1,438 samples, they found 24.2% exceeded the WHO guideline value with 50% of deep wells and 90% of hot spring samples exceeding the WHO guideline value. Areas along the Awash River were the most affected. Here, the sugar plantations attracted large settler communities and long-term use of deep borehole water has resulted in development of high rates of both dental and skeletal fluorosis (Tekle-Haimanot et al. 2006).

In a study of the Wonji/Shoa and Metahara areas of the Awash Valley in Ethiopia, Ashley and Burley (1994) found concentrations up to 26 mg L^{-1} in groundwater from volcanic rocks of the Pliocene to Recent Aden Volcanic Series and up to 4.3 mg L^{-1} in recent alluvium overlying the volcanic rocks. They concluded that the fluoride originated from the volcanic rocks and that high concentrations in solution were maintained by the dominance of Na ions over Ca in the groundwaters. Concentrations diminished slightly in the overlying alluvium as a result of increased groundwater Ca concentrations following likely reaction with carbonate minerals in the sediments.

Alkaline and crater lakes in the East African Rift have some of the highest recorded dissolved fluoride concentrations. In their survey, Tekle-Haimanot et al. (2006) observed highest fluoride values in the Rift Valley lakes Shala (264 mg L^{-1}) and Abijata (202 mg L^{-1}). In Lake Magadi in Kenya, Jones et al. (1977) reported concentrations up to 1,980 mg L^{-1} in surface brines (Table 13.3). The lake waters have evolved towards saline compositions, in some cases with dissolved solids in excess of 300 g L^{-1}, principally by surface evaporative concentration in the low-lying lake basin, which has acted as an evaporating pan for the lake waters. Elevated concentrations of fluoride are also achieved by very low Ca concentrations, following precipitation of carbonate minerals. The lake waters appear to be largely saturated with respect to fluorite and the mineral is an abundant accessory authigenic phase in many of the Magadi lake sediments (Jones et al. 1977). Jones et al. (1977) reported extremely high fluoride concentrations (up to 2,170 mg L^{-1}) in saline groundwaters from boreholes in Magadi lake sediments. The occurrence of fluoride in the natural waters of Kenya is also further explored by Gaciri and Davies (1993) and Wambu and Muthakia (2011).

Nanyaro et al. (1984) found concentrations of 60–690 mg L^{-1} in alkaline lakes and ponds of the Momella Lakes Group (Mount Meru area) in Tanzania (Table 13.3). The lake waters are also brackish and the most alkaline

and saline compositions are found in the largest lakes of the group. Evaporation is also considered a major factor in the development of the extreme fluoride enrichments and the higher concentrations in the larger lakes of the group may be related to a smaller influence of dilution from runoff during wet periods and the long history of evolution of the lakes (Nanyaro et al. 1984). Also in Tanzania, concentrations up to 72 mg L^{-1} have been reported for Lake Natron (Table 13.3). Ashley and Burley (1994) also found 36 mg L^{-1} fluoride in Lake Besaka in the Awash Valley of Ethiopia. Variable concentrations were obtained for lake waters from East Africa by Kilham and Hecky (1973). The highest concentrations (up to 437 mg L^{-1} reported for Lake Tulusia in the Momella Group and up to 437 mg L^{-1} in Lake Magadi) were found in lake waters of Na-HCO_3 (low Cl) composition. Kilham and Hecky (1973) postulated that the fluoride most likely originates from weathering of volcanic rocks and/or from geothermal solutions in the rift system. As with other lakes in the Rift, the maintenance of high dissolved concentrations is possible because they have low Ca concentrations resulting from loss of Ca by precipitation of travertine in local streams and springs.

It is an unfortunate fact that this region of anomalously high fluoride concentrations, like many arid areas, also suffers from significant problems with water scarcity. The arid climate, combined with often poor coverage by, or poor maintenance of, abstraction boreholes in many of the affected countries means that the emphasis is usually on water availability and that drinking-water quality is of lower priority.

13.5.6 Gaspé Peninsula, Canada – Depth Control Where Glacial Deposits Overlie Bedrock

Increasing fluoride concentrations are found with depth in groundwaters in many parts of the world. Such areas may represent increasing reaction times and decreasing circulation rates, or alternatively a change in geology. One such area is the Gaspé Peninsula in Canada, where Quaternary glacial till overlies older bedrock. This is likely to be a widespread geological occurrence in North America, Europe and northern Asia and this example serves as a type area for other glacial terrain.

A farming community relying on shallow wells for water supply occupies the town of Maria in the Gaspé Peninsula of Quebec, Canada. The community has only recently suffered from cases of chronic skeletal fluorosis, resulting from fluoride concentrations in the water supplies up to 28 mg L^{-1} (Boyle and Chagnon 1995). The area is situated on a flat coastal plain covered by colluvial, alluvial and glacial sediments, varying in thickness from 10 to 30 m. The coastal plain is underlain by Carboniferous sediments comprising conglomerates, sandstones, mudstones and commonly containing calcareous cements. High-fluoride waters occur throughout the region and there is a clear correlation between fluoride concentrations and depth. Thus, almost all the wells completed in the Carboniferous sediments contain excessive fluoride; the mean concentration in water from the overlying sediments was around 0.1 mg L^{-1}, compared to 10.9 mg L^{-1} in the bedrock. The high fluoride concentrations are explained by the high cation-exchange capacities of the sediments, which give rise to softening of the groundwaters and exchange of Ca^{2+} for $2Na^{+}$, in a similar way to the Lincolnshire Limestone described above. Tritium analyses of groundwater also suggest that the groundwaters in this low-lying area have long residence times.

Farmers relied originally on shallow dug wells, but were advised in the 1970s to deepen these to increase yields (to satisfy the same demands as the townspeople who were supplied by mains water). Two cases of skeletal fluorosis drew attention to the potentially widespread nature of the problem, with the affected individuals only having been exposed to the fluoride-rich water for a period of 6 years. This study draws attention to the need for water-quality as well as water-quantity assessments during groundwater investigations, as well as an awareness of the geochemical properties of the sediments and a consideration of groundwater residence times.

13.6 Remediation of High-Fluoride Groundwater

As a means of combating the fluoride problem, several methods of water treatment using various media have been tested and a number are in use in various parts of the world. Some of the common methods are listed in Table 13.5 and reviews have been published by Heidweiller (1990) and Mohapatra et al. (2009). Most low-technology methods rely on precipitation or flocculation, or adsorption/ion-exchange processes. Probably the best-known and most established method is the Nalgonda technique (Nawlakhe and Bulusu 1989), named after the Nalgonda District in Andhra Pradesh, India where the technique has been developed and is in use today. The technique uses a combination of alum (or aluminium chloride) and lime (or sodium aluminate), together with bleaching powder, which are added to high-fluoride water, stirred and left to settle. Fluoride is subsequently removed by flocculation, sedimentation and filtration. The method can be used at a domestic scale (in buckets) or community scale (fill-and-draw type defluoridation plants; Nawlakhe and Bulusu 1989). It has moderate costs and uses materials which are usually easily available. It is therefore preferred for small defluoridation

Table 13.5 Popular removal methods for fluoride from drinking water

Removal method	Capacity/ dose	Working pH	Interferences	Advantages	Disadvantages	Relative cost
Precipitation						
Alum (aluminium sulphate)	150 mg/ mg F	Non-specific	–	Established process	Sludge produced, treated water is acidic, residual Al present	Med-high
Lime	30 mg/mg F	Non-specific	–	Established process	Sludge produced, treated water is alkaline	Med-high
Alum + lime ('Nalgonda')	150 mg alum + 7 mg lime/mg F	Non-specific, optimum 6.5	–	Low-tech, established process	Sludge produced, high chemical dose, residual Al present	Med-high
Gypsum + fluorite	5 mg gypsum + <2 mg fluorite/mg F	Non-specific	–	Simple	Requires trained operators Low efficiency, high residual Ca, SO_4	Low-med
Calcium chloride	3 mg $CaCl_2$/mg F	6.5–8.0	–	Simple	Requires additional flocculent (e.g. $FeCl_3$)	Med-high
Adsorption/ion exchange						
Activated carbon	Variable	<3	Many	–	Large pH changes before and after treatment	High
Plant carbon	300 mg F/kg	7	–	Locally available	Requires soaking in potassium hydroxide	Low-med
Zeolites	100 mg F/kg	Non-specific	–		Poor capacity	High
Defluoron 2	360 g F/m^3	Non-specific	Alkalinity		Disposal of chemicals used in resin regeneration	Medium
Clay pots	80 mg F/kg	Non-specific	–	Locally available	Low capacity, slow	Low
Activated alumina	1,200 g F/m^3	5.5	Alkalinity	Effective, well-established	Needs trained operators, chemicals not always available	Medium
Bone	900 g F/m^3	>7	Arsenic	Locally available	May give taste; degenerates not universally accepted	Low
Bone char	1,000 g F/m^3	>7	Arsenic	Locally available high capacity	Not universally accepted; may give adverse colour, taste	Low
Other						
Electrodialysis	High	Non-specific	Turbidity	Can remove other ions. Used for high salinity	Skilled operators high cost. Not much used	Very high
Reverse osmosis/ nanofiltration	High	Non-specific	Turbidity	Can remove other ions. Used for high salinity	Skilled operators high cost	Very high

After Heidweiller (1990), Richards et al. (2010), Solsona (1985), Vanderdonck and Van Kesteren (1993)

units used exclusively for drinking water. Aluminium polychloride sulphate has been found to have technical advantages over alum as use of alum results in increased concentrations of SO_4 and suspended particles in the treated water (N'dao et al. 1992).

Other precipitation methods include the use of gypsum, dolomite or calcium chloride. Most methods tested are capable in principle of reducing fluoride in treated water to a concentration below the WHO guideline value. Vanderdonck and Van Kesteren (1993) reported that the calcium chloride method is capable of reducing fluoride concentrations up to 20 mg L^{-1} to acceptable concentrations.

The most common ion-exchange removal methods tested are activated carbon, activated alumina (Barbier and Mazounie 1984), ion-exchange resins (e.g. Defluoron 2), plant carbon, clay minerals (Chaturvedi et al. 1988), clay pots, crushed bone or bone char (Rao et al. 2009). Activated alumina and bone materials are among the most effective appropriate-technology removal methods (with highest removal capacity, Table 13.5). These also have drawbacks however: activated alumina may not always be available or affordable and bone products are not readily acceptable in some cultures (Tekle-Haimanot et al. 2006). Use of fired clay-pot shards has proved a promising approach in some

developing countries (Moges et al. 1996). The use of fly ash has also been tested favourably for fluoride removal (Chaturvedi et al. 1990), although the concentrations of other solutes in the water after treatment with fly ash need to be ascertained. Some success has also been demonstrated experimentally using aluminium-rich volcanic soils (Ando soils) from the East African Rift of Kenya (Zevenbergen et al. 1996), although field testing of the technique has not been reported.

Other highly efficient methods of removal include electro-dialysis, reverse osmosis and nanofiltration (Banasiak and Schaefer 2009; Richards et al. 2010; Schneiter and Middlebrooks 1983). These methods tend to involve higher technology and higher costs (Table 13.5) and are therefore less suitable for many applications in developing countries.

Most methods designed for village-scale fluoride removal have drawbacks in terms of removal efficiency, cost, local availability of materials, chemistry of resultant treated water and disposal of treatment chemicals. Local circumstances will dictate which methods, if any, are the most appropriate. In addition, many of the defluoridation methods have only been tested at pilot scale or in the laboratory. In practice, remediation techniques such as those outlined above often meet with disappointing success when put into operation. Success rates depend on efficacy, user acceptance, ease of maintenance, and degree of community participation, availability and cost of raw materials. An additional problem lies with monitoring. In most rural communities, it will be almost impossible to monitor the initial fluoride concentrations so that dosing can be accurate. It is also not practicable at village level to chemically monitor the progress of treatment schemes, such as knowing when to recycle or replace the media involved in the treatment.

As an example of problems with the water treatment approach, defluoridation schemes have been in operation in Wonji, Ethiopia since 1962. In the past, these principally involved bone char, but most recently a resin-activated-alumina adsorption method has been used (Kloos and Tekle-Haimanot 1999). Only two of twelve plants were still in operation in 1997 and in the years of operation, only four of the plants consistently reduced fluoride below 1.5 mg L^{-1}. The major problems encountered included efficiencies in the operation of the plant, lack of spare parts and materials and lack of community involvement. The affected areas are currently served by a piped water scheme, but in more remote areas, interest in defluoridation continues.

Given the potential drawbacks of water treatment, alternative fluoride mitigation approaches could prove more effective, such as careful borehole siting (Ayenew 2008), groundwater management and use of alternative sources such as rainwater harvesting. It should be borne in mind that the main issue is for the provision of rather small amounts of water for drinking purposes so rainwater harvesting from roof collection or augmenting aquifer recharge provides a straightforward and low-cost solution.

Factors worth considering in borehole siting are local geology and variations in groundwater fluoride concentration with depth. Groundwater management includes consideration of optimum pumping rates, especially where there exists the possibility of mixing of groundwater with deep fluoride-rich groundwater (e.g. old groundwater or hydrothermal solutions), which would be increasingly drawn upwards at high pumping rates (Carrillo-Rivera et al. 2002). Possibilities for enhanced recharge of low-fluoride surface water to aquifers could also give benefits to shallow groundwater quality. Examples include the charco dams of Tanzania. A fully-fledged artificial recharge scheme would offer similar benefits. Considerations with rainwater harvesting include availability of rainfall in arid and semi-arid areas that tend to be worst-affected by fluoride problems, and in maintenance of clean storage facilities for the water supply.

In view of the site-specific nature of each incidence of high fluoride and also the problems of applying remediation, especially in small communities, the supply of small quantities of imported low-fluoride water should also be considered. This could be achieved by importation of water in bottles or tankers to central distribution points and may prove cost-effective, since the high-fluoride source can still be used for sanitation and possibly also for most forms of agriculture.

In low-rainfall areas with fluoride problems, it is important to assess the hydrogeological situation very carefully. The generation of high-fluoride groundwaters usually requires considerable residence times in the aquifer, for reasons outlined in the preceding sections. Thus, it is likely that younger, shallow groundwaters, for example those recharged rapidly below wadis or stream channels, may have lower fluoride concentrations than the bulk groundwater. They may be exploitable as a resource overlying older groundwater. Exploitation may require 'skimming' the shallow water table rather than abstraction from deeper penetrating boreholes. The harvesting of rainwater, either directly in cisterns or by collection in small recharge dams, offers a potentially attractive alternative solution.

Conclusions

This review has summarised the occurrence of fluorine in natural water and the environment and outlined its behaviour in terms of the hydrogeochemical cycle. The principal sources of fluoride in rainfall are marine aerosols, volcanic emissions and anthropogenic inputs. Concentrations in rain typically lie in the range 0.02–0.2 mg L^{-1} with pristine and continental rains lying at the lower end of the range. Surface waters

generally also reflect these sources and concentrations in most are low, except where affected by extensive evapotranspiration or geothermal inputs. Surface waters rarely have fluoride concentrations sufficiently high to be detrimental to health and if fluoride-related health problems exist, they are more likely to be linked to deficiency.

Groundwaters are generally more vulnerable to the build up of fluoride and although concentrations in these are also mostly low (<1 mg L^{-1}), a number of conditions can give rise to detrimentally high concentrations. Geology plays a key role in defining fluoride concentrations. Areas with potentially high-fluoride groundwater include crystalline basement rocks, especially those of granitic composition. These contain relatively high concentrations of fluorine-bearing minerals (e.g. apatite, mica, hornblende, occasionally fluorite) and have low calcium concentrations. Such areas occur over large parts of India, Sri Lanka and Africa for example. The active volcanic province of the East African Rift Valley is also an anomalous area with high fluoride concentrations on a regional scale. High concentrations result from both weathering of volcanic rocks and geothermal inputs. In groundwaters from the East African Rift, fluoride concentrations in excess of 1 mg L^{-1} are common, while concentrations up to several hundred mg L^{-1} have been recorded in some hot springs, and concentrations in excess of 1,000 mg L^{-1} have been found in some alkaline lakes. Some sedimentary aquifers also contain high-fluoride groundwaters, especially where calcium concentrations are low and Na-HCO_3 waters dominate. These can occur under arid conditions and in aquifers affected by ion exchange.

The influence of low recharge on dissolved fluoride concentrations in groundwater is also clear from documented studies. The tendency for fluoride accumulation in groundwater from arid and semi-arid areas is well-illustrated from crystalline basement aquifers of India and Sri Lanka and many sedimentary aquifers. The accumulations under arid conditions are the result of rock weathering, evapotranspiration and low recharge. Excessive groundwater fluoride concentrations are relatively rare in tropical and even temperate regions.

Groundwaters in a number of aquifers display progressive increases in fluoride along flow lines and demonstrate the importance of residence time and their position in the flow and reaction sequence. The phenomenon has been demonstrated in the English Lincolnshire Limestone aquifer where dissolved fluoride concentrations increase progressively from 0.11 mg L^{-1} in the unconfined aquifer to 5.6 mg L^{-1} in the deep confined aquifer in response to time-dependent mineral dissolution and ion-exchange reactions. In many aquifers, the evolution of groundwater down the flow gradient has taken place over centuries or millennia, with water being abstracted having had significant opportunity for equilibration with host aquifer minerals. The accumulation of fluoride in water is ultimately limited by mineral solubility. In groundwaters where calcium is abundant, fluoride concentrations are limited by saturation with the mineral fluorite. In cases where calcium concentrations are low, or where calcium is removed by ion exchange, fluoride may build up to excessive and dangerous concentrations.

Given these key controls on fluoride occurrence and distribution, it is possible to anticipate broadly where areas of regionally high fluoride concentrations are likely to exist. Such an understanding of the fluoride occurrence is important for the management of the fluoride-related epidemiological problems. Water-supply programmes in potentially high-fluoride areas should have geological and hydrological guidance, including chemical analyses and geological maps. High fluoride incidence may not be universal within a given area and remediation strategies may include identification of areas of distinct geology, residence time or selective recharge where fluoride concentrations are locally lower. Nevertheless, even taking into account all available hydrogeological information, groundwater fluoride concentrations are frequently so variable on a local scale that prediction of concentrations in individual wells is difficult. Hence, for the purposes of compliance, testing of each well used for drinking water should be undertaken in fluoride-prone areas. Testing can be achieved effectively either in the field or laboratory by ion-selective electrodes, colorimetry or ion chromatography. Chemical and other forms of intervention and treatment should generally be a last resort, especially since their monitoring in rural communities is virtually impossible.

See Also the Following Chapters. Chapter 2 (Natural Distribution and Abundance of Elements) • Chapter 10 (Volcanic Emissions and Health) • Chapter 28 (GIS in Human Health Studies) • Chapter 30 (Mineralogy of Bone) • Chapter 33 (Modeling Groundwater Flow and Quality)

References

Ainsworth NJ (1933) Mottled teeth. Br Dent J 55:233–250

Apambire WB, Boyle DR, Michel FA (1997) Geochemistry, genesis, and health implications of fluoriferous groundwaters in the upper regions of Ghana. Environ Geol 33:13–24

Ashley PP, Burley MJ (1994) Controls on the occurrence of fluoride in groundwater in the Rift Valley of Ethiopia. In: Nash H, McCall GJH (eds) Groundwater quality. Chapman & Hall, London, pp 45–54

Ayenew T (2008) The distribution and hydrogeological controls of fluoride in the groundwater of central Ethiopian rift and adjacent highlands. Environ Geol 54:1313–1324

Banasiak LJ, Schaefer AI (2009) Removal of boron, fluoride and nitrate by electrodialysis in the presence of organic matter. J Membr Sci 334:101–109

Banks D, Frengstad B, Midtgard AK, Krog JR, Strand T (1998) The chemistry of Norwegian groundwaters: I. The distribution of radon, major and minor elements in 1604 crystalline bedrock groundwaters. Sci Total Environ 222:71–91

Barbier JP, Mazounie P (1984) Methods of reducing high fluoride content in drinking water: fluoride removal methods–filtration through activated alumina: a recommended technique. Water Supply 2:3–4

Bårdsen A, Bjorvatn K, Selvig KA (1996) Variability in fluoride content of subsurface water reservoirs. Acta Odontol Scand 54:343–347

Barnard WR, Nordstrom DK (1982) Fluoride in precipitation–II. Implication for the geochemical cycling of fluorine. Atmos Environ 16:105–111

Botchway CA, Ansa-Asare OD, Antwi LAK (1996) Natural fluoride and trace metal levels of ground and surface waters in the greater Accra region of Ghana. West Afr J Med 15:204–209

Boyle DR, Chagnon M (1995) An incidence of skeletal fluorosis associated with groundwaters of the maritime Carboniferous basin, Gaspé region, Quebec, Canada. Environ Geochem Health 17:5–12

Bugaisa SL (1971) Significance of fluorine in Tanzania drinking water. In: Gerhard T (ed) Proceedings of the Conference on Rural Water Supply in East Africa. Bureau of resource assessment and land use planning. University of Dar Es Salaam, Tanzania, 20:107–113

Carrillo-Rivera JJ, Cardona A, Edmunds WM (2002) Use of abstraction regime and knowledge of hydrogeological conditions to control high-fluoride concentration in abstracted groundwater: San Luis Potosí basin, Mexico. J Hydrol 261:24–47

Chakraborti D, Chanda CR, Samanta G, Chowdhury UK, Mukherjee SC, Pal AB, Sharma B, Mahanta KJ, Ahmed HA, Sing B (2000) Fluorosis in Assam, India. Curr Sci 78:1421–1423

Chaturvedi AK, Pathak KC, Singh VN (1988) Fluoride removal from water by adsorption on China clay. Appl Clay Sci 3:337–346

Chaturvedi AK, Yadava KP, Pathak KC, Singh VN (1990) Defluoridation of water by adsorption on fly ash. Water Air Soil Pollut 49:51–61

Connett P (2007) Professionals mobilize to end water fluoridation worldwide. Fluoride 40:155–158

Cook JM, Edmunds WM, Robins NS (1991) Groundwater contributions to an acid upland lake (Loch Fleet, Scotland) and the possibilities for amelioration. J Hydrol 125:111–128

Corbett RG, Manner BM (1984) Fluoride in ground water of northeastern Ohio. Ground Water 22:13–17

Datta DK, Gupta LP, Subramanian V (2000) Dissolved fluoride in the Lower Ganges-Brahmaputra-Meghna River system in the Bengal Basin, Bangladesh. Environ Geol 39:1163–1168

Dean HT, Elvove E (1937) Further studies on the minimal threshold of chronic endemic dental fluorosis. Public Health Rep (Wash) 52:1249–1264

Deng Y, Nordstrom DK, McCleskey RB (2011) Fluoride geochemistry of thermal waters in Yellowstone National Park: I. Aqueous fluoride speciation. Geochimica Et Cosmochimica Acta 75:4476–4489

Diaz-Barriga F, Leyva R, Quistian J, Loyola-Rodriguez JB, Pozos A, Grimaldo M (1997) Endemic fluorosis in San Luis Potosi, Mexico. Fluoride 30:219–222

Diesendorf M (1986) The mystery of declining tooth decay. Nature 322:125–129

Dissanayake CB (1991) The fluoride problem in the groundwater of Sri Lanka–environmental management and health. Int J Environ Stud 38:137–156

Edmunds WM (1973) Trace element variations across an oxidation-reduction barrier in a limestone aquifer. In: Proceedings of a symposium hydrogeochemistry, Clarke Co., Washington, Tokyo (1970), pp 500–526

Edmunds WM (1994) Characterization of groundwaters in semi-arid and arid zones using minor elements. In: Nash H, McCall GJH (eds) Groundwater quality. Chapman & Hall, London, pp 19–30

Edmunds WM, Walton N (1983) The Lincolnshire Limestone–hydrogeochemical evolution over a ten-year period. J Hydrol 61:201–211

Edmunds WM, Cook JM, Darling WG, Kinniburgh DG, Miles DG, Bath AH, Morgan-Jones M, Andrews JN (1987) Baseline geochemical conditions in the Chalk aquifer, Berkshire, UK: a basis for groundwater quality management. Appl Geochem 2:251–274

Edmunds WM, Cook JM, Kinniburgh DG, Miles DG, Trafford JM (1989) Trace element occurrence in British groundwaters, British geological survey research report SD/89/3. British Geological Survey, Keyworth, p 424

Ellis AJ (1973) Chemical processes in hydrothermal systems–a review. In: Ingerson E (ed) Proceedings of a symposium on hydrogeochemistry. Clarke Co., Washington DC, pp 1–26

Ellis AJ, Mahon WAJ (1977) Chemistry and geothermal systems. Academic, New York

Eyre R, Feltbower RG, Mubwandarikwa E, Eden TOB, McNally RJQ (2009) Epidemiology of bone tumours in children and young adults. Pediatr Blood Cancer 53:941–952

Fantong WY, Satake H, Ayonghe SN, Suh EC, Adelana SMA, Fantong EBS, Banseka HS, Gwanfogbe CD, Woincham LN, Uehara Y, Zhang J (2010) Geochemical provenance and spatial distribution of fluoride in groundwater of Mayo Tsanaga River Basin, Far North Region, Cameroon: implications for incidence of fluorosis and optimal consumption dose. Environ Geochem Health 32:147–163

Finkelman RB, Belkin HE, Zheng B (1999) Health impacts of domestic coal use in China. Proc Natl Acad Sci USA 96:3427–3431

Fritz B (1981) Etude thermodynamique et modélisation des réactions hydrothermales et diagénétiques. Sciences Géologiques. Strasbourg, Strasbourg, p 197

Fuge R, Andrews MJ (1988) Fluorine in the UK environment. Environ Geochem Health 10:96–104

Fuhong R, Shuquin J (1988) Distribution and formation of high-fluorine groundwater in China. Environ Geol Water Sci 12:3–10

Gaciri SJ, Davies TC (1993) The occurrence and geochemistry of fluoride in some natural waters of Kenya. J Hydrol 143:395–412

Goni J, Greffard J, Leleu M, Monitron L (1973) Le fluor dans les eaux de boisson. Rapport 73 SGN 429 AME, Bureau de Recherches Géologiques et Miniéres (BRGM), Orléans, France

Guo Q, Wang Y, Liu W (2009) Hydrogeochemistry and environmental impact of geothermal waters from Yangyi of Tibet, China. J Volcanol Geotherm Res 180:9–20

Gupta SC, Rathore GS, Doshi CS (1993) Fluoride distribution in groundwaters of southeastern Rajasthan. Indian J Environ Health 35:97–109

Gupta MK, Singh V, Rajwanshi P, Agarwal M, Rai K, Srivastava S, Srivastav R, Dass S (1999) Groundwater quality assessment of Tehsil Kheragarh, Agra (India) with special reference to fluoride. Environ Monit Assess 59:275–285

Hamilton M (1992) Water fluoridation: a risk assessment perspective. J Environ Health 54:27–32

Handa BK (1975) Geochemistry and genesis of fluoride-containing ground waters in India. Ground Water 13:275–281

Heidweiller VML (1990) Fluoride removal methods. In: Frencken JE (ed) Proceedings of a symposium on endemic fluorosis in developing countries: causes, effects and possible solutions. NIPG-TNO, Leiden, pp 51–85

Hem JD (1985) Study and interpretation of the chemical characteristics of natural water, USGS water supply paper 2254. U.S. Geological Survey, Alexandria, p 263

Hitchon B (1995) Fluorine in formation waters in Canada. Appl Geochem 10:357–367

Hoather RC (1953) Fluorides in water supplies. J R Sanit Inst 73:202–223
Jacob B (1975) Contribution à l'étude géochimique des eaux froides et thermominérales du Massif du Mont-Dore. Université de Paris VI, Paris, p 205
Jones BF, Eugster HP, Reitig SL (1977) Hydrochemistry of the Lake Magadi basin, Kenya. Geochim Cosmochim Acta 41:53–72
Kabata-Pendias A, Pendias H (1984) Trace elements in soils and plants. CRC Press, Boca Raton
Kafri U, Arad A, Halicz L, Ganor E (1989) Fluorine enrichment in groundwater recharged through loess and dust deposits, southern Israel. J Hydrol 110:373–376
Kilham P, Hecky RE (1973) Fluoride: geochemical and ecological significance in East African waters and sediments. Limnol Oceanogr 18:932–945
Kloos H, Tekle-Haimanot R (1999) Distribution of fluoride and fluorosis in Ethiopia and prospects for control. Trop Med Int Health 4:355–364
Kotoky P, Barooah PK, Baruah MK, Goswami A, Borah GC, Gogoi HM, Ahmed F, Gogoi A, Paul AB (2008) Fluoride and endemic fluorosis in the Karbianglong district, Assam, India. Fluoride 41:72–75
Lamont P (1958) Interim report on an investigation of the chemical characteristics of the underground waters from the Lincolnshire Limestone in the area of supply, with special reference to the possibility of locating new sources yielding a more suitable water for public supplies. Spalding, Lincolnshire
Lavado RS, Reinaudi N (1979) Fluoride in salt affected soils of La Pampa (Republica Argentina). Fluoride 12:28–32
Mahon WAJ (1964) Fluorine in the natural thermal waters of New Zealand. N Z J Sci 7:3–28
Maier FJ (1950) Fluoridation of public health supplies. J Am Waterwork Assoc 42:1120–1132
Maithani PB, Gurjar R, Banerjee R, Balaji BK, Ramachandran S, Singh R (1998) Anomalous fluoride in groundwater from western part of Sirohi district, Rajasthan, and its crippling effects on human health. Curr Sci 74:773–777
Majumdar KK (2011) Health impact of supplying safe drinking water containing fluoride below permissible level on flourosis patients in a fluoride-endemic rural area of West Bengal. Indian J Public Health 55:303–308
Marshall E (1990) The fluoride debate: one more time. Science 247:276–277
Martinez DE, Quiroz Londono OM, Massone HE, Palacio Buitrago P, Lima L (2012) Hydrogeochemistry of fluoride in the Quequen river basin: natural pollutants distribution in the argentine pampa. Environ Earth Sci 65:411–420
McCaffrey LP (1998) Distribution and causes of high fluoride groundwater in the western Bushveld area of South Africa, Unpublished PhD thesis, Cape Town, South Africa
Moges G, Zewge F, Socher M (1996) Preliminary investigations on the defluoridation of water using fired clay chips. J Afr Earth Sci 22:479–482
Mohapatra M, Anand S, Mishra BK, Giles DE, Singh P (2009) Review of fluoride removal from drinking water. J Environ Manage 91:67–77
Murray RJ (1960) The geology of the "Zuarungu" E field sheet, Geological survey Ghana bulletin no. 25. Geological Survey Ghana, Accra, p 118
N'dao I, Lagaude A, Travi Y (1992) Défluoruration expérimentale des eaux souterraines du Sénégal par le sulphate d'aluminium et le polychlorosulphate basique d'aluminium. Sciences et Techniques de l'Eau 26:243–249
Nanyaro JT, Aswathanarayana U, Mungure JS (1984) A geochemical model for the abnormal fluoride concentrations in waters in parts of northern Tanzania. J Afr Earth Sci 2:129–140
Nawlakhe WG, Bulusu KR (1989) Nalgonda technique–a process for removal of excess fluoride from water. Water Qual Bull 14:218–220
Neal C (1989) Fluorine variations in Welsh streams and soil waters. Sci Total Environ 80:213–223
Neal C, Wilkinson J, Neal M, Harrow M, Wickham H, Hill L, Morfitt C (1997) The hydrochemistry of the River Severn, Plynlimon. Hydrol Earth Syst Sci 1:583–617
Nordstrom DK, Jenne EA (1977) Fluorite solubility equilibria in selected geothermal waters. Geochim Cosmochim Acta 41:175–198
Parkhurst DL, Appelo CAJ (1999) User's guide to PHREEQC (Version 2)–A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations, Water-resources investigations report 99–4259. USGS, Denver
Pickering WF (1985) The mobility of soluble fluoride in soils. Environ Pollut (Ser B) 9:281–308
Pizzo G, Piscopo MR, Pizzo I, Giuliana G (2007) Community water fluoridation and caries prevention: a critical review. Clin Oral Investig 11:189–193
Queste A, Lacombe M, Hellmeier W, Hillerman F, Bortulussi B, Kaup M, Ott O, Mathys W (2001) High concentrations of fluoride and boron in drinking water wells in the Muenster region–results of a preliminary investigation. Int J Hyg Environ Health 203: 221–224
Rafique T, Naseem S, Usmani TH, Bashir E, Khan FA, Bhanger MI (2009) Geochemical factors controlling the occurrence of high fluoride groundwater in the Nagar Parkar area, Sindh, Pakistan. J Hazard Mater 171:424–430
Rao NS (2002) Geochemistry of groundwater in parts of Guntur District, Andhra Pradesh, India. Environ Geol 41:552–562
Rao ASM, Rajayalakshmi K, Sastry KRK, Rao PS, Rao VVK, Rao KN (1974) Incidence of fluorides in drinking water resources, in Andhra Pradesh. Fluorosis, Hyderabad, pp 227–235
Rao SM, Reddy BVV, Lakshmikanth S, Ambika NS (2009) Re-use of fluoride contaminated bone char sludge in concrete. J Hazard Mater 166:751–756
Rea RE (1979) A rapid method for the determination of fluoride in sewage sludges. Water Pollut Control 78:139–142
Reddy DV, Nagabhushanam P, Sukhija BS, Reddy AGS, Smedley PL (2010) Fluoride dynamics in the granitic aquifer of the Wailapally watershed, Nalgonda District, India. Chem Geol 269:278–289
Richards LA, Vuachere M, Schaefer AI (2010) Impact of pH on the removal of fluoride, nitrate and boron by nanofiltration/reverse osmosis. Desalination 261:331–337
Saether OM, Andreassen BT (1989) Fluoride in precipitation in southern Norway: amounts and sources, Report 89/106. Geological Survey of Norway, Norway
Schneiter RW, Middlebrooks EJ (1983) Arsenic and fluoride removal from groundwater by reverse osmosis. Environ Int 9:289–292
Smedley PL, Edmunds WM, West JM, Gardner SJ, Pelig-Ba KB (1995) Vulnerability of shallow groundwater quality due to natural geochemical environment. 2. Health problems related to groundwater in the Obuasi and Bolgatanga areas, Ghana. BGS technical report WC/95/43, British Geological Survey, Keyworth, Nottinghamshire
Smedley PL, Nicolli HB, Macdonald DMJ, Barros AJ, Tullio JO (2002) Hydrogeochemistry of arsenic and other inorganic constituents in groundwaters from La Pampa, Argentina. Appl Geochem 17:259–284
Smet J (1990) Fluoride in drinking water. In: Frencken JE (ed) proceedings of a symposium on Endemic Fluorosis in developing countries: causes, effects and possible solutions, NIPG-TNO, Leiden, pp 51–85
Solsona F (1985) Water defluoridation in the Rift Valley, Ethiopia. UNICEF, Addis Ababa, p 27
Suma Latha S, Anbika SR, Prasad SJ (1999) Fluoride contamination status of groundwater in Karnataka. Curr Sci 76:730–734

Tekle-Haimanot R, Melaku Z, Kloos H, Reimann C, Fantaye W, Zerihun L, Bjorvatn K (2006) The geographic distribution of fluoride in surface and groundwater in Ethiopia with an emphasis on the Rift Valley. Sci Total Environ 367:182–190

Ten Haven HL, Konings R, Schoonen MAA, Jansen JBH, Vriend SP, Van der Weijden CH, Buitenkamp J (1985) Geochemical studies in the drainage basin of the Rio Vouga (Portugal). Chem Geol 51:225–238

Travi Y (1993) Hydrogéologie et hydrochimie des aquifères du Sénégal, Sciences Géologiques, Memoire 95. Université de Paris-Sud, Paris

Turner BD, Binning P, Stipp SLS (2005) Fluoride removal by calcite: evidence for fluorite precipitation and surface adsorption. Environ Sci Technol 39:9561–9568

US Public Health Service (1943) Public health drinking water standards. J Am Waterwork Assoc 35:93–104

US Public Health Service (1946) Drinking water standards. J Am Waterwork Assoc 38:361–370

Vanderdonck P, Van Kesteren X (1993) Utilisation du chlorure de calcium. Tribune de l'Eau 562:39–40

Vijaya Kumar V, Sai CST, Rao PLKM, Rao CS (1991) Studies on the distribution of fluoride in drinking water sources in Medchal Block, Ranga Reddy District, Andra Pradesh, India. J Fluor Chem 55:229–236

Viswanathan G, Jaswanth A, Gopalakrishnan S, Ilango SS, Aditya G (2009) Determining the optimal fluoride concentration in drinking water for fluoride endemic regions in South India. Sci Total Environ 407:5298–5307

Wambu EW, Muthakia GK (2011) High fluoride water in the Gilgil area of Nakuru County, Kenya. Fluoride 44:37–41

Wenzel WW, Blum WEH (1992) Fluorine speciation and mobility in F-contaminated soils. Soil Sci 153:357–364

WHO (2011) Guidelines for drinking-water quality, 4th edn. World Health Organization, Geneva

Woo NC, Moon JW, Won JS, Hahn JS, Lin XY, Zhao YS (2000) Water quality and pollution in the Hunchun Basin, China. Environ Geochem Health 22:1–18

Wuyi W, Ribang L, Jian'an T, Kunli L, Lisheng Y, Hairong L, Yonghua L (2002) Adsorption and leaching of fluoride in soils of China. Fluoride 35:122–129

Yeung CA (2008) A systematic review of the efficacy and safety of fluoridation. Evid Based Dent 9:39–43

Zan-dao W, Yan W (2002) Fluoridation in China: a clouded future. Fluoride 35:1–4

Zevenbergen C, van Reeuwijk LP, Frapporti G, Louws RJ, Schuiling RD (1996) A simple method for defluoridation of drinking water at village level by adsorption on Ando soil in Kenya. Sci Total Environ 188:225–232

Zubair L (2003) Sensitivity of Kelani streamflow in Sri Lanka to ENSO. Hydrol Process 17:2439–2448

Water Hardness and Health Effects

14

Eva Rubenowitz-Lundin and Kevin M. Hiscock

Contents

14.1 Definition of Water Hardness

Water hardness is the traditional measure of the capacity of water to react with soap and describes the ability of water to bind soap to form lather, which is a chemical reaction detrimental to the washing process. Hardness has little significance in terms of hydrochemical studies, but it is an important parameter for water users. Today, the technical significance of water hardness is more concerned with the corrosive effects on water pipes that carry soft water.

Despite the wide usage of the term, the property of hardness is difficult to define exactly. Water hardness is not caused by a single substance but by a variety of dissolved polyvalent metallic ions—predominantly calcium and magnesium—although other ions, for example, aluminum, barium, iron, manganese, strontium, and zinc, also contribute. The source of the metallic ions are typically sedimentary rocks, and the most common are limestone ($CaCO_3$) and dolomite ($CaMg(CO_3)_2$).

Hardness is normally expressed as the total concentration of calcium and magnesium ions in water in units of mg L^{-1} as equivalent $CaCO_3$. Hardness can be determined by substituting the concentration of calcium and magnesium, expressed in mg L^{-1}, in the following equation:

$$\text{Total hardness} = 2.5\left(Ca^{2+}\right) + 4.1\left(Mg^{2+}\right) \qquad (14.1)$$

Each concentration is multiplied by the ratio of the formula weight of $CaCO_3$ to the atomic weight of the ion; hence, the factors 2.5 and 4.1 are included in the hardness relation (Freeze and Cherry 1979).

In Europe, water hardness is often expressed in terms of degrees of hardness. One French degree is equivalent to 10 mg L^{-1} as $CaCO_3$, one German degree to 17.8 mg L^{-1} as $CaCO_3$, and one English of Clark degree to 14.3 mg L^{-1} as $CaCO_3$. One German degree of hardness (dH) is equal to 1 mg of calcium oxide (CaO) or 0.72 mg of magnesium oxide (MgO) per 100 mL of water.

E. Rubenowitz-Lundin
Department of Environmental Medicine, Göteborg University, Göteborg, Sweden

K.M. Hiscock (✉)
School of Environmental Sciences, University of East Anglia, Norwich, UK
e-mail: k.hiscock@uea.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_14,

Where reported, carbonate hardness includes that part of the total hardness equivalent to the bicarbonate and carbonate (or alkalinity). If the hardness exceeds the alkalinity, the excess is called the non-carbonate hardness, and is a measure of the calcium and magnesium sulfates. In older publications, the terms "temporary" and "permanent" are used in place of carbonate and non-carbonate. Temporary hardness reflects the fact that the ions responsible may be precipitated by boiling, such that:

$$Ca^{2+} + 2HCO_3^- \rightarrow CaCO_3\downarrow + H_2O + CO_2\uparrow \qquad (14.2)$$

A number of attempts have been made to classify water hardness. Water with hardness values greater than 150 mg L^{-1} is designated as very hard. Soft water has values of less than 60 mg L^{-1}. Groundwaters in contact with limestone or gypsum ($CaSO_4.2H_2O$) rocks can commonly attain levels of 200–300 mg L^{-1}. In water from gypsiferous formations, 1,000 mg L^{-1} or more of total hardness may be present (Hem 1985).

Hardness in water used for domestic purposes does not become particularly troublesome until a level of 100 mg L^{-1} is exceeded. Depending on pH and alkalinity, hardness of about 200 mg L^{-1} can result in scale deposition, particularly on heating, and increased soap consumption. Soft waters with a hardness of less than about 100 mg L^{-1} have a low buffering capacity and may be more corrosive to water pipes resulting in the presence of heavy metals such as cadmium, copper, lead, and zinc in drinking water. This depends on the pH, alkalinity, and dissolved oxygen concentration of the water.

No health-based guideline value is proposed for hardness because it is considered that the available data on the inverse relationship between the hardness of drinking water and cerebrovascular disease (CVD) are inadequate to permit the conclusion that the association is causal (WHO 2002). However, a concentration of 500 mg L^{-1} is at the upper limit of aesthetic acceptability.

14.2 Natural Hydrochemical Evolution of Groundwater

The combination of geology and hydrology of a river catchment is important in determining the hardness of water. As illustrated in Fig. 14.1a, catchments underlain by impermeable rocks that are resistant to erosion generate surface runoff with little time for weathering to occur. As a result, the surface water has a chemical composition similar to dilute rainfall and is characterized as soft water. In contrast, and as illustrated in Fig. 14.1b, catchments underlain by permeable rocks allow water to infiltrate below the ground surface such that groundwater in contact with the rock mass promotes solutional weathering, which potentially leads to the development of fissures and conduits. As a result, the groundwater attains a high concentration of dissolved constituents and is characterized as hard water.

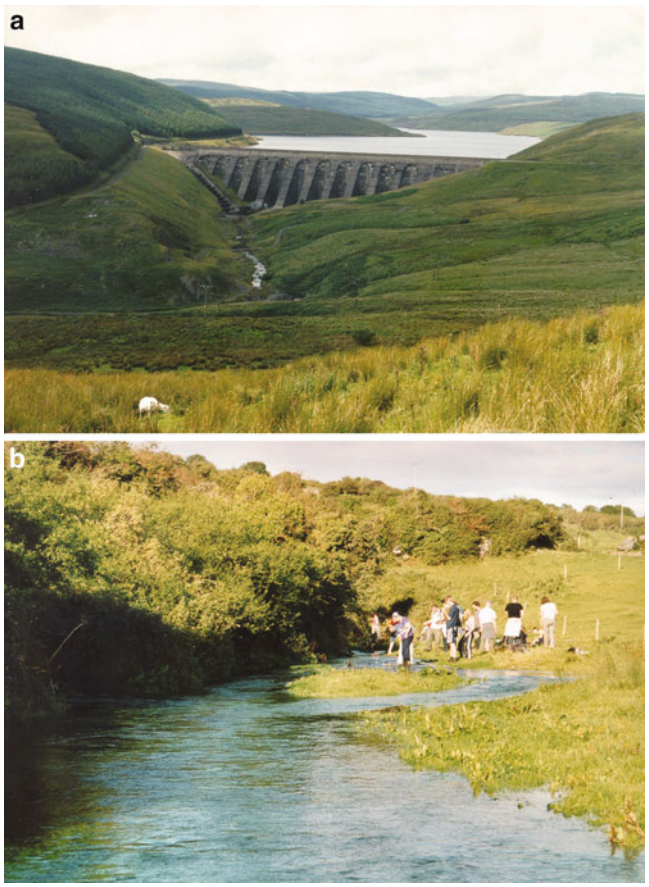


Fig. 14.1 Illustration of two types of water resources and the influence of catchment geology on water hardness. (**a**) surface water runoff from the impermeable Silurian mudstones and siltstones to the Dinas reservoir in west Wales is a soft water with a hardness of 19 mg L^{-1} as $CaCO_3$, pH of 7.0, and electrical conductivity of 73 µScm^{-1} (sample date 1998; Hiscock and Paci 2000). (**b**) groundwater discharge from the Fergus River Cave springs developed in Carboniferous limestone in County Clare, Ireland, is a very hard water with a hardness of 256 mg L^{-1} as $CaCO_3$, pH of 7.7, and electrical conductivity of 440 µScm^{-1} (sample date 2002)

In natural aqueous systems, the chemical composition of water is continually changing and is controlled predominantly by geological factors. The geochemistry of rock weathering determines the release of elements into water. In groundwaters, a natural evolution of chemical composition is recognized. Dilute rainwater with a sodium chloride water type and containing CO_2 enters the soil zone whereupon further CO_2, produced by the decay of organic matter, is dissolved in the infiltrating water to form carbonic acid. Within the soil and unsaturated zone of sedimentary rocks, this weak acid dissolves soluble calcium and magnesium carbonates such as calcite and dolomite to give high

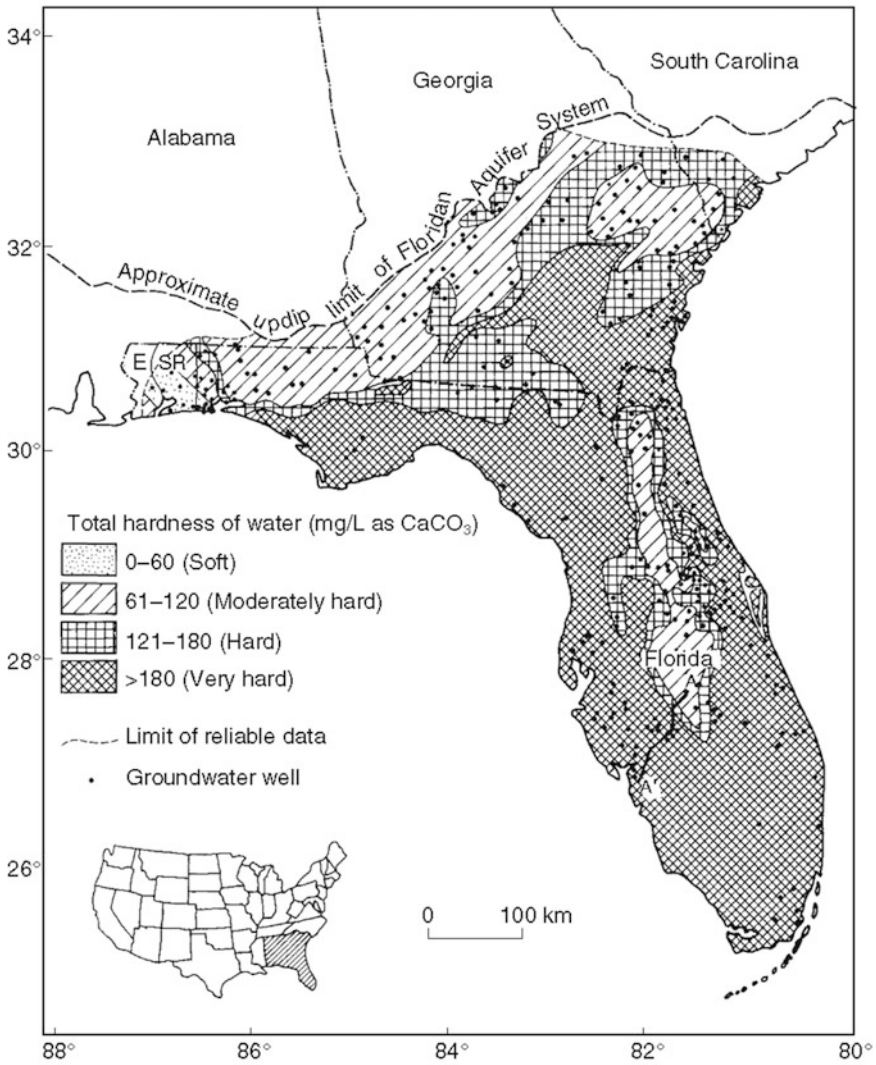


Fig. 14.2 Map of the extent of the Upper Floridan Aquifer showing the spatial distribution of total hardness of water. Escambia and Santa Rosa Counties, Florida, are located at positions E and SR, respectively (After Sprinkle 1989)

concentrations of calcium, magnesium, and bicarbonate. In crystalline igneous and metamorphic rocks, slow weathering of silicates by the attack of carbonic acid releases low concentrations of calcium, magnesium, potassium, and sodium in the infiltrating soil water or groundwater and also produces bicarbonate.

Away from the supply of oxygen in the soil and unsaturated zone, infiltrating water becomes increasingly anoxic as a result of progressive bacterial reduction of oxygen. Below the water table and with increasing reducing conditions, iron and manganese become mobilized and then later precipitated as metal sulfides. In the presence of disseminated clay material within an aquifer, ion exchange replaces calcium for sodium as the water evolves to a sodium bicarbonate water type. Hence, the groundwater is naturally softened by ion exchange reactions. In the deeper, confined section of aquifers, mixing with saline water may occur to produce a sodium chloride water type before a region of static water and aquifer diagenesis is reached. Either part or all of this classic sequence of hydrochemical change is identified in a number of aquifers including the Floridan aquifer system. The Floridan aquifer system occurs in the southeast United States (Fig. 14.2) and is one of the most productive aquifers in the world. The aquifer system is a vertically continuous sequence of Tertiary carbonate rocks of generally high permeability. Limestones and dolomites are the principal rock types, although in southwestern and northeastern Georgia and in South Carolina the limestones grade into lime-rich sands and clays. The Floridan aquifer is composed primarily of calcite and dolomite with minor gypsum, apatite, glauconite, quartz, clay minerals, and trace amounts of metallic oxides and sulfides.

The total hardness of water in the Upper Floridan aquifer varies from <50 to >5,000 mg L^{-1} as $CaCO_3$. Generally, where the system is composed only of limestone, the total hardness is equivalent to carbonate hardness and is <120 mg L^{-1}. Groundwater with higher total hardness usually results from (1) dissolution of other aquifer minerals,

primarily dolomite and gypsum; (2) mixing of fresh water with residual saline groundwater; (3) encroachment and mixing of modern seawater; or (4) contamination. Natural softening of the groundwater by cation exchange is thought to be responsible for the low hardness in Escambia and Santa Rosa Counties, Florida (Sprinkle 1989).

A sequence of hydrochemical evolution is identified by Sprinkle (1989), which starts with calcite dissolution in recharge areas that produces a calcium-bicarbonate dominated water type with a total dissolved solids (TDS) concentration of generally less than 250 mg L^{-1}. Downgradient, dissolution of dolomite leads to a calcium-magnesium-bicarbonate water type. Where gypsum is abundant, sulfate becomes the predominant anion. In coastal areas, as shown in Fig. 14.3, seawater increases the TDS concentration and the water type changes to sodium-chloride. In the western panhandle of Florida, cation exchange leads to the development of a sodium-bicarbonate water type and a less hard groundwater.

For additional reading on groundwater modeling, see Chap. 33, this volume.

14.3 The Hard-Water Story

The history behind what is considered today to be a commonly accepted fact—that hard water protects against CVD—is often referred to as the "hard-water story." The hard-water story started in 1957 with Jun Kobayashi, a Japanese agricultural chemist. He had for many years been engaged in studies of the nature of irrigation water from an agricultural point of view. In these studies he found a close relation between the chemical composition of river water and the death rate from "apoplexy" (CVD). The death rate of apoplexy in Japan was extraordinarily high compared to other countries and the biggest cause of death in Japan. Kobayashi found that it was the ratio of sulfur to carbonate (SO_4/CaO_3) that was related to the death rate from apoplexy. He suggested that inorganic acid $CaCO_3$ might induce or prevent apoplexy (Kobayashi 1957).

Three years later a study was presented by Schroeder (1960), comprising the 163 largest cities in the United States. He found inverse correlations between water hardness and CVD in both men and women. He also studied the relation between different water constituents and coronary heart disease (CHD) among 45- to 64-year-old men. He found significant correlations between death rate from CHD and sulfate and bicarbonate, respectively, but not for the ratio of the agents. He also found significant negative correlations between deaths from CHD and magnesium, calcium, fluoride, and pH. The coefficient of correlation for magnesium in drinking water was $r = -0.30$, $p < 0.01$, and for calcium $r = -0.27$, $p < 0.01$ (Schroeder 1960).

During the following decades several studies on the relation between water hardness and CVD were presented. Initially the studies dealt with the question of whether there was a toxic effect in soft water or a protective effect in hard water. As described above, soft waters usually have low buffering capacity and are more corrosive, which leads to higher amounts of toxic trace elements. It has been proposed that this could account for the observed increased mortality. However, the results from previous studies have not supported this hypothesis (Marier 1986a). For example, in a nationwide survey of more than 500 tap waters in Canada, no significant correlation was found between mortality and trace elements like lead, cadmium, cobalt, lithium, mercury, molybdenum, nickel, or vanadium (Neri et al. 1975). Gradually it became obvious that it was a protective effect from hard water that was responsible for the relations seen. In different parts of the world, studies have been made on the relation between magnesium and calcium in the local drinking water and CVD mortality. These studies were generally based upon death registers and water data at regional or municipality levels. The results of these studies are summarized in Table 14.1.

Even with these studies, the results were not conclusive as to the role of magnesium and calcium in drinking water for CVD. Most of the studies showed a relation between different cardiovascular deaths and either magnesium or calcium, or both, but some studies showed no relation at all. However, most of these studies were ecological meaning that the exposure to water constituents was determined at group levels with a high risk of misclassification. Often, very large groups, for example, all inhabitants in large cities or areas, were assigned the same value of water magnesium and calcium, despite the presence of several waterworks or private wells.

In addition, the disease diagnoses studied were sometimes unspecific, with wide definitions that included both cardiac and cerebrovascular diseases. In some studies, it is also unclear whether the range of magnesium and calcium in drinking water was large enough to allow for appropriate analyses.

One of the most comprehensive studies of the geographic variations in cardiovascular mortality was the British Regional Heart Study. The first phase of this study (Pocock et al. 1980) applied multiple regression analysis to the geographical variations in CVD for men and women aged 35–74 in 253 urban areas in England, Wales, and Scotland from 1969 to 1973. The investigation showed that the effect of water hardness was nonlinear; much greater in the range from very soft to medium-hard water than from medium to very hard water. The geometric mean for the standardized mortality ratio (SMR) for CVD for towns grouped according to water hardness both with and without adjustments (by analysis of covariance) for the effects of four climatic and

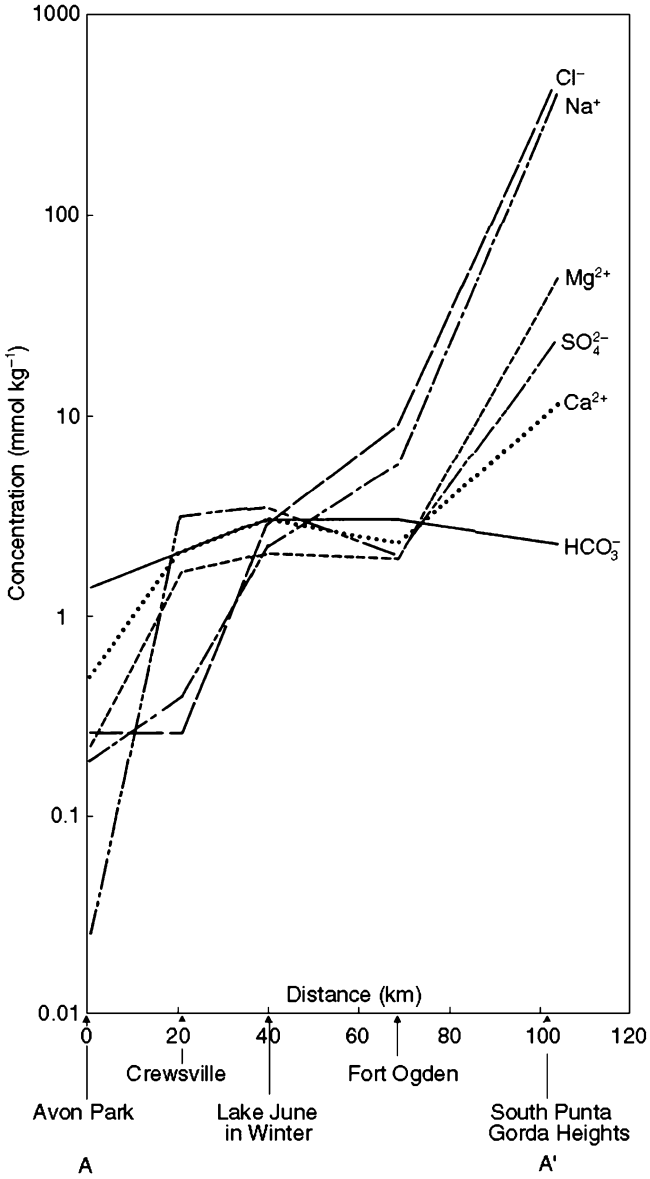


Fig. 14.3 Hydrochemical section along line A-A′ (for location see Fig. 14.2) showing the variation in concentrations of major ions downgradient in the direction of groundwater flow. Relative to calcium and magnesium, the groundwater experiences an increasing concentration of sodium as a result of ion ex-change and mixing with saline water in the coastal zone (After Sprinkle 1989)

Table 14.1 The effect of high Magnesium (Mg) or Calcium (Ca) levels in drinking water on the mortality from cardiovascular disease (CVD), coronary heart disease (CHD), or cerebrovascular disease (CD) in different studies published since 1975

	Mg			Ca		
Studies	CVD	CHD	CD	CVD	CHD	CD
North America						
Dawson et al. (1978)	¬			¬		
Neri et al. (1975)		¬			—	
Great Britain						
Shaper et al. (1980)	—			¬		
Maheswaran et al. (1999)		—			—	
Germany						
Teitge (1990)		¬				
Sweden						
Nerbrand et al. (1992)	—			¬		
Rylander et al. (1991)		¬	—		¬	¬
South Africa						
Leary et al. (1983)		¬				

¬ = lower mortality, — = no difference

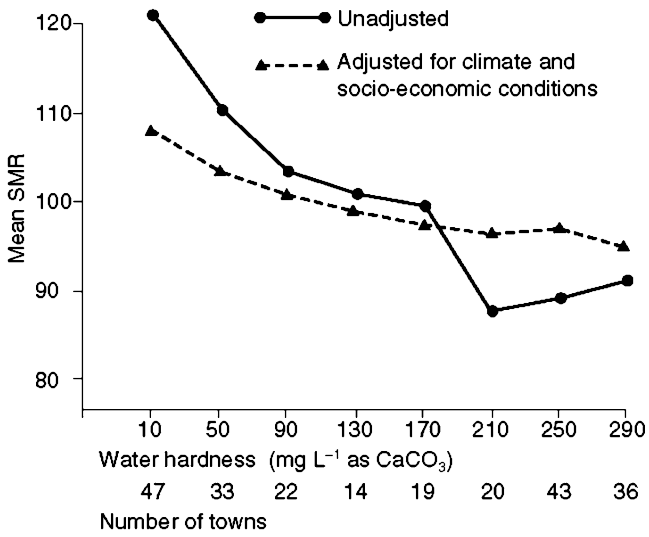


Fig. 14.4 Geometric means of the standardized mortality ratio (SMR) (for all men and women aged 35–74 with CVD) for towns in England, Wales, and Scotland grouped according to water hardness (in concentration units of mgL^{-1} as $CaCO_3$ equivalent) (After Pocock et al. 1980)

socioeconomic variables (percentage of days with rain, mean daily maximum temperature, percentage of manual workers, car ownership) is shown in Fig. 14.4. The adjusted SMR decreased steadily in moving from a hardness of 10–170 mg L^{-1} but changed little between 170–290 mg L^{-1} or greater. After adjustment, CVD in areas with very soft water, around 25 mg L^{-1}, was estimated to be 10–15% higher than in areas with medium-hard water, around 170 mg L^{-1}, whereas any further increase in hardness beyond 170 mg L^{-1} did not additionally lower CVD mortality. Hence, it appeared that the maximum effect on CVD was principally between the very soft and medium-hard waters. Adjusting for climatic and socioeconomic differences considerably reduced the apparent magnitude of the effect of water hardness (Pocock et al. 1980).

A problem with correlation studies such as the British Regional Heart Study, as argued by Jones and Moon (1987), is the failure of much of the research to consider the causal mechanism that links independent variables to the disease outcome. Also, many of the calibrated models presented in the literature are socially blind in including only variables

pertaining to the physical environment, which often contributes a large number of water quality elements. Even in those better analyses that have included social variables, as in the case of the British Regional Heart Study, the relatively strong correlation found for calcium in England and Wales may be a result of calcium acting as a very good surrogate for social variables. The soft water areas of the north and west of the British Isles equate to the areas of early industrialization, and today these areas house a disproportionate percentage of the socially disadvantaged (Jones and Moon 1987). Therefore, it is important that further studies undertake the challenge of quantitatively analyzing the separate effects of social variables from those of water hardness.

Fewer studies have been carried out in developing countries but Dissanayake et al. (1982), for example, found a negative correlation between water hardness and various forms of CVD and leukemia in Sri Lanka. More recent studies (Dissanayake 1991; Rajasooriyar 2003) have highlighted the problem of high fluoride concentrations and associated dental fluorosis in areas of hard water abstracted from crystalline bedrock aquifers in Sri Lanka.

Rajasooriyar (2003) measured total hardness in dug wells and tubewells in the Uda Walawe basin of southern Sri Lanka in the wide range of 7–3,579 mg L^{-1} as $CaCO_3$ with an average of 395 mg L^{-1} as $CaCO_3$ (Fig. 14.5). Compared with the government water quality limit of 600 mg L^{-1} as $CaCO_3$, 12% of the 102 samples collected during the wet season in 2001 were in excess of the limit and are considered too hard to drink (values above 100–150 mg L^{-1} as $CaCO_3$ are locally considered too hard as a water supply). Soft waters are found in areas with a dense irrigation network supplied by rain-fed surface reservoirs. Irrigation canal waters in the Suriyawewa and Uda Walawe subcatchments

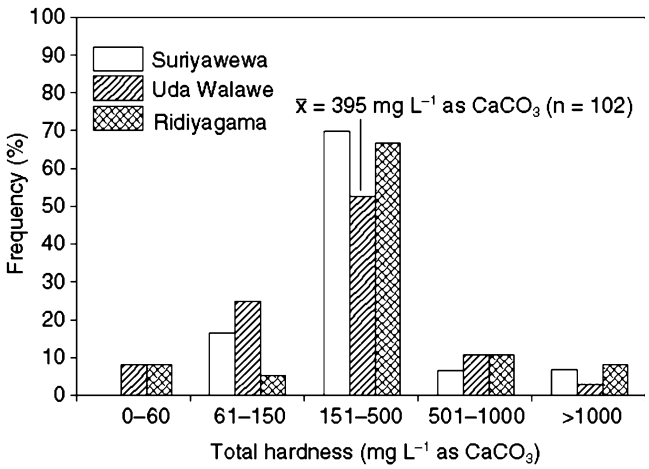


Fig. 14.5 Histogram of total hardness values recorded for groundwaters sampled during the wet season (January and February 2001) from dug wells and tubewells in three subcatchments of the Uda Walawe basin of Sri Lanka (Data courtesy of L. Rajasooriyar)

were measured in the dry season to have a total hardness in the range 40–90 mg L^{-1} as $CaCO_3$ (Rajasooriyar 2003), and it is leakage of this water source that leads to the softening of shallow groundwater. The majority of the groundwaters (63%) in the fractured aquifer represent very hard water with carbonate hardness in the range 151–500 mg L^{-1} as $CaCO_3$ contributed by the weathering of ferromagnesian minerals, anorthite, calcite, and dolomite. The products of this weathering lead to high concentrations of dissolved calcium, magnesium, and bicarbonate in groundwaters. Exceptionally high values of hardness (>1,000 mg L^{-1} as $CaCO_3$) typically occur in non-irrigated areas with additional non-carbonate hardness contributed by pyrite oxidation and, in the case of the coastal Ridiyagama coastal catchment, by salt water inputs.

14.4 Studies at an Individual Level

In recent years some studies have been made with a higher precision as regards exposure classification than the previously cited ecological studies. Most of them have been conducted with a case–control design. This means that the exposure to water magnesium and calcium has been estimated for individuals who have suffered from the disease as well as for healthy control persons, and the difference in risk calculated. A similar type of study design is the prospective cohort study, where the subjects are followed over time.

14.4.1 Studies in Finland

In 1979 Punsar and Karvonen made a cohort study in which they compared the death rates from CHD in two rural areas in western and eastern Finland (Punsar and Karvonen 1979). The cohort comprised men 40–59 years old in 1959, and they were followed for 15 years. The water data were not truly individual but were median values in ten subareas in the western area and 33 subareas in the eastern area. The ranges of subarea medians were 6.9–27.8 mg L^{-1} of water magnesium in the western area and 0.6–7.3 mg L^{-1} in the eastern area. The cohort in eastern Finland had a death rate from CHD which was 1.7 times higher than the western cohort. Calcium was not investigated.

A few years later, Luoma et al. (1983) published a case–control study conducted in the southeastern region of Finland. Cases were men 30–64 years of age with a first acute myocardial infarction (AMI; alive or deceased), who were pair-matched with hospital controls for age and region (rural vs. urban). In addition, population controls were selected and matched for age and municipality. All subjects submitted a sample of their drinking water. The range of magnesium in the drinking water was 1.0–57.5 mg L^{-1} for

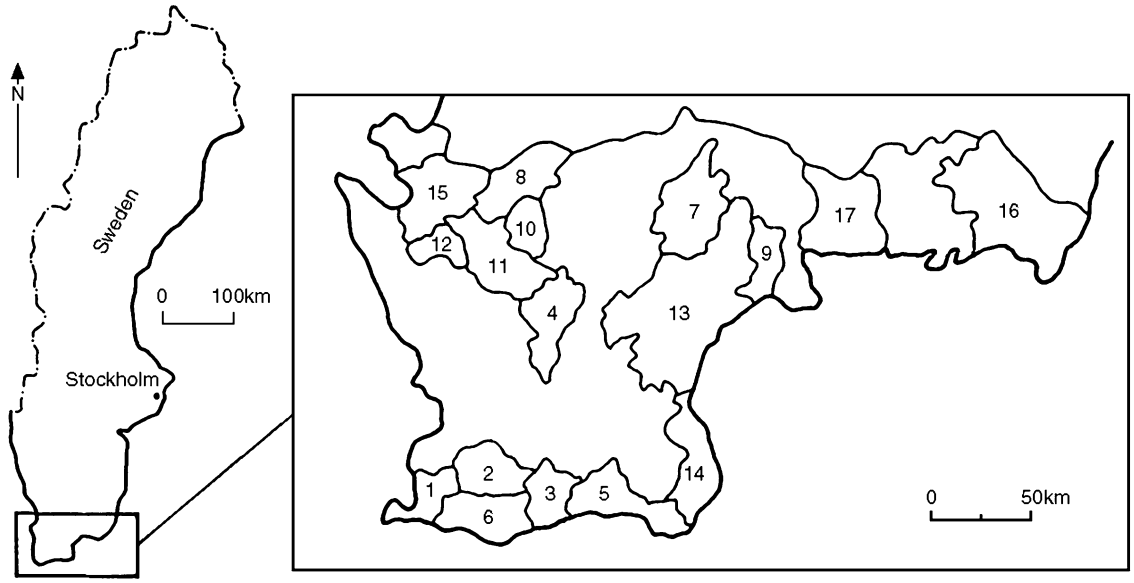


Fig. 14.6 The country of Sweden with Skåne and Blekinge counties enlarged. The 17 municipalities in the studies are numbered corresponding to Table 14.2, i.e., 1 Vellinge, 2 Svedala, 3 Skurup, 4 Höör, 5 Ystad, 6 Trelleborg, 7 Göinge, 8 Örkelljunga, 9 Bromölla, 10 Perstorp, 11 Klippan, 12 Åstorp, 13 Kristianstad, 14 Simrishamn, 15 Ängelholm, 16 Karlskrona, and 17 Karlshamn.

the cases and 0.75–30.0 mg L^{-1} and 1.0–16.0 mg L^{-1} for the hospital controls and population controls, respectively. For case-population control comparisons, the relative risk (RR) with 95% confidence limits was 4.7 (95% confidence interval [CI] 1.3–25.3) for magnesium levels lower than 1.2 mg L^{-1}. This means that those with the lowest magnesium levels had a risk of AMI almost five times higher than those with higher magnesium levels. For the case-hospital control comparisons, the RR was 2.0 (95% CI 0.7–6.5), that is double the risk. They also found inverse relations with fluoride levels but no relation to calcium.

14.4.2 Studies in Sweden

In Sweden, three case–control studies have been conducted during the last decade. First, using mortality registers, the relation between death from AMI and the level of magnesium and calcium in drinking water was examined among men. A few years later a study with a similar design was made comprising women. The studies were conducted in southern Sweden, in a relatively small geographic area, where there was a great difference between as well as within the municipalities regarding magnesium and calcium content in drinking water. The advantage with this limited study area was that the possible risk of such confounding factors as climate, geographical, cultural, and socioeconomic differences was minimized.

Seventeen municipalities were identified whose water quality with respect to water hardness, acidity, and treatment procedures had been basically unchanged (change of hardness <10% and pH <5%) during the most recent 10 years. Figure 14.6 shows the location of the 17 municipalities that comprised the study area.

Cases were men (n = 854) and women (n = 378) in 17 municipalities in southern Sweden who had died of AMI between ages 50 and 69 years. Controls were men (n = 989) and women (n = 1,368) of the same age group who had died of cancer. Individual water data were collected. Table 14.2 shows the number of waterworks, the range of magnesium in drinking water, and the amount of magnesium in water supplied to the most densely populated area in each municipality.

The subjects were divided into quartiles according to the levels of magnesium. Odds ratios were calculated in relation to the group with the lowest exposure. Adjustments for age were made in all analyses. The results show odds ratios of 0.65 for men and 0.70 for women in the quartile with highest magnesium levels in the drinking water (≥9.8–9.9 mg L^{-1}) (Fig. 14.7). This means that the risk of dying from AMI was about 30% lower compared with the risk for those who used drinking water with the lowest levels of magnesium (Rubenowitz et al. 1996, 1999).

A few years later a prospective interview study was conducted in the same area where men and women who suffered from AMI from 1994 to 1996 were compared with population controls (Rubenowitz et al. 2000). The results showed that magnesium in drinking water protected against

Table 14.2 Number of waterworks, range of Magnesium (Mg) concentrations in drinking water (mgL^{-1}), and amount of Mg in water supplied to the most densely populated areas (mgL^{-1})

Municipality no.[a]	Number of waterworks in the study	Range of water Mg (mg L^{-1})	Water Mg in the most densely populated area (mg L^{-1})
1	2	5.9–10.0	10.0
2	2	6.8–20.0	20.0
3	4	2.6–10.0	9.0
4	2	8.8–13.0	8.8
5	5	5.1–11.9	7.5
6	4	6.5–18.0	16.0
7	9	3.3–13.5	5.0
8	3	3.0–6.9	6.9
9	5	1.3–7.6	1.3
10	2	4.0–9.0	9.0
11	2	5.5–9.0	8.0
12	1	10.0	10.0
13	15	1.3–14.4	6.7
14	9	3.8–13.4	9.7
15	2	7.0–11.0	11.0
16	9	2.0–13.0	2.0
17	2	1.9–5.0	2.0

[a]Municipality numbers correspond to Fig. 14.6

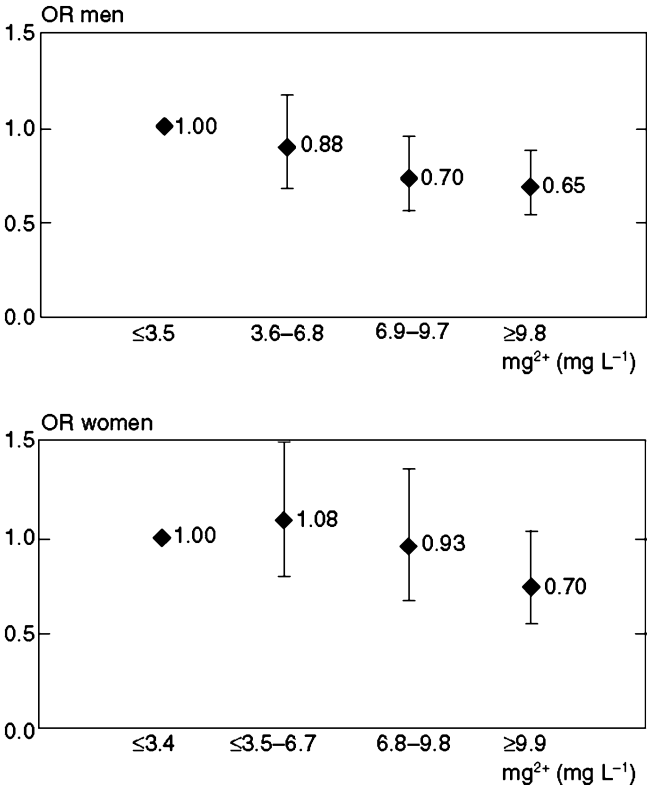


Fig. 14.7 Odds ratios (OR) with 95% confidence intervals for death from AMI in relation to magnesium in drinking water (adjusted for age and calcium) in men and women

death from AMI, but the total incidence was not affected. In particular, the number of deaths outside hospitals was lower in the quartile with high magnesium levels. This supports the hypothesis that magnesium prevents sudden death from AMI rather than all CHD deaths. The mechanisms that could explain these findings are discussed below.

14.4.3 Studies in Taiwan

In a study from Taiwan (Yang 1998) all eligible cerebrovascular deaths (17,133 cases) of Taiwan residents from 1989 to 1993 were included and compared with residents who had died from other causes (17,133 controls). The controls were pair matched to the cases by sex, year of birth, and year of death. The levels of calcium and magnesium in drinking water of these residents were determined. The adjusted odds ratios (95% confidence interval) were 0.75 (0.65–0.85) for the group with water magnesium levels between 7.4 and 13.4 mg/L and 0.60 (0.52–0.70) for the group with magnesium levels of 13.5 mg/L or more. After adjustment for magnesium levels in drinking water, there was no difference between the groups with different levels of calcium.

In the same way all hypertension deaths (2,336 cases) of Taiwan residents from 1990 to 1994 were compared with deaths from other causes (2,336 controls), and related to the levels of calcium and magnesium in the drinking water (Yang and Chiu 1999). The adjusted odds ratios for the highest versus lowest quintiles of exposure was 0.63 (0.47–0.84). No effect of calcium was shown. The risk reduction of mortality from stroke and hypertensive disease in relation to drinking water magneisum was thus similar to the risk reduction of death from acute myocardial infarction shown in the Swedish studies.

14.5 Physiological Importance of Magnesium

14.5.1 Physiological Properties of Magnesium in Humans

Magnesium is involved in several important enzymatic reactions. All reactions that involve ATP (adenosine triphosphate), that is energy demanding reactions, have an absolute magnesium requirement (Reinhardt 1988). Important processes in the body mediated by magnesium are, for example, synthesis of protein, nucleic acid and fat, glucose use, neuromuscular transmission, muscular contraction, and transport over cell membranes (Altura and Altura 1996) (see also Chap. 32, this volume).

Magnesium is essential to the cardiovascular system, and two properties are especially important: stabilizing the cardiac electric system (preventing cardiac arrhythmia) and regulating vascular tone (Reinhardt 1991). There are multiple mechanisms behind these properties.

Magnesium is needed to maintain the normal gradient of potassium and calcium over cell membranes (Altura et al. 1981). It is well known that magnesium is necessary to maintain intracellular levels of potassium. This is done by blocking the outward passage of potassium through the cell membrane and by activating the enzyme Na/K-ATPase. It has a similar function in Ca-ATPase. Magnesium also has a direct effect on potassium and calcium channels in the cell membranes (Reinhardt 1991).

Furthermore, as regards impact on vascular tone, magnesium is a necessary activator for the synthesis of cyclic adenosine monophosphate (c-AMP), which is a vasodilator. It also acts as a natural calcium antagonist by competing for calcium binding sites in the vascular smooth muscle and thus reducing the constrictive effect of calcium in the blood vessels (Reinhardt 1991). In addition, the vasoconstrictive actions of hormones such as angiotensin, serotonin, and acetylcholine are enhanced in the case of magnesium deficiency (Altura et al. 1981).

14.6 Magnesium Intake

14.6.1 Magnesium Deficiency

Marginal magnesium deficiency probably affects a large proportion of the population whose dietary intake does not reach the recommended amount. Although severe magnesium deficiency is not common in the population, hypomagnesemia is often present among hospitalized patients (Altura 1994). A survey showed that hypomagnesemia was the most common electrolyte abnormality in patients entering the intensive care unit and was present in 20% of the patients (Reinhardt 1988) (see also Chap. 9, this volume).

The causes of hypomagnesemia include reduced intake caused by starvation or intravenous therapy without magnesium supplement; impaired absorption owing to chronic diarrhea or malabsorption syndromes; increased renal loss caused by diseases such as diabetes or renal diseases or by the use of alcohol or such drugs as diuretics or antibiotics (Reinhardt 1988).

14.6.2 Magnesium Intake from Food

The largest part of the total magnesium intake is from foodstuffs. Magnesium is present in many foodstuffs and in large amounts in nuts, beans, green leafy vegetables, and whole grain cereals. Mainly because of increased industrial treatment, which decreases magnesium levels by 80–95% (Marier 1986b), it has been suggested that the majority of people today have a lower magnesium intake than the recommended dietary amount (RDA) of 6 mg kg^{-1} day^{-1} (Marier 1986b; Seelig 1986; Durlach 1989). Table 14.3 shows the magnesium content in some foodstuffs and beverages.

Approximately 40% of magnesium ingested in food is normally absorbed (Hardwick et al. 1990). The proportion absorbed is inversely related to the amount ingested, however, and has been shown to range between approximately 10 and 70% (Fine et al. 1991). The absorption rate also depends on the intake of other foodstuffs. Elements such as calcium, phosphorus, fibers, and phytic acids are known to diminish the absorption rate of magnesium (Seelig 1986). Most magnesium is absorbed in the distal small intestine, mainly by passive diffusion through the paracellular pathway, but also to a smaller extent by solvent drag and active transport (Hardwick et al. 1990).

14.7 Water Magnesium and Body Magnesium Status

14.7.1 Magnesium Intake from Water—Importance for Body Magnesium Status

Magnesium has a natural source in water from the weathering of a range of rock types. In igneous rock, magnesium is typically a constituent of the dark colored ferromagnesian minerals, which include olivine, pyroxenes, amphiboles, and dark colored micas. In altered rocks, magnesian mineral species also occur such as chlorite, montmorillonite, and serpentine. Sedimentary forms of magnesium include magnesite ($MgCO_3$) and dolomite ($CaMg(CO_3)_2$). Magnesium is substantially less abundant than calcium in all rock types and so in most natural waters the magnesium concentration is much lower, usually by 5–10 times, than

Table 14.3 Magnesium (Mg) content in food and beverages

Food items and beverages	Mg (mg 100 g^{-1})	Food items and beverages	Mg (mg 100 g^{-1})
Food items		*Food items*	
Bread, white	23	Bananas	33
Bread, fiber-rich	50	Fish	25
Muesli	100	Shrimp	42
Corn flakes	16	Meat	25
Cheese	34	Eggs	13
Common beans, white, dried	184	Almonds	280
Common beans, brown, dried	131	Peanuts	188
Corn	23	Dark chocolate	130
Spinach	79	Milk chocolate	60
Avocado	39		
Broccoli	23	*Beverages*	
Tomato	10	Milk	12
Cucumber	10	Coffee[a]	5
Mushrooms	13	Tea[a]	3
Brown rice	110	Beer, low alcohol content	8
White rice	34	Beer, high alcohol content	11
Potatoes	24	Red wine	12
Apples	5	White wine	3
Oranges	10	Spirits	0

[a]Magnesium content in the water not included

the calcium concentration. Concentrations of magnesium in fresh waters are controlled by solution and precipitation reactions involving magnesium-bearing silicate and carbonate minerals with concentrations typically less than 50 mg L^{-1}, although values above 100 mg L^{-1} are recorded.

Magnesium intake via water depends on the level in drinking water. An important issue in discussions about the relation between water magnesium and AMI is whether magnesium in drinking water can be critical for the body magnesium status, as the main part of the magnesium intake derives from food (Neutra 1999). It has been suggested that the quantitative contribution of water magnesium may be crucial for body magnesium status for those who have a low dietary intake and use water with high magnesium levels (Durlach et al. 1989). In addition, cooking food in magnesium-poor water leaches out magnesium, while cooking in magnesium-rich water diminishes this loss (Haring and van Delft 1981).

Furthermore, it has been suggested that magnesium in water, appearing as hydrated ions, has a higher bioavailability than magnesium in food, which is bound in different compounds that are less easily absorbed (Durlach et al. 1989; Theophanides et al. 1990). However, simultaneous intake of other agents could diminish the absorption rate, as discussed above.

Plants cultivated in areas with magnesium-rich water may have higher magnesium content, especially if the soil is magnesium-rich and the land is irrigated with magnesium-rich water. People living in such areas who eat locally grown vegetables and fruits may also benefit from an addition to the total magnesium intake, especially during summer months. However, genetic factors appear to have a greater effect on plant magnesium composition than do soil and environmental factors (Wilkinson et al. 1987).

Some previous studies have shown relations between water magnesium and body magnesium content. In a study of baboons, tap water was more effective than dietary supplementation in increasing serum levels of magnesium and zinc (Robbins and Sly 1981). Anderson et al. (1975) found relations between water magnesium and magnesium content in the heart muscle. The myocardial magnesium was approximately 7% lower among residents of cities with soft water. There was, however, no difference in magnesium content in the diaphragm or the pectoralis major muscle. Other studies have shown relations between water magnesium levels and the magnesium content in skeletal muscle and in coronary arteries (Landin et al. 1989).

In a loading test, ten subjects who normally used water with a magnesium level of 1.6 mg L^{-1}, instead were given drinking water with 20 mg L^{-1} magnesium. After 6 weeks of supplementation with magnesium-enriched water, the excretion of magnesium was increased, which indicated improved body magnesium status (Rubenowitz et al. 1998).

Another study demonstrated that the ionized serum magnesium level was raised after only 6 days of an increased dietary load of magnesium, but not the total body magnesium.

However, a correlation between ionized serum and ionized intracellular magnesium was shown (Altura and Altura 1996).

From the available data it is apparent that the quantitative contribution from water may be crucial. The magnesium levels in drinking water in the Swedish case–control studies ranged from 0 to 44 mg L^{-1}. With a total daily intake of drinking water of 2 l, the proportional magnesium contribution from water thus ranges from 0 to 88 mg day^{-1}. This is a percentage contribution between 0 and 25% of the RDA of 350 mg day^{-1}. For those who have a daily intake lower than the RDA and use magnesium-rich water, the contribution would be even more important.

In the prospective study, the calculated intake of magnesium from food ranged from 157 to 658 mg day^{-1}, median 356 mg day^{-1}. This means that a large number of the subjects had a lower intake than the RDA. One subject with an intake of 157 mg magnesium per day used drinking water with 3.5 mg L^{-1}, which means a 4.5% addition to the magnesium intake from food. If she instead had used water with 40 mg L^{-1}, the addition would have been 50%, and the total daily intake 240 mg day^{-1}.

14.8 Calcium in Drinking Water and Cardiovascular Disease

The majority of the previous ecological studies showed an inverse relation between calcium in drinking water and CVD, as reported above. The physiological mechanisms that could explain the relationship are not clear. There is, however, evidence that calcium deficiency can cause hypertension, which is a well-known risk factor for both stroke and AMI (Lau and Eby 1985; Moore 1989; Waeber and Brunner 1994).

14.8.1 Calcium Deficiency

Calcium deficiency is common among the elderly, especially women. The absorption and renal conservation of calcium decreases with age. The absorption of calcium from food varies between 15 and 75% (Schaafsma 1992), but in menopausal women the absorption is only about 20–30% (Heany and Recker 1985). Furthermore, calcium intake is often decreased among the elderly (Harlan et al. 1984). The RDA for calcium in Sweden is 800 mg day^{-1} for adult women and 600 mg day^{-1} for men. In the United States, the RDA is 1,000–1,500 mg day^{-1} for adults.

A study comprising 61,000 women in Sweden aged 40–76 showed that calcium intake decreased with age and that a majority of postmenopausal women had a deficient calcium intake (Michaelsson 1996). Several studies conducted in the United States have also shown an intake lower than recommended, especially among women (Fleming and Heimbach 1994). For individuals with a deficiency, the additional calcium from water could be crucial to prevent this. Along with the contribution of drinking water, cooking food in calcium-rich water has been shown not only to prevent leaching, but even to increase calcium levels in the food (Haring and van Delft 1981).

14.8.2 Calcium and Blood Pressure

Several studies have shown an inverse relation between dietary calcium intake and blood pressure. Meta-analysis comprising nearly 40,000 people has shown that a high calcium intake lowered both systolic and diastolic blood pressure (Cappucio et al. 1995). Low serum concentrations of ionized calcium have been measured in patients with hypertension (McCarron 1982). There are several possible mechanisms that could explain how calcium lowers blood pressure.

One mechanism may be that hypocalcemia inhibits Ca-ATPase activity, which leads to an increase in free intracellular calcium and contraction of vascular smooth muscles (McCarron 1985). Calcium supplementation has been shown to be efficacious, especially among salt-sensitive hypertensives with a deficient basal calcium intake (Sowers et al. 1991). These individuals often have increased levels of the calcium regulatory hormones parathyroid hormone and active vitamin D (1,25-$(OH)_2$-D), which can cause increased peripheral resistance. Dietary calcium suppresses these hormones which causes the blood pressure to decrease. Calcium and calcium regulatory hormones may also influence blood pressure regulation via the central nervous system. Calcium also induces natriuresis, which has been shown to lower blood pressure in postmenopausal hypertensive women (Johnson et al. 1985).

The case–control studies reported above showed no clear relation between calcium and AMI.

14.9 Water Hardness and Other Health Effects

The results of several studies have suggested that a variety of other diseases are also correlated with water hardness, which include various types of cancer. Incidences of a correlation between cancer and water hardness have been reported from studies in Finland and Taiwan. In northern Finland, correlations of age-adjusted incidences of various forms of cancer with the geochemical composition of groundwater were undertaken by comparing geochemical maps showing the hardness, uranium, iron, and nitrate content of water and maps showing the areal distribution of the incidences of ten

forms of cancer (Piispanen 1991). A statistically significant positive correlation was identified between water hardness and several forms of cancer, especially for all forms of cancer combined in the female population ($r = 0.66$). Piispanen (1991) suggested that drinking hard water may be an initiator and promoter of cancer, although it is admitted that a positive correlation between the geochemical and medical variables does not necessarily prove a cause-and-effect relationship between these variables.

However, in a case–control study, Yang et al. (2002) compared a group of Taiwan residents who had died from esophageal cancer with a control group that consisted of people who had died from causes other than cancer, matched by sex, year of birth, and year of death. Data on calcium and magnesium levels in drinking water were obtained. The results showed that the risk of death from esophageal cancer decreased significantly with higher magnesium levels. Compared to those with magnesium levels below 7.0 mg/l, the adjusted odd ratios were 0.75 (0.62–0.91) for the group with water magnesium levels between 7.3 and 14.0 mg/l and 0.71 (0.57–0.88) for the group with magnesium levels of 14.3 mg/l or more. This means about 25–30% lower risk with higher magnesium levels. No effect of water calcium levels was seen.

In a Japanese study of water hardness, regional geological features, and the incidence of struvite stones, Kohri et al. (1993) found a positive correlation between the magnesium-calcium ratio of tap water and the incidence of struvite stones. The incidence of struvite stones was both high in regions of basalt and sedimentary rock and low in granite and limestone areas.

In an ecological study of the relation between domestic water hardness and the prevalence of eczema among primary-school-age children in Nottingham, England, McNally et al. (1998) found a significant direct relation between a 1-year period and lifetime prevalence of eczema and water hardness, both before and after adjustment for confounding factors (sex, age, socioeconomic status, access to health care). Eczema prevalence trends in the secondary-school population were not significant.

World health organization (WHO 2009) recently published a comprehensive survey about public health significans of calcium and magnesium in drinking water. The report contains a summary of the study results showing a reduction of cardiovascular disease mortality risk with increasing magnesium levels in drinking water. It also discusses a protective relation between magnesium intake and hypertension, cardiac arrythmias, pre-eclampsia (hypertension after 20 weeks of gestation), atherosclerosis and diabetes mellitus. Regarding calcium the survey reports a possible positive effect on osteoporosis, with increasing bone mass with higher calcium intake. The incidence on kidney stones have been suggested to be lower with calcium intake from water and food but higher with calcium supplements. In addition, a positive effect on insulin resistance, and hypertension and stroke is discussed.

14.10 Conclusions and Consequences

The significance of earlier studies of the link between water hardness and various health effects is unclear, and it is suggested that the reported associations may reflect disease patterns that can be explained by social, climatological, and environmental factors rather than by the hardness of the water. However, several studies during recent years have shown a significant protective effect of magnesium in drinking on cardiovascular disease mortality. Catling et al. (2005) have made an systematic review including several of the above mentioned epidemiological cohort and case–control studies. They made a meta-analysis of the case–control studies and concluded that the review found evidence to support an inverse association between cardiovascular disease mortality and drinking water magnesium levels. There was limited evidence to support an association between cardiovascular mortality and calcium levels. As mentioned above these results have also been summarized in a WHO report (WHO 2009).

14.10.1 A Public Health Perspective

Marier and Neri (1985) attempted to quantify the importance of water magnesium using a number of the abovementioned epidemiological studies. They estimated that an increase in water magnesium level of 6 mg L^{-1} would decrease CHD mortality by approximately 10%.

The data collected in the Swedish case–control studies can be used to estimate the impact of water magnesium on the incidence of myocardial infarction in the study population. If everyone in the male study base were to drink water from the highest quartile ($\geq$9.8 mg L^{-1}), the decrease in mortality from AMI would be about 19%. This means that the age-specific incidence of death from myocardial infarction in the study area would change from about 350/100,000 $year^{-1}$ to 285/100,000 $year^{-1}$. The decrease of the incidence per mg L^{-1} magnesium can be calculated to be approximately 10/100,000 $year^{-1}$, which is an even larger decrease than Marier and Neri (1985) estimated.

For women, the corresponding decrease in mortality from AMI would be about 25% if everyone were to drink water from the highest quartile ($\geq$9.9 mg L^{-1}). The age-specific mortality among women in the study area would change from about 92/100,000 $year^{-1}$ to 69/100,000 $year^{-1}$.

14.10.2 Consequences

What should be the consequences of today's knowledge of a recommended intake of magnesium? Is there sufficient evidence to recommend an increased intake of magnesium? If so, should the recommendation be applicable to the whole population or only to certain risk groups? Furthermore, what would be the best way to increase magnesium intake?

The consumption of magnesium-rich food and water can be encouraged, and the use of water softeners in areas with hard water discouraged. At least one tap with unfiltered water for drinking should always be left. Although such recommendations are hardly controversial, it is difficult to make them effective, especially if the target is the population as a whole.

Theoretically, a possible way of increasing magnesium intake would be to add magnesium to drinking water, especially in areas with naturally soft water. This is already done in some waterworks in order to reduce corrosiveness. To do so on a larger scale would, however, be expensive and politically difficult.

Recommendations to use oral magnesium supplementation would not be practically feasible for the whole population, but could be considered to be directed toward certain risk groups.

Before a general prevention program can be accepted, large-scale intervention studies must be conducted to accurately evaluate the preventive effect of magnesium. This would require several thousands of subjects, however, and would thus be cumbersome and expensive. Nevertheless, in view of the significant implications for public health, such studies should be conducted. The possibility of a simple and harmless way of reducing AMI mortality rate must not be overlooked.

See Also the Following Chapters. Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 9 (Biological Responses of Elements) • Chapter 24 (Environmental Medicine) • Chapter 25 (Environmental Pathology) • Chapter 33 (Modeling Groundwater Flow and Quality)

Further Reading

Altura BM (1994) Introduction: importance of magnesium in physiology and medicine and the need for Ion selective electrodes. Scand J Clin Lab Invest 54(Supplement 217):5–9

Altura BM, Altura BT (1996) Role of magnesium in patophysiological processes and the clinical utility of magnesium ion selective electrodes. Scand J Clin Lab Invest 56(Supplement 224):211–234

Altura BM, Altura BT, Carella A, Turlapaty PDMV (1981) Hypomagnesemia and vasoconstriction: possible relation to etiology of sudden death ischemic heart disease and hypertensive vascular disease. Artery 9:212–231

Anderson TW, Neri LC, Schreiber GB, Talbot FDF, Zdrojewski A (1975) Ischemic heart disease, water hardness and myocardial magnesium. Can Med Assoc J 113:199–203

Cappucio F, Elliot P, Allender PS, Pryer J, Follman PA, Cutler JA (1995) Epidemiologic association between dietary calcium intake and blood pressure: a meta-analysis of published data. Am J Epidemiol 142:935–945

Catling LA, Abubakar I, Lake IR, Swift L, Hunter PR (2005) A systematic review of analytical observational studies investigating the association between cardiovascular disease and drinking water hardness. J Water Health 6(4):433–442

Dawson EB, Frey MJ, Moore TD, McGanity J (1978) Relationship of metal metabolism to vascular disease mortality rates in Texas. Am J Clin Nutr 31:1188–1197

Dissanayake CB (1991) The fluoride problem in the groundwater of Sri Lanka – environmental management and health. Int J Environ Stud 38:137–156

Dissanayake CB, Senaratne A, Weerasooriya VR (1982) Geochemistry of well water and cardiovascular diseases in Sri Lanka. Int J Environ Stud 19:195–203

Durlach J (1989) Recommended dietary amounts of magnesium: magnesium RDA. Magnes Res 2:195–203

Durlach J, Bara M, Guiet-Bara A (1989) Magnesium level in drinking water: its importance in cardiovascular risk. In: Itokawa Y, Durlach J (eds) Magnesium in health and disease. John Libbey & Co., London

Fine KD, Santa Ana CA, Porter JL, Fordtran JS (1991) Intestinal absorption of magnesium from food and supplements. J Clin Invest 88:396–402

Fleming KH, Heimbach JT (1994) Consumption of calcium from the US: food sources and intake levels. J Nutr 124:1426S–1430S

Freeze RA, Cherry JA (1979) Groundwater. Prentice Hall, Englewood Cliffs

Hardwick LL, Jones MR, Brautbar N, Lee DBN (1990) Site and mechanism of intestinal magnesium absorption. Min Electrolyte Metab 16:174–180

Haring BSA, van Delft W (1981) Changes in the mineral composition of food as a result of cookingin "hard" and "soft" waters. Arch Environ Health 36:33–35

Harlan WR, Hull AL, Schmouder RL, Landis JR, Thompson FE, Larkin FA (1984) Blood pressure and nutrition in adults. Am J Epidemiol 120:17–28

Heany RP, Recker RR (1985) Estimation of true calcium absorption. Ann Int Med 103:516–521

Hem JD (1985) Study and interpretation of the chemical characteristics of natural water, 3rd edn. United States geological survey water supply paper 2254

Hiscock K, Paci A (2000) Groundwater resources in the quaternary deposits and lower Paleozoic bedrock of the Rheidol catchment, West Wales. In: Robins NS, Misstear BDR (eds) Groundwater in the Celtic regions: studies in hard rock and quaternary hydrogeology. Geological Society, London, Special Publications, 182: 141–155

Johnson NE, Smith IL, Freudenhim JL (1985) Effects on blood pressure of calcium supplementation of women. Am J Clin Nutr 42:12–17

Jones K, Moon G (1987) Health, disease and society: a critical medical geography. Routledge & Kegan Paul, London, pp 134–140

Kobayashi J (1957) On geographical relations between the chemical nature of river water and death rate from apoplexy. Ber Ohara Inst Landwirtsch Biol 11:12–21

Kohri K, Ishikawa Y, Iguchi M, Kurita T, Okada Y, Yoshida O (1993) Relationship between the incidence of infection stones and the magnesium-calcium ratio of tap water. Urol Res 21:269–272

Landin K, Bonevik H, Rylander R, Sandström B (1989) Skeletal muscle magnesium and drinking water magnesium level. Magnes Bull 11:177–179
Lau K, Eby B (1985) The role of calcium in genetic hypertension. Hypertension 7:657–667
Leary WP, Reyes AJ, Lockett CJ, Arbuckle DD, van der Byl K (1983) Magnesium and deaths ascribed to ischaemic heart disease in South Africa. S Afr Med J 64:775–776
Luoma H, Aromaa A, Helminen S, Murtomaa H, Kiviluoto L, Punsar S, Knekt P (1983) Risk of myocardial infarction in Finnish Men in relation to fluoride, magnesium and calcium concentration in drinking water. Acta Med Scand 213:171–176
Maheswaran R, Morris S, Falconer S, Grossinho A, Perry I, Wakefield J, Elliot P (1999) Magnesium in drinking water supplies and mortality from acute myocardial infarction in North West England. Heart 82:455–460
Marier JR (1986a) Role of magnesium in the "hard-water story". Magnes Bull 8:194–198
Marier JR (1986b) Magnesium content of the food supply in the modern-Day world. Magnesium 5:1–8
Marier JR, Neri LL (1985) Quantifying the role of magnesium in the interrelation between human mortality/morbidity and water hardness. Magnesium 4:53–59
McCarron DA (1982) Low serum concentrations of ionized calcium in patients with hypertension. N Engl J Med 307:226–228
McCarron DA (1985) Is calcium more important than sodium in the pathogenesis of essential hypertension? Hypertension 7:607–627
McNally NJ, Williams HC, Phillips DR, Smallman-Raynor M, Lewis S, Venn A, Britton J (1998) Atopic eczema and domestic water hardness. Lancet 352:527–531
Michaelsson K (1996) Diet and osteoporosis. Medical dissertations, Uppsala University
Moore TJ (1989) The role of dietary electrolytes in hypertension. J Am Coll Nutr 8:68S–80S
Nerbrand C, Svärdsudd K, Ek J, Tibblin G (1992) Cardiovascular mortality and morbidity in seven counties in Sweden in relation to water hardness and geological settings. Eur Heart J 13:721–727
Neri LC, Hewitt D, Schreiber GB, Anderson TW, Mandel JS, Zdrojewsky A (1975) Health aspects of hard and soft waters. J Am Water Works Assoc 67:403–409
Neutra RR (1999) Epidemiology vs. Physiology? drinking water magnesium and cardiac mortality. Epidemiology 10:4–6
Piispanen R (1991) Correlation of cancer incidence with groundwater geochemistry in Northern Finland. Environ Geochem Health 13:66–69
Pocock SJ, Shaper AG, Cook DG, Packham RF, Lacey RF, Powell P, Russell PF (1980) British regional heart study: geographic variations in cardiovascular mortality, and role of water quality. Br Med J 280:1243–1249
Punsar S, Karvonen MJ (1979) Drinking water quality and sudden death: observations from West and East Finland. Cardiology 64:24–34
Rajasooriyar LD (2003) A study of the hydrochemistry of the Uda Walawe Basin, Sri Lanka, and the factors that influence groundwater quality. Ph.D. thesis, University of East Anglia, Norwich
Reinhardt RA (1988) Magnesium metabolism: a review with special reference to the relationship between intracellular content and serum levels. Arch Int Med 148:2415–2420
Reinhardt RA (1991) Clinical correlates of the molecular and cellular actions of magnesium on the cardiovascular system. Am Heart J 121:1513–1521
Robbins DJ, Sly MR (1981) Serum zinc and demineralized water. Am J Clin Nutr 34:962–963
Rubenowitz E, Axelsson G, Rylander R (1996) Magnesium in drinking water and death from acute myocardial infarction. Am J Epidemiol 143:456–462
Rubenowitz E, Axelsson G, Rylander R (1998) Magnesium in drinking water and body magnesium status, measured using an oral loading test. Scand J Clin Lab Invest 58:423–428
Rubenowitz E, Axelsson G, Rylander R (1999) Magnesium in drinking water and death from acute myocardial infarction among women. Epidemiology 10:31–36
Rubenowitz E, Molin I, Axelsson G, Rylander R (2000) Magnesium in drinking water in relation to morbidity and mortality from acute myocardial infarction. Epidemiology 11:416–421
Rylander R, Bonevik H, Rubenowitz E (1991) Magnesium and calcium in drinking water and cardiovascular mortality. Scand J Work Environ Health 17:91–94
Schaafsma G (1992) The scientific basis of recommended dietary allowances for calcium. J Int Med 231:187–194
Schroeder HA (1960) Relation between mortality from cardiovascular disease and treated water supplies. Variations in states and 163 largest municipalities of the United States. J Am Med Assoc 172:1902–1908
Seelig MS (1986) Nutritional status and requirements of magnesium: with consideration of individual differences and prevention of cardiovascular disease. Magnes Bull 8:170–185
Shaper AG, Packham RF, Pocock SJ (1980) The British regional heart study: cardiovascular mortality and water quality. J Environ Pathol Toxicol 3:89–111
Sowers JR, Zemel MB, Zemel PC, Standley PR (1991) Calcium metabolism and dietary calcium in salt sensitive hypertension. Am J Hypertens 4:557–563
Sprinkle CL (1989) Geochemistry of the Florida Aquifer System in Florida and in parts of Georgia, South Carolina, and Alabama, United States geological survey professional paper 1403-I
Teitge JE (1990) Incidence of myocardial infarction and the mineral content of drinking water (in German). Z Gesamte Iinn Med Ihre Grenzgeb 45:478–485
Theophanides T, Angiboust J-F, Polissiou M, Anastassopoulous J, Manfait M (1990) Possible role of water structure in biological magnesium systems. Magnes Res 3:5–13
Waeber B, Brunner HR (1994) Calcium deficiency in the elderly: a factor contributing to the development of hypertension. Eur J Endocrinol 130:433
WHO (2002) World Health Organization guidelines for drinking water quality. Hardness. http://www.who.int/water_sanitation_health/GDWQ/Chemicals/hardnfull.htm
WHO (2009) Calcium and Magnesium in drinking water: public health significance. Cotruvo J, Bartram J (eds) ISBN 978 92 4 156355 0. Geneva, Worls Health organization
Wilkinson SR, Stuedemann JA, Grunes DL, Devine OJ (1987) Relation of soil and plant magnesium to nutrition of animals and man. Magnesium 6:74–90
Yang CY (1998) Calcium and magnesium in drinking water and risk of death from cerebrovascular disease. Stroke 29(2):411–414
Yang CY, Chiu HF (1999) Calcium and magnesium in drinking water and the risk of death from hypertension. Am J Hypertens 12:894–899
Yang CY, Chiu HF, Tsai SS, Wu TN, Chang CC (2002) Magnesium and calcium in drinking water and the risk of death from esophageal cancer. Magnes Res 15(3–4):215–222

Bioavailability of Elements in Soil

15

Brian J. Alloway

Contents

15.1 Introduction

The natural abundance of elements in the soil parent material and the factors controlling their availability to plants and animals provide major links between geology and human medicine and the health of plants and animals in natural and agricultural ecosystems. This is the rock-soil-plant-animal/human pathway, and it is of major importance in the study of medical geology.

The formation of soils (pedogenesis) is closely linked to the weathering of rock-forming minerals and the structural arrangement of organic (mainly humic) and mineral materials to form a soil profile. However, soil formation depends on the involvement of plants, microorganisms, and soil fauna, whereas rock weathering can occur without the biological component and produce a regolith of decomposed rock rather than a soil.

Within the context of soil–plant interactions, the main soil processes controlling the availability of both naturally occurring and contaminant trace and major elements to plants and their leaching down the soil profile to the groundwater are those which influence the sorption and desorption of these elements within the soil. Sorption is the collective term for the retention of metal ions on the surfaces of the solid phase of the soil system. The sorption mechanisms include:

1. Cation or anion exchange in which ions are attracted to oppositely charged sites on the solid surfaces (e.g., negatively charged surfaces in the case of cation exchange and positively charged surfaces in anion exchange)
2. Specific adsorption in which certain metal cations and most anions are held by ligand exchange in the form of covalent bonds
3. Co-precipitation in which ions are precipitated on surfaces simultaneously with other inorganic compounds, such as iron, aluminum, and manganese oxides, or calcium carbonate
4. Insoluble precipitates of elements on surfaces, including the formation of insoluble carbonates, sulfides, phosphates, and hydroxides

B.J. Alloway (✉)
Department of Soil Sciences, The University of Reading, Berkshire, UK
e-mail: b.j.alloway@reading.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_15,

5. Organic complexation with solid-state organic matter ligands (in contrast to the formation of low molecular weight soluble complexes)

Desorption is the term for the release of ions from these sorbed forms due to a change in pH, and redox conditions, or the release of plant root exudates (Alloway 1995).

The factors controlling adsorption and desorption of ions in soils, include:

1. The properties, speciation, and concentration of specific trace and major elements
2. The composition of the soil, especially the relative abundance of clay minerals of different types, and total contents of iron, aluminum and manganese oxides, free calcium carbonate, and organic matter
3. Soil physicochemical conditions, which include pH, redox status, and concentrations of other cations and anions

Plant factors, especially genotype, also play a major role in determining the extent to which elements are accumulated in plant tissues, including edible parts, and, ultimately, in the food chain.

However, elements differ considerably in the relative extent to which they are taken up from the soil and accumulated in plant tissues even allowing for the differences in soil properties and plant genotype. The soil–plant transfer of different elements varies in orders of magnitude from relatively unavailable metals such as barium, to the more readily accumulated elements such as cadmium. In general, the essential trace elements with relatively low transfer coefficients (Tf = M Plant/M Soil) are likely to be more prone to being deficient, whereas both essential and non-essential elements with high transfer coefficients would be more likely to pose a problem of toxicity.

15.2 Soil Formation

Pedogenesis is the process by which a thin surface layer of soil develops on weathered rock material, gradually increases in thickness, and undergoes vertical differentiation in morphology to form a soil profile. The soil profile comprises distinct horizontal layers (called horizons), which differ in color, texture, structure, and organic matter content. This soil profile (or solum) is the unit of classification of soils. The basis of soil mapping is to delineate areas of soils with distinct profile characteristics.

Pedogenesis is essentially the processes of chemical weathering of rock fragments in an environment which is generally rich in atmospheric oxygen, moisture, carbon dioxide, humic material, and biochemicals from living and decomposing vegetation (the biosphere). These pedogenic processes are strongly influenced by climatic factors, which include temperature, precipitation, and evaporation. They are also affected by the drainage of the site at which the soil is forming and this depends on the topography (shedding or receiving water) and the permeability of the layers of weathering rock material.

Humus is the name given to the complex organic molecules formed in the soil as a result of microbial action on dead plant material (litter) in the soil. The solid-state humic substances generally have relatively high molecular weights and are distinctly different from biochemical substances found in living and recently dead plant and animal tissues. Low molecular weight, soluble forms of soil organic matter are referred to as dissolved organic compounds and are important in forming soluble complexes containing metals. These soluble organic compounds can desorb metal ions from sorption sites on solid surfaces and thus increase the mobility and plant availability of many elements. Solid-state humus plays several very important roles in the soil. Physically it contributes to the binding of particles to create soil aggregates and a pore system comprising linked voids of varying diameter which are involved in both water transmission and storage and gaseous exchanges with the atmosphere. Chemically, soil humus is a major reserve of carbon, nitrogen, phosphorus, and sulfur which were constituents of the plant tissues that underwent humification. However, humic substances also have relatively strong adsorptive capacities for cations and the presence of humus in soil therefore adds considerably to the sorptive properties of the soil contributed by clay minerals, oxides of iron, and manganese and carbonate minerals.

Soils are formed as a result of the interactions between the geological parent material, climate, vegetation, and topography (especially with regard to drainage status) over time. Jenny (1941) called these the "state factors" of soil formation and expressed them in the form of an equation:

$$\text{Soil} = f(cl,\ o,\ r,\ p,\ t)$$

where f is a function, cl is climate, o is organisms (vegetation), r is relief (topography), p is parent material, and t is time.

It is important to note that the parent material is the weathered rock at the surface of the Earth and in many cases this may not be the underlying solid geology but the "drift material" that can be fragmented unaltered or chemically weathered rock, which has been transported and deposited on top of the solid geology. This transport may have been by wind (loess), rivers (alluvium), glaciers (glacial sands, gravels, and boulder clay), and downslope movement (colluvium). However, where the solid geology is exposed at the surface the soils will be formed on weathered fragments of this material.

A typical profile of an uncultivated soil comprises a layer of organic litter (desiccated but undecomposed plant material) on the surface (called the L horizon). Below this can occur an A horizon which is a relatively dark colored mixture of humic and mineral material (formed by the actions of earthworms). In cultivated soils where mixing has occurred often over many centuries through plowing and other tillage, the surface layer of dark colored soil is referred to as the plow layer (or Ap horizon). However, under acid conditions in unplowed soils, dark-colored layers of decomposing plant material may exist beneath the litter horizon. These organic layers differ in the extent to which the plant tissues have decomposed and are designated O1 and O2, etc., with O1 only partially decomposed and O2 more decomposed and so on. The O horizons will often overlie an A horizon, but it will not be as thick or deep as in the less acid soils where deep mixing of the soil by large numbers of earthworms or where regular plowing and cultivation has created a thick Ap horizon.

Beneath the A horizon there may be a zone of a lighter color and different texture called an eluvial (E) horizon. This is the zone in the soil profile from which clay minerals and iron have been removed and translocated down the profile. Beneath the E horizon, a B horizon of accumulation can occur. The B horizon can be designated a Bt horizon where clay has accumulated, or a Bs horizon where sesquioxides of iron and aluminum have been deposited and, in podsol profiles, a Bfe and a Bo horizon can also be found. Beneath this B horizon is found the C horizon—the weathering rock material on which the soil is forming. If solid rock is found beneath this C horizon then it is designated the R horizon. Figure 15.1 shows a diagrammatic soil profile with both the FAO/UNESCO and the USDA Soil Taxonomy horizon nomenclature.

There are two important modifying factors that can affect a soil horizon's appearance and chemical properties, and these are identified in the designation of soil horizons. These are where free calcium carbonate occurs in soil horizons (calcimorphic soils) and where soils are affected by either permanent or intermittent waterlogging (hydromorphic soils). In calcimorphic soils (also called calcareous soils, or calcisols), horizons containing visible calcium carbonate are described as calcareous and identified with a lowercase "ca" as in Aca or Bca and so on. This calcium carbonate (calcite) may have originated either from limestone in the parent material or from precipitation of calcite in the pores of soils in semi-arid areas. In hydromorphic soils (usually called gleys) poor drainage results in the onset of reducing conditions which affect both the appearance and the physico-chemical properties of the soil. The waterlogging, which gives rise to the creation of reducing conditions in gley soils can be due to impermeability caused by either fine soil texture (where there is a high proportion of clay-size particles), and/or

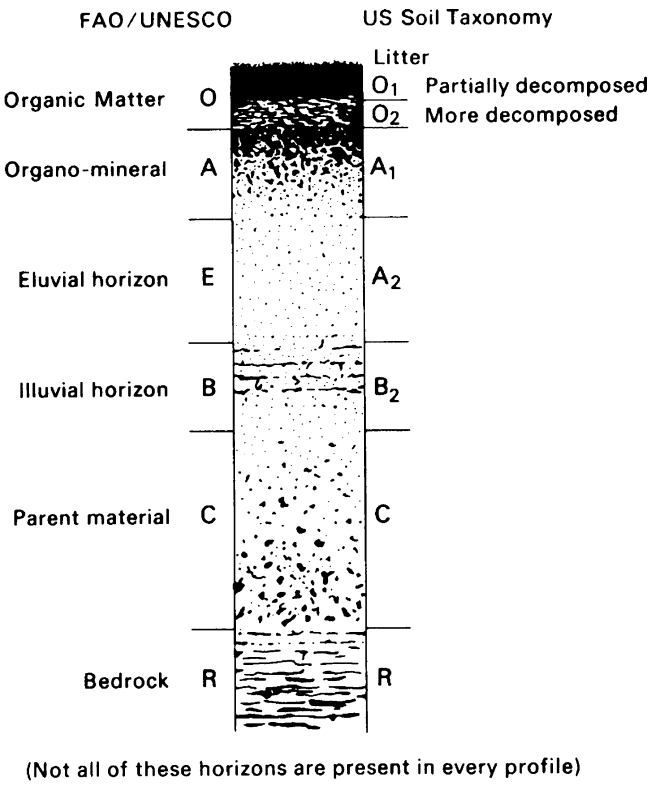


Fig. 15.1 A diagrammatic soil profile with FAO/UNESCO and USDA soil taxonomy horizon nomenclature (From Alloway 1995)

poor structure, or poor site drainage (e.g., in a hollow or at the foot of a slope). Gleying in soils is characterized by pale colors (light brown, gray to bluish green) caused by the reduction of iron oxides. Soils with aerobic, or oxic, conditions normally have darker brown and reddish colors due to the presence of abundant iron oxides.

It should be pointed out that not all soil profiles contain all of the types of distinct horizons mentioned here, but different soil types are characterized by certain combinations of horizons. For example, acidic and strongly leached soils developed on permeable parent material, such as sands or weathered granite, in areas with a high precipitation–evapotranspiration ratio can have a characteristic podsol (or spodosol in the USDA Soil Taxonomy) profile. This podsol comprises a combination of thick organic horizons at the top of the profile—a strongly bleached E horizon, underlain by a Bo horizon of deposited organic matter, over a thin, distinct hard layer of deposited iron (iron pan, Bfe)—which in turn overlies a relatively thick yellowish brown horizon of deposited iron and aluminum sesquioxides (Bs horizon) and this, in turn, overlies the C horizon which is the parent material. This is the distinct profile (L, O1, O2, A, Ea, Bo, Bfe, Bs, C) of a podsol found in many parts of the world with a cool humid climate, such as areas of natural coniferous forest. In contrast, areas underlain by chalk (soft limestone) tend to have a much simpler and shallower profile with a

characteristic deep, dark-colored and humus-rich A horizon underlain by a C horizon of chalk fragments. This AC profile belongs to a shallow calcareous soil called a rendzina found on limestones in central and western Europe. In arid and semi-arid areas, soils with a calcium-carbonate-enriched "calcic" horizon (Bca, or Cca) are commonly found. These calcareous soils are called calcisols and a calcic horizon with more than 15% calcium carbonate (and at least 5% more than in the underlying horizon) is their defining characteristic. Calcisols can develop on limestone parent material or on other parent material in low-lying areas where a high water table with calcium-rich groundwater leads to evaporation from the soil surface and precipitation of calcium carbonate within the pore structure of the soil profile.

The chemical weathering of the geological parent material involves the processes of hydrolysis, hydration, dissolution, oxidation, reduction, ion-exchange, and carbonation. These, together with the physical disintegration of rock, bring about the decomposition of rock-forming minerals, which release cations and anions into solution. New secondary minerals (including clay minerals and iron oxides) can be formed from the products of weathering and precipitates of compounds formed when their solubility products are exceeded. Some of the ions released will remain in solution and be leached down the soil profile. These will either reach the groundwater and move to water courses in humid regions or accumulate in low-lying areas and move up through the soil profile as evapotranspiration of water occurs in arid environments. The latter situation results in the formation of calcified and/or saline soils, which are characteristic of arid and semi-arid regions. Soils with high contents of calcium and a high pH often have low concentrations of plant-available trace elements, such as zinc, and acute deficiencies of this and other essential trace elements can occur in crops growing on them. Low availability of elements to plants can be the result of either low total concentrations in the soil (sometimes referred to as a primary deficiency), to a high degree of sorption leading to low soluble/plant available concentrations, or to antagonistic effects between two or more elements, e.g., copper and zinc, which are both absorbed by the same pathways into plant roots. A relative excess of one element can induce a reduction in the uptake and availability of the other.

The main soil constituents responsible for the adsorption of cations and anions are certain clay minerals, hydrous oxides of iron, manganese and aluminum, calcite, and humus. Therefore, the soil horizons that have relatively high concentrations of these mineral and organic materials will tend to retain ions released from weathering together with ions introduced in fertilizers (especially phosphorus, potassium, and sulfur). Many environmental pollutants (especially trace metals and persistent organic pollutants) will also be retained in these horizons. Nitrogen added to the soil in fertilizers and manures is not retained to any significant extent. It is usually lost from the soil by leaching (mainly as nitrate ions or lost to the atmosphere in gaseous oxide forms [NO_x]). The organic-rich surface horizon, especially in cultivated soils (Ap horizon, topsoil) is usually the horizon with the greatest adsorptive capacity. This horizon is the rooting zone for plants and receives elements cycled through plants and added in fertilizers, manures, and agrichemicals and is also a sink for atmospherically deposited pollutants.

Conversely, horizons of elution where materials, especially clays, have been removed by being washed down the profile, will tend to have lower adsorptive capacities and contents of both nutrient and non-essential elements. In general, sandy textured, well-drained soils tend to have lower adsorptive capacities than soils with a higher proportion of silt and clay-sized mineral particles or iron oxides. However, the humus content can greatly modify this. Intensively cultivated soils generally have lower organic matter contents than soils under permanent grassland or rough grazing in the same climatic zone. Therefore, the texture, organic matter content, and mineral composition of intensively cultivated soils tends to be more important in the dynamics of ions of all types than in grassland soils. In addition to the macronutrients required by plants (nitrogen, phosphorus, potassium, magnesium, calcium, sulfur, and sodium), the trace element content of the parent material will also differ between soils developed on different types of weathered rock. This will include both essential trace elements (boron, cobalt, chlorine chromium, copper, iron, iodine, manganese, molybdenum, nickel, selenium, and zinc) and elements with no known essential function (including arsenic, cadmium, mercury, lead, thallium, and uranium). Therefore, differences in mineralogy and geochemistry of soil parent materials will result in differences in the total concentrations of both major and trace elements found in topsoils and whole soil profiles. Examples of this include soils developed on ultramafic (e.g., serpentinite) rocks where there is a relative excess of magnesium com-pared to calcium and therefore possible problems with a deficiency of calcium and an excess of magnesium and anomalously high concentrations of cobalt, chromium, and nickel. This combination of high magnesium to calcium ratio (Mg:Ca) and elevated concentrations of certain metals has resulted in the evolution of specialized serpentine flora and if non-adapted plants are grown on these soils they could be affected by nutritional imbalances and possibly toxicity.

Of wider importance with regard to areas of land affected are the soils developed directly on clay or shale strata with distinct sedimentary layers (facies) of marine black shales within them and also soils developed on surface drift deposits containing these materials. The marine black shales generally contain anomalously high concentrations of silver, arsenic, cadmium, copper, molybdenum, lead, selenium,

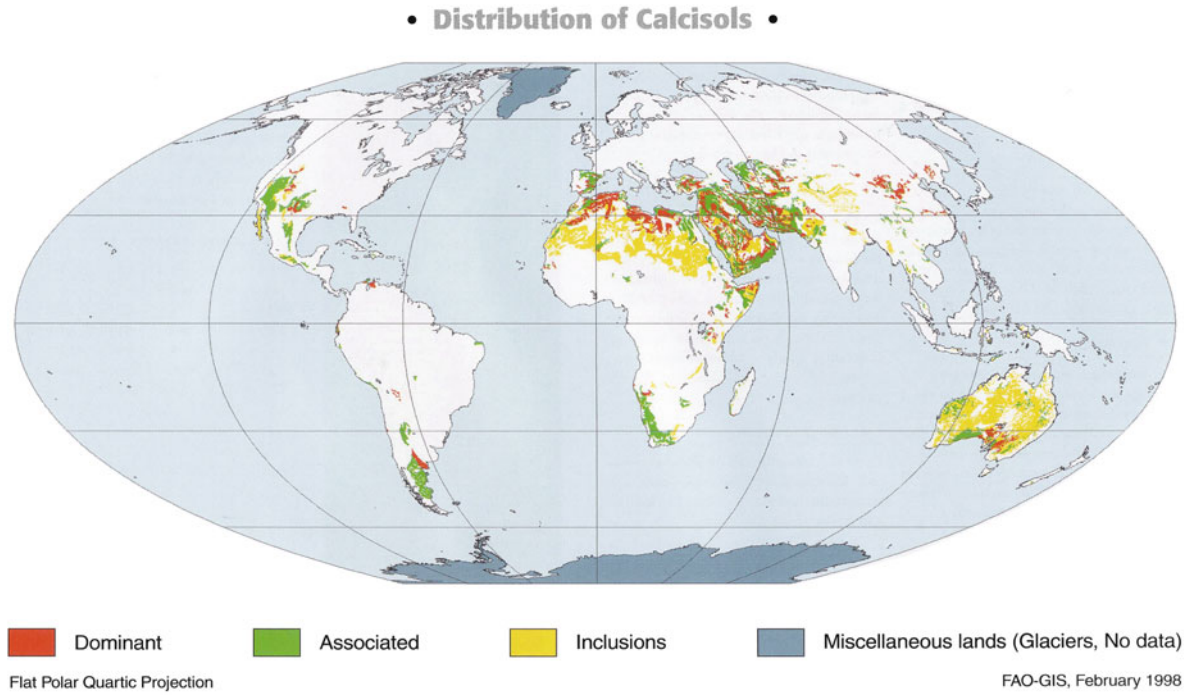


Fig. 15.2 World distribution of calcisols (calcareous soils) (Reproduced from the world reference base for soil resources Atlas, ISSS working group RB 1998, Bridges et al., Eds., by permission of the editors and publisher)

uranium, vanadium, and zinc. There is a distinct possibility of these soils giving rise to elevated concentrations of some or all of these elements in food crops and livestock herbage. Apart from the risk of excess available concentrations of some of these elements, there is also the possibility of antagonistic effects occurring due to interactions between some of the elements present in high concentrations. For example, high molybdenum concentrations can induce copper deficiency in ruminant livestock even though copper itself may be present in elevated total concentrations in the soil and herbage.

In general, soils developed on clay and mudstone formations or drift derived from them tend to have higher concentrations of most essential trace elements than those developed on sandstones and sandy drift deposits. This is due to the clays and mudstones having inherently higher concentrations of these elements and also to the soils having greater adsorptive capacities to retain them against leaching compared with more sandy soils. These geochemical differences, together with variations in soil chemical properties associated with the variations in parent materials and soil forming-factors, will also help to determine the concentrations of elements (both essential and non-essential) that are available to plants (bioavailable fraction). In addition to these geochemical associations, inputs of fertilizers, agrichemicals, and environmental pollutants will also add to the total concentrations, but the nature of the elements concerned and the physicochemical properties of the soil will determine the bioavailability of these substances.

The global distribution of calcisols (ISSS 1998) is shown in Fig. 15.2. These soils are important with regard to the mobility and plant availability of trace elements because their alkaline pH and the presence of free calcium carbonate result in the strong adsorption of many cations. This has serious implications in the case of essential trace elements, which include zinc, iron, copper, and manganese because inadequate concentrations of available forms of these elements can cause deficiencies in both crops and livestock. When crops, such as wheat and rice are affected by deficiencies of essential trace elements, such as zinc, there is both the problem of reduced crop yields and also low concentrations of the elements in the human diet. Therefore, the health of regular consumers could also be affected. Zinc deficiency is a major problem in wheat and other staple crops in arid countries such as Turkey, Syria, Iraq, India, and Pakistan which all have large areas of calcisols.

Soils with a high percentage (>65%) of sand grains are called "arenosols," but in contrast to calcisols, they do not necessarily have one particularly distinctive horizon nor do they always have an alkaline soil pH. They are classified by their distinctive sandy texture, which has certain soil physical and chemical properties associated with it. These sandy soils also give rise to deficiencies of essential trace elements, which include zinc, iron, copper, boron,

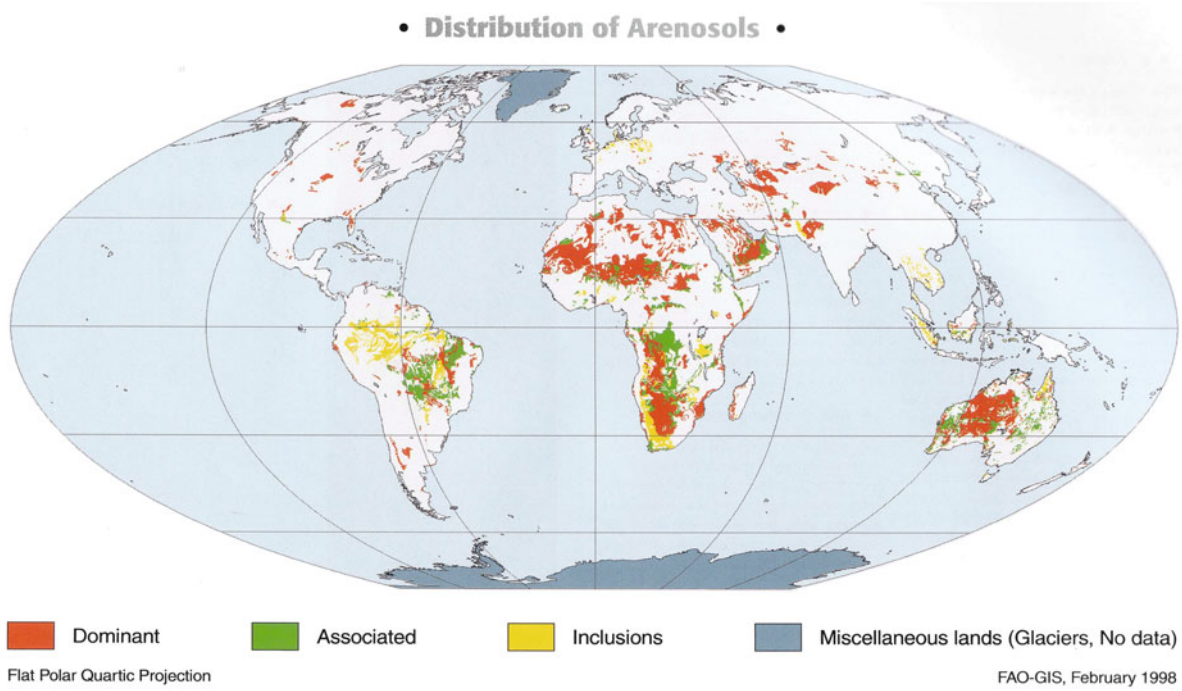


Fig. 15.3 World distribution of arenosols (sandy soils) (Reproduced from the world reference base for soil resources Atlas, ISSS working group RB 1998, Bridges et al., Eds., by permission of the editors and publisher)

manganese, selenium, and cobalt. In this case the cause of the deficiency is the inherently low total concentrations in the sand grains of the soil parent material and the relatively low sorptive capacity of these soils. Their low clay content results in the loss by leaching of nutrient elements reaching the soil by various pathways. Arenosols are commonly found in deserts, but in areas where they are cultivated (usually where the climate is humid or where irrigation is available) deficiencies of essential trace elements are commonly found. The global distribution of arenosols (ISSS 1998) is shown in Fig. 15.3.

Soils rich in iron oxides are characteristic of tropical regions of the world. These oxides are the result of the accumulation of iron and aluminum after all the weatherable minerals have decomposed under the aggressive tropical weathering environment with its high temperatures, high precipitation, and low pH, which favor the chemical weathering of geological minerals. These soils are referred to as oxisols in the USDA Soil Taxonomy. They are also sometimes called ferralitic soils, and they are classified as ferralsols in the *World Reference Base for Soil Resources* (ISSS 1998) and in the FAO/UNESCO World Soil Map. These ferralsols rich in iron oxide are also a potential trace element deficiency problem as a result of both the complete weathering of the original parent material and subsequent leaching and/or the sorption of ions on the iron oxides. The global distribution of ferralsols (ISSS 1998) is shown in Fig. 15.4.

15.3 Soil–Plant Transfer of Trace Elements

The soil–plant transfer coefficient (Tf), also called the bioaccumulation ratio (BR), is a convenient method of expressing the relative ease with which elements in soils (total concentrations) are taken up by plants and accumulated in their above-ground tissues. The coefficient is obtained by dividing the concentration of an element in the plant tissues by the total concentration of the same element in the topsoil (Tf = M in plant/M in soil). Plants differ considerably in their ability to absorb trace elements, but Kabata-Pendias and Pendias (1992) have quoted work that considered the relative ranges of BRs or transfer factors for a large number of plant species from data published in the literature. These range from the least accumulated elements such as barium and titanium to those most readily accumulated (cadmium and boron). However, considerable differences in trace element accumulation occur between plant species and between varieties within a species (inter- and intraspecific variations):

1. Elements lacking accumulation (only slightly available) (Tf < 0.001)
   barium > titanium > scandium > zirconium > bismuth > gallium > iron > selenium
2. Elements showing slight accumulation: (Tf 0.001–0.01)
   antimony > beryllium > chromium > iodine > vanadium > fluorine > lithium > nickel > manganese

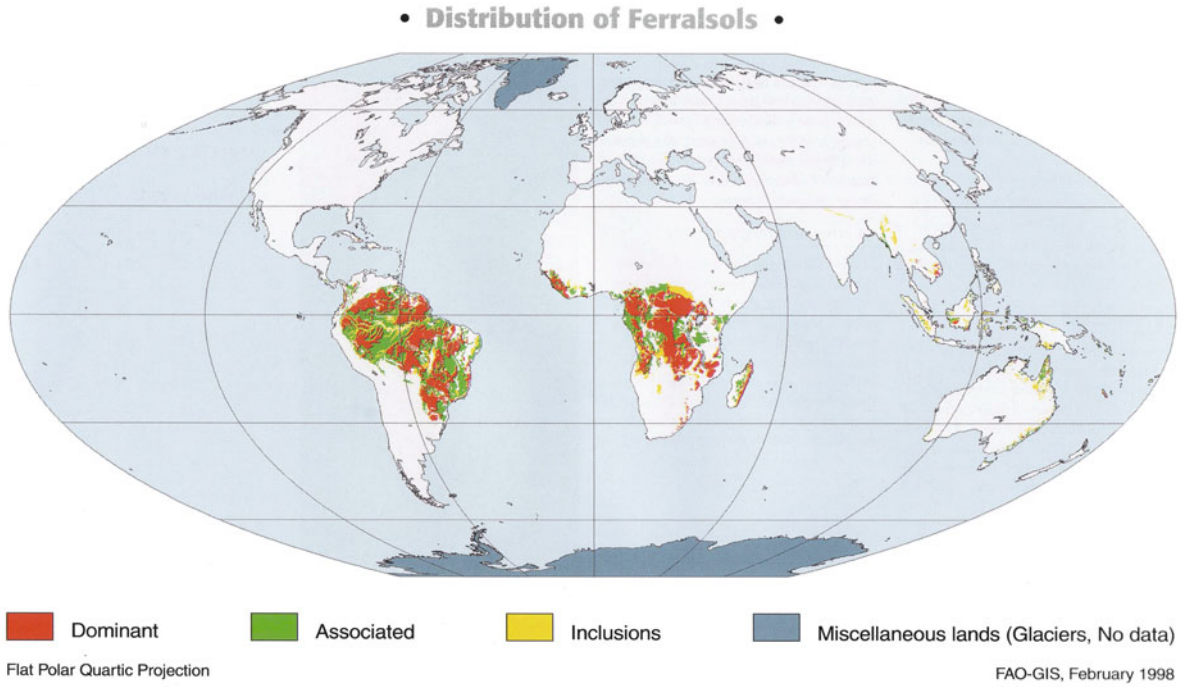


Fig. 15.4 World distribution of ferralsols (iron oxide-rich soils) (Reproduced from the world reference base for soil resources Atlas, ISSS working group RB 1998, Bridges et al., Eds., by permission of the editors and publisher)

3. Elements showing medium accumulation: (Tf 0.01–1.0) cobalt > arsenic > germanium > tellurium > silver > strontium > lead > coppe > mercury > molybdenum > zinc
4. Elements which are intensively accumulated (Tf 1–10) rubidium > caesium > bromine > boron > cadmium

In general, the elements that tend to have relatively low soil–plant transfer coefficients are those that tend to be more strongly sorbed onto the surfaces of soil solids. Those that are more readily accumulated are less strongly adsorbed due to the properties of the ionic forms of these elements. The ions that are strongly sorbed will tend to have lower free ion concentrations in the soil solution and, therefore, be less available for plant uptake and translocation to the aerial tissues of plants.

15.4 Soil Chemical Properties and the Bioavailability of Trace and Major Elements

The soil is a dynamic system comprising solid, liquid, and gaseous phases, which is subject to short-term fluctuations such as variations in moisture status, temperature, pH, and redox conditions. In addition to the non-living mineral and organic constituents, the topsoil contains a highly diverse microbial and mesofauna population and a high density of roots of many different plant species. These living organisms are all intricately involved in the physical, chemical, and biological processes taking place in the soil. In addition to short-term changes, soils also undergo gradual alterations in response to changes in management and environmental factors. Examples of these longer term changes in soil properties include a decrease in the content of organic matter with increasing cultivation and/or increasing temperatures and gradual acidification due to acid precipitation or lack of regular liming in areas of high precipitation relative to evaporation. These short- and long-term changes in soil properties can each have significant effects on the form and bioavailability of trace and major elements (of both indigenous geochemical and external pollution origin). Soils are also inherently heterogeneous at the micro (profile) and macro (field/catchment) scales. For the purposes of soil testing to assess the available concentration of both essential and potentially toxic elements, the spatial variability in soil physical and chemical properties necessitates appropriate and thorough sampling. It is essential to take sufficient samples which include the range of variability in parameters at any site investigated.

15.4.1 Key Soil Chemical Properties

15.4.1.1 Soil pH

The soil reaction (pH) is the most important factor controlling the chemical behavior of ions and many other important

processes in soils. The pH of a soil applies to the hydrogen ion (H^+) concentration in the solution present in the soil pores. This is in dynamic equilibrium with the predominantly negatively charged surfaces of the soil particles with which it is in contact. Positively charged hydrogen ions are strongly attracted to surface negative charges and they have the power to replace most other cations. There is a diffuse layer close to the negatively charged surfaces, which has a higher concentration of H^+ than the bulk solution. When the soil solution is diluted (as when a suspension of soil in water is made to measure pH or in the field after heavy rain), the diffuse layer expands, causing the pH of the bulk solution to increase. This has important implications for soil testing to measure soil pH. It is the normal practice to mix soil with 2.5 times its weight of distilled water. This generally gives a pH value which is 1–1.5 units higher than that of the soil solution in the diffuse layer near the solid surfaces where the reactions take place. However, this dilution effect can be overcome by measuring the pH in a dilute suspension of a neutral salt such as calcium chloride or potassium chloride, which helps to recreate the ionic strength of the real soil solution. When reporting soil pH values, the method used for measuring it is usually given with the results, but if the solvent used is not named, it is normally assumed that the pH was measured in distilled water.

Soil pH is affected by variations in redox potential that periodically occur when the drainage status of soils changes from waterlogged to more freely drained and vice versa. Reducing conditions (gleying) generally cause an increase in pH and oxidation decreases pH. Fluctuations of up to two pH units can occur over a year in soils prone to periodic waterlogging. Oxidation of iron pyrites (FeS_2) in a soil parent material, such as a weathering black shale or drained marshland, can cause a marked decrease in pH due to the formation of sulfate ions and sulfuric acid.

There are several mechanisms in soils that have the effect of buffering pH changes. These include the formation of hydroxy-aluminum ions ($Al(OH)^{2+}$), partial pressure of CO_2, and formation and dissolution of carbonates and cation exchange reactions. However, even with these buffering mechanisms soil pH can vary by as much as one unit diurnally and spatially due to localized conditions in different parts of a field. In humid regions, soil pH usually increases with depth due to the leaching of bases down the soil profile. In arid environments, pH can decrease with increasing depth due to the accumulation of salts of calcium, sodium, and magnesium in the surface horizons through evaporation of the soil solution.

Soils generally have pH values (measured in water) from 4 to 8.5 due to the buffering by aluminum at the lower end and by calcium carbonate at the upper end of the range. Brady (1984) states that the normal pH range of soils in humid regions is 5–7 and 7–9 in the soils of arid regions. The maximum range of pH conditions found in uncontaminated soils is 2–10.5, but some contaminated soils containing large amounts of cement or mortar may have even higher pH values. In a typical temperate environment, such as the UK, the pH range generally found in topsoils is 4–8. The optimum pH for most arable crops is 6.5 and for grassland mineral soils it is 6.0, but for peaty soils, the optimum pH for grassland is 5.5. Soil pH can be relatively easily raised by liming (with calcium carbonate or calcium hydroxide), but it is normally impractical to acidify agricultural soils more alkaline than these values.

In general, most divalent cationic forms of elements are more mobile and more available to plants; therefore, they are less strongly adsorbed on soil solids under acid conditions than they are at neutral and slightly alkaline pH values. However, the pH of the soil also has a major control on the solubility of soil organic matter, which in turn can modify the behavior of trace elements in the soil. As the soil pH increases toward neutrality and beyond, there is an increase in the amount of DOC. This can have a modifying influence on the solubility of several elements which readily form chelate complexes with soluble organic matter. For example, copper and lead are strongly complexed by soil organic matter so they may become more available with increasing soil pH in soils with a relatively high organic matter content due to the increased concentration of organic ligands in the DOC.

The overall importance of the soil pH in controlling the availability/mobility of elements is due to its effects on (1) the solubility of soil organic matter; (2) the speciation and solubility of elements in the soil solution; (3) the polarity of the charge on solids in soils, such as Fe oxides which have variable charges; (4) the cation exchange capacity of soil solids; (5) the dissolution of certain precipitates and minerals (e.g., calcium carbonate); and (6) effects on soil microorganisms and fauna.

15.4.2 Soil Organic Matter

Soil can be distinguished from regolith (decomposed rock) by the presence of living organisms, organic debris (mainly from plants), and humus. All soils (by definition) contain organic matter although the amount and type may vary considerably. Colloidal organic matter has a major influence on soil physical and chemical properties and can be divided into “non-humic” and “humic” substances. The non-humic substances comprise unaltered biochemicals such as amino acids, carbohydrates, organic acids, lignin, lipids, and waxes that have not changed their form since they were synthesized by living organisms. In contrast, humic substances are a series of acidic substances, yellow to black colored, polyelectrolytes of moderately high molecular weight. They are formed by secondary synthesis reactions involving

microorganisms and have characteristics that are dissimilar to any compounds in living organisms. They have several functional groups, which include carboxyl, phenolic hydroxyl, carbonyl, ester, and possibly quinone and methoxy groups (Hayes and Swift 1978; Stevenson 1979). Although mainly composed of humic substances, soil humus also contains some biochemicals bound to humic polymers. The elemental composition of humus (on an ash-free basis) is 44–53% carbon, 3.6–5.4% hydrogen, 1.8–3.6% nitrogen, and 40–47% oxygen. In general, soil organic matter (humus plus non-humic material) contains about 58–60% organic carbon.

The organic matter contents of soils can vary widely from <1% in intensively cultivated arable soils or soils in semi-arid areas, to more than 10% in permanent grassland soils in cool humid conditions. In poorly drained (gleyed) sites, the soil may be formed on peat and comprise more than 70% organic matter. In general, in hotter drier climates organic matter contents are much lower than those found in cooler humid regions. Within the soil profile, organic matter is always found concentrated in the surface horizon. In cultivated soils this will have been mixed within a deeper layer due to plowing (Ap horizon) and in podsols (usually found in cool humid conditions) and in vertisols (found in hot humid conditions) some organic matter will have been translocated down the profile.

Humus is frequently described as comprising three fractions forming a continuum of compounds varying in molecular weight, carbon, oxygen and nitrogen contents, acidity, and cation exchange capacity in the order: humin > humic acid > fulvic acid. The humic acid fraction has a molecular weight in the range of 20,000–100,000 Da and fulvic acid generally consists of lower molecular weight compounds (some of which may be decomposition products of humic acid) with up to 10% polysaccharides. Humins are humic acid type compounds strongly bound onto minerals.

The contribution of organic matter (humic and non-humic) to the chemical properties of a soil are

1. In the adsorption of cations to negatively charged sites (created by deprotonation of carboxyl and phenol groups)
2. In the mobility and protection of some metal ions from adsorption through the formation of soluble complexes (e.g., chelates) with low molecular weight humic substances (DOC)
3. In the retention of many elements in the higher molecular weight, solid forms of humus by chelation

Soil organic matter is the main reservoir of carbon, nitrogen, phosphorus, and sulfur in soils. These can be gradually released as leachable ionic or gaseous forms (not phosphorus) through the action of soil microorganisms. The carbon locked up in the organic matter of the world's soils is a major consideration in model predictions of global climate change.

Table 15.1 Summarized values for pH and organic carbon in soils in England and Wales, Northern Europe (Baltic area), and the United States

Country	Number of samples	Minimum	Median	Maximum
England and Wales[a]				
pH (water)	5,679	3.1	6.0	9.2
Organic C (%)	5,666	0.1	3.6	65.9
Baltic area[b]				
pH (water)	774	3.7	6.1	8.7
Organic C (%)	774	1.8	7.1	97.2
United States[c]				
pH (water)	3,045	3.9	6.1	8.9
Organic C (%)	3,045	0.09	1.05	63.0

[a]Soils collected on 5 × 5 km grid (From McGrath and Loveland 1992)
[b]Soils collected on a 2,500-km grid over an area of 1,800,000 km^2 from 10 European countries surrounding the Baltic Sea (From Reimann et al. 2000)
[c]Soils from sites with healthy crops remote from obvious contamination (Holmgren et al. 1993)

Several trace metals show particularly high affinities for soil organic matter. These include cobalt, copper, mercury, nickel, and lead and these are probably held principally in chelated form within humus (Adriano 2001). Other trace metals, such as cadmium appear to be mainly sorbed in the soil by cation exchange and specific adsorption and are not retained as strongly on solid-state soil organic matter.

Table 15.1 gives the summarized results of large-scale surveys of soils in England and Wales, a large area of northern Europe surrounding the Baltic Sea, and the United States for soil pH (measured in water) and percentage organic carbon. The surveys differed in that the samples from England and Wales were collected on a formal 5 × 5 km grid, the Baltic area on a grid of one sample per 2,500 km^2, and the American samples from sites selected as free from obvious contamination. The pH data are remarkably similar but the organic matter data show a much lower median value for the American samples, which is probably a reflection of the hotter climate in many agricultural areas of the United States. In contrast, northern Europe has a cooler, more humid climate which is reflected in more peaty soils and pastures with relatively high organic matter contents.

15.4.3 Chemically Active Mineral Constituents

The inorganic constituents of soils usually comprise more than 90% of the mass of soils, and it is the adsorption and desorption of ions on the surfaces of these materials that exerts an important effect on the plant availability and mobility of macro elements and trace elements. The inorganic fraction can comprise a wide range of rock fragments and minerals undergoing weathering; newly synthesized and recycled clay minerals; oxides of iron, aluminum, and

manganese; free carbonates of calcium and magnesium; and, in more arid regions, crystals of salts such as calcium sulfate and sodium chloride.

15.4.3.1 Clay Minerals

Clay minerals are either products of rock weathering or are synthesized as new minerals from the products of weathering. They have marked effects on both the physical and chemical properties of soils. Their contribution to soil chemical properties results from their comparatively large surface area and permanent negative charge on their surfaces which adsorb cations. The clay fraction of a soil is defined as the mass of the dispersed inorganic constituents which are less than 2 μm in diameter. Although this is based on particle size rather than mineralogy, in most cases it is the mineralogically distinct group of clay minerals that comprise most of the material in this size fraction (together with iron oxides in many cases). Due to space limitations in this book, it is not possible to cover the crystallography of these phyllosilicate minerals; however, they all share two main types of building blocks in their structure. These are a continuous sheet of silicon (Si) oxygen (O) tetrahedra (the silica unit) and another of aluminum (Al) hydroxide (OH) octahedra (the gibbsite unit). In many cases some of the silicon and aluminum ions in the crystal lattice of the minerals may be replaced by other ions.

The most common types of clay mineral include: (1) the kaolinites with one silica sheet and one gibbsite sheet (a 1:1 clay); (2) the illites which contain two silica sheets with one gibbsite sheet between them (a 2:1 clay); (3) the smectites which also have two silica sheets and one gibbsite sheet (2:1 clay); and (4) the vermiculites which have two silica sheets, one gibbsite sheet, and one brucite sheet containing magnesium, which is not found in the other types of clay minerals (2:2 clay). In all the clay minerals except kaolinite, isomorphous substitution within the mineral lattice leads to a permanent charge imbalance which gives rise to a net negative charge on the surface of the mineral. For example, this can be caused by Al^{3+} substituting for Si^{4+}, and Mg^{2+} or Fe^{2+} substituting for Al^{3+}.

In kaolinites the 1:1 units are tightly bound together by hydrogen bonds between hydrogen and oxygen atoms of adjacent layers. These kaolinites have a smaller surface area than the other clay minerals (5–40 $m^2\ g^{-1}$) and their cation exchange capacity is relatively low (3–20 $cmols_c\ kg^{-1}$) because little isomorphous substitution has occurred.

Illites have their 2:1 units bonded by potassium ions and their specific surface and cation exchange capacity are larger than those of kaolinites (100–200 $m^2\ g^{-1}$ and 10–40 $cmols_c\ kg^{-1}$, respectively). Smectites have the largest specific surfaces (700–800 $m^2\ g^{-1}$) due to relatively weak interlayer bonding, which allows them to expand when they are wetted, and consequently they have high cation exchange capacities (80–120 $cmols_c\ kg^{-1}$). They shrink on drying and can give rise to cracks during prolonged dry periods in soils in which they predominate. Vermiculites have an intermediate surface area and a high cation exchange capacity (100–150 $cmols_c\ kg^{-1}$).

15.4.3.2 Oxides of Iron, Manganese, and Aluminum

Oxides of iron, manganese, and aluminum in soils are often referred to as the hydrous oxides, or in the case of iron and aluminum, as sesquioxides. They play important roles in the chemical properties of soils. In temperate regions they generally occur in the clay size fraction (<2 μm) mixed with the clay minerals and have a disordered structure. The main iron oxides include gelatinous ferrihydrite ($Fe_2O_3.9H_2O$), goethite (FeOOH), hematite (α-Fe_2O_3), and lepidocrocite (γ-FeOOH). Precipitation of Fe^{3+} is initially in the form of ferrihydrite, which gradually dehydrates with aging to form more stable minerals. However, ferrihydrite is more likely to be subsequently dissolved again than the other iron oxide minerals when a decrease in redox potential (Eh) or pH occurs. These oxides generally occur as mixtures and are often referred to generically as "iron oxides" or "hydrous iron oxides," but goethite is the iron oxide mineral most frequently found in soils. In tropical regions where there is a more aggressive chemical weathering regime and the soils are usually much older (10^4–10^6 years), the oxides of iron and aluminum are often the predominant soil minerals. This is because all the primary (rock-forming) minerals and most of the clay minerals will have been chemically decomposed (weathered). The characteristic brown color of most soils throughout the world is due to the presence of iron oxides. Nevertheless, the colors of soils vary widely depending on the amounts of iron in the parent material, the presence of other strongly colored minerals, the drainage (redox) status, and the organic matter and calcium carbonate contents.

Gibbsite is the common form of aluminum oxide found in soils, but it is much less abundant than iron oxides. Manganese oxide minerals are generally present in smaller quantities than iron oxides but have stronger adsorptive properties for several cations.

Freshly deposited soil oxides are the most active in adsorbing and co-precipitating trace and major elements. As a result of its large surface area, freshly deposited ferrihydrite acts as a scavenger, sorbing both cations and anions, especially phosphate and arsenate ions (HPO_4^{2-}, H_2PO_4, and AsO_4^{3-}). If oxidizing conditions persist for a long time and/or temperatures remain high, the oxide crystals age and become dehydrated and less strongly charged. Therefore retention of metal ions by oxide surfaces is inversely related to the degree of crystallinity of the oxide minerals (Okazaki et al. 1986).

The adsorptive properties of iron and manganese oxides depend on the soil pH, which determines whether they are positively or negatively charged. Generally speaking, their charge is negative under neutral-alkaline conditions and positive under acid conditions. The pH at which the charge is neutral (called the point of zero charge, PZC) varies for the pure forms of different oxides. The PZC for iron oxides lies in the region pH 7–10, for gibbsite it is in the pH range 8–9.4, and for manganese oxides the pH range is 1.5–4.6. However, when mixed with clay minerals in soils, the PZC values tend to be much lower than these. In acid soils, the positively charged iron oxides are the main adsorptive medium for soil phosphate and arsenic anions. Soil organic matter also has a pH-dependent charge, but its PZC is around pH 2 which is not normally encountered in soils, so this material is nearly always negatively charged. It is, therefore, an important contributor to the cation exchange capacity of a soil.

The presence of chemically active forms of metal oxides in soils is very dependent on the drainage status of the soil at any site. Where waterlogging occurs either as a result of an impermeable soil mineral fraction, such as a high proportion of swelling clays, or topographic position, as in a depression or receiving site, then reducing conditions will predominate. In soils with developing reducing conditions (called gleying), oxides will be reduced and the iron and manganese ions will be mobilized together with the ions of other elements that had been adsorbed on their surfaces. Permanently waterlogged (gleyed) soils have low concentrations of iron and manganese because the oxides of these elements will have been dissolved, transported away, and deposited in an oxygenated environment over a long period of time. Fluctuating redox conditions (such as in paddy rice soils) can result in the periodic mobilization and precipitation of oxides. Certain specialized bacteria, including *Thiobacillus ferrooxidans* and *Metallogenum* spp. are also involved in the precipitation of iron and manganese oxides, respectively.

15.4.3.3 Free Carbonates

The predominant carbonate mineral found in soils is calcite ($CaCO_3$) and more rarely dolomite (Ca, $Mg(CO_3)_2$). These minerals can be present as a result of soils forming on weathering limestone rocks and, in addition, in arid regions from the accumulation of calcite in the pores of soils due to the evaporation of water vapor from calcium-rich ground waters. This latter process is referred to as calcification.

The presence of free carbonate in soils has such a dominant effect on the soil's morphology and chemical properties that they are specially classified as members of the "calcimorphic" group of soils. These include the calcisols (Fig. 15.2) and other soils with relatively high calcium carbonate contents, such as rendzinas, which do not have a horizon of distinct calcium carbonate enrichment within their profile. The carbonate minerals control the soil pH and tend to keep it between 7 and 8.5, and they bind a wide range of cations and certain anions to their surface. Major elements such as phosphorus are strongly bound forming apatite minerals (calcium phosphates) and many trace elements, which include cadmium and zinc are sorbed and rendered less available and less mobile. The agricultural practice of liming soils, where either calcium carbonate or calcium hydroxide is added to soils, has the effect of converting the soil to a near neutral or calcareous soil with a significant proportion of free carbonate surface area. Liming is frequently carried out to remediate soils contaminated with trace metals which include cadmium, copper, lead, and zinc. In naturally calcareous soils, the plant-available concentrations of essential trace elements, including boron, iron, copper, manganese, and zinc, are often inadequate for many agricultural crops. These crops are found to suffer from deficiencies of essential trace elements which can cause major reductions in yield and sometimes crop failure unless appropriate prophylaxis is carried out. This usually involves applying salts of the elements, such as zinc sulfate, copper sulfate, manganese sulfate, and sodium borate to the soil, or other forms, including chelates, as foliar applications to the crops themselves.

15.4.4 Redox Conditions

The balance of reducing and oxidizing conditions (redox status) in the soil is important due to its effects on the speciation of several very important elements including: carbon, nitrogen, sulfur, iron, manganese, chromium, copper, arsenic, silver, mercury, and lead. All of these elements can exist in soils in more than one oxidation state. The main factor determining the redox status is the degree of waterlogging which prevents the movement of oxygen through the soil profile in the larger diameter, air-filled pores. Redox equilibria are controlled by the aqueous free electron activity which can be expressed as either a pE value (negative log of the electron activity) or as a redox potential, Eh, (the millivolt difference in potential between a Pt electrode and a standard H electrode). Large positive values of pE or Eh (+300 to +800 mV) indicate the presence of oxidized species and low or negative Eh values (+118 to −414 mV) are associated with reducing conditions. Eh can be converted to pE by the factor: Eh (mV) = 59.2 pE (Lindsay 1979).

Redox conditions control the precipitation and dissolution of iron and manganese oxides and, therefore, have a direct influence on the soil's adsorptive capacity for anions and cations. In gleyed (periodically or permanently waterlogged) soils, those elements which are normally sorbed to iron oxides tend to be more bioavailable than in freely

drained (oxic) soils. These include cobalt, iron, nickel, vanadium, copper, and manganese. However, boron, cobalt, molybdenum, and zinc do not undergo changes in valency themselves when changing redox conditions but are sorbed or desorbed and co-precipitated or dissolved according to the effects of redox on the iron and manganese oxides.

Oxic soils tend to have red and brown colors, whereas gleyed soils with strong reducing conditions tend to have pale colors, such as grayish brown and gray to blue-green colors determined by the presence of ferrous ions (Fe^{2+}). However, strong colored parent materials may mask the color changes to a certain extent. Under severely reducing conditions, several elements may be precipitated as sulfides which renders them insoluble and therefore unavailable to plants. The sulfide ion (S^{2-}) comes from the reduction of sulfate (SO_4^{2-}) or from the degradation of sulfur-containing compounds in soil organic matter. In general, the soil pH in acid soils increases slightly under reducing conditions, but in alkaline soils there is a slight pH decrease. If there are cyclic reducing-oxidizing conditions, such as occur in a rice paddy soil with periodic flooding and drying, sulfides, such as iron pyrites (FeS_2) and cadmium sulfide (CdS) (if the soil has been contaminated by this element) will become oxidized and the metal ions released. However, the formation of sulfate ions (SO_4^{2-}) from the oxidation of sulfide when a waterlogged soil dries out results in the soils becoming strongly acidified and elements such as cadmium rendered highly mobile and bioavailable to plants. This situation was responsible for the heavily contaminated paddy soils of the Jinzu Valley in Japan, which caused the outbreak of itai-itai disease in women who had subsisted on locally grown rice (see Sect. 15.7.3).

15.4.5 Adsorption and Desorption of Ions in Soils

The surfaces of the organic and mineral colloidal solids in soils are able to retain ions by several different mechanisms. Frequently, this retention is referred to by the general term "sorption" because in addition to true adsorption by attraction forces, some ions may reside on surfaces as a result of the formation of insoluble precipitates and by chelation.

15.4.5.1 Cation and Anion Exchange

Cation exchange is the term applied to the electrostatic attraction of positively charged cations to negatively charged surfaces. This occurs on several clay minerals, organic matter, and Fe oxides at higher pH values. It is the formation of outer sphere complexes with the surface functional groups to which they are bound electrostatically. The "exchange" part of this mechanism is due to the exchange between counter ions in the soil solution near the charged surface. The ability of an adsorbent (mixtures of colloidal clays, oxides, and organic matter) to attract and retain cations is refer-red to as the cation exchange capacity (CEC) and its units of measurement are centimoles of charge per kg ($cmols_c\ kg^{-1}$). The CEC of mineral soils can range from around 3 to 60 $cmols_c\ kg^{-1}$, but in organic soils this can rise to 200 $cmols_c\ kg^{-1}$. The importance of pH and PZC in determining the charge of oxides was discussed above. In general, oxides contribute little to the CEC of soils below pH 7 although they can be involved in specific adsorption reactions in acid soils.

Anion exchange occurs where negatively charged anions, such as Cl^-, SO_4^-, and NO_3^- are attracted to positively charged sites on soil solids. These are usually variable charge sites at pH values below the PZC. Soil in which positively charged surfaces predominate, such as ferralsol, at low pH will have the ability to retain various anions against leaching. Maximum anion exchange capacities (AEC) for sesquioxides (iron and aluminum oxides) are 30–50 $cmol_c\ kg^{-1}$ (White 1997). The exchange of Cl^- and NO_3^- is straightforward, but the exchange of other anions including sulfates, phosphates, and molybdates is more complicated due to specific reactions between these anions and the adsorbent. The CEC of soils is generally greater than their AEC due to the greater number of negative charges on the colloid surfaces.

Table 15.2 gives typical CEC values for colloidal constituents of soils and shows that soil organic matter has a much higher CEC than all other soil constituents except vermiculite clays aluminum.

Cation exchange has the following characteristics: it is reversible, diffusion controlled, stoichiometric, and there is some degree of selectivity for one ion over another by the adsorbing surface. This selectivity gives rise to an order of replacement determined by the concentration of ions, their valency, their degree of hydration, and hydrated radius. The higher the valency of an ion, the greater its replacing power. However, the only exception to this is H^+ which behaves like a polyvalent ion. The greater the degree of hydration, the lower the replacing power of an ion, other things equal. The commonly quoted order of replaceability on the cation

Table 15.2 Typical cation exchange capacity values for soil constituents

Soil constituent	CEC ($cmols_c kg^{-1}$)
Soil organic matter	150–300
Kaolinite (clay)	2–5
Illite (clay)	15–40
Montmorillonite (clay)	80–100
Vermiculite (clay)	150
Hydrous oxides of iron, manganese and aluminum	4

From Ross (1989)

Table 15.3 Typical orders of replacement of trace element cations on various soil constituents

Soil constituent	Selectivity order	Ref.
Montmorillonite	Ca > Pb > Cu > Mg > Cd > Zn	Bittel and Miller (1974)
Illite	Pb > Cu > Zn > Ca > Cd > Mg	Bittel and Miller (1974)
Kaolinite	Pb > Ca > Cu > Mg > Zn > Cd	Bittel and Miller (1974)
Smectite, vermiculite and kaolinite	Zn > Mn > Cd > Hg	Stuanes (1976)
(ferrihydrite)	Pb > Cu > Zn > Ni > Co > Sr > Mg	Kinniburgh et al. (1976)
Fe oxides- (hematite)	Pb > Cu > Zn > Co > Ni	MaKenzie (1980)
(goethite)	Cu > Pb > Zn > Co > Cd	Forbes et al. (1976)
Peat	Pb > Cu > Cd = Zn > Ca	Bunzl et al. (1976)
Fulvic acid	Fe^{3+} > Cu > Zn > Mn > Ca > Mg	Murray and Lindler (1983)
Humic substances	Cu > Pb > Zn = Ni > Co > Cd > Mn > Ca > Mg	Tipping and Hurley (1992)

exchange complex (comprising colloidal organic matter, clay, and oxides) is:

$$\begin{aligned}\text{Lithium } (Li^+) &= \text{sodium } (Na^+) > \text{potassium } (K^+)\\ &= \text{ammonium } (NH_4{}^+) > \text{rubidium } (Rb^+)\\ &> \text{caesium } (Cs^+) > \text{magnesium } (Mg^{2+})\\ &> \text{calcium } (Ca^{2+}) > \text{strontium}\\ (Sr^{2+}) &= \text{barium } (Ba^{2+}) > \text{lanthanum } (La^{3+})\\ &= \text{hydrogen } (H^+) = \text{aluminium } (Al^{3+})\\ &> \text{thorium } (Th^{4+})\end{aligned}$$

Examples of replacing power of different trace element ions on specific soil constituents are shown in Table 15.3.

From Table 15.4 it can be seen that metal ions such as cadmium and zinc, which were shown in Sect. 15.3 to have relatively high soil–plant Tfs, tend to have low replacing powers and are therefore not strongly retained on soil surfaces. These ions are therefore more readily available for uptake by plants and are also more easily leached down the soil profile than ions with higher replacing powers such as lead and copper.

15.4.5.2 Specific Adsorption

This mechanism involves the exchange of cations of several elements and most anions with surface ligands on solids to form partly covalent bonds with lattice ions. This mechanism is highly pH-dependent and is related to the hydrolysis of the sorbed ions. The pK (equilibrium constant) values of the reaction $M^{2+} + H_2O = MOH^+ + H^+$ determine the adsorption behavior of different ions. Specific adsorption increases with decreasing pK value. However, where the pK values are the same, the ion with the greater radius will be the more strongly adsorbed. Brummer (1986) gave the order for increasing specific adsorption as:

$$\begin{aligned}&\text{Cd } (pK = 10.1) < \text{Ni } (pK = 9.9) < \text{Co } (pK = 9.7) < \text{Zn}\\ &(pK = 9.0) << \text{Cu } (pK = 7.7) < \text{Pb } (pK = 7.7) < \text{Hg } (pK = 3.4).\end{aligned}$$

Those ions retained by specific adsorption are held much more strongly than they would be by cation exchange and the CEC of the soil constituent may not reflect the extent of sorption by this different mechanism. For example, it has been shown that the sorptive capacities of iron and aluminum oxides were between 7 and 26 times greater than their CECs at pH 7.6 (Brummer 1986).

In addition to sorption on colloid surfaces, some ions can diffuse into minerals, such as iron and manganese oxides, illite and smectite clays, and calcite. The rate of diffusion into the minerals increases with pH up to a maximum, which is equal to the pK value for when $M^{2+} + H_2O = MOH^+ + H^+$ on the mineral surface. Above this pH, the $MOH^+ > M^{2+}$ and diffusion rate decreases. This can be related to the ionic radius of the ions involved. For example, the maximum relative diffusion rates for cadmium, nickel, and zinc decrease in the order: Ni > Zn > Cd where the ionic radii are Ni 0.69 nm, Zn 0.74 nm, and Cd 0.97 nm (Brummer 1986).

15.4.5.3 Co-precipitation

Co-precipitation is defined as the simultaneous precipitation of a chemical in conjunction with other elements by any mechanism and at any rate (Sposito 1983). The types of mixed solids formed include clay minerals, iron and manganese oxides, and calcite in which isomorphous substitution has occurred. In addition to co-precipitation, replacement of Ca^{2+} in $CaCO_3$ by other elements can occur. For example, cadmium can diffuse into $CaCO_3$ and form cadmium carbonate ($CdCO_3$) (Papadopoulos and Rowell 1988). Typical co-precipitated elements in different minerals are shown in Table 15.4.

15.4.5.4 Insoluble Precipitates of Elements in Soils

When the concentrations of cations and anions in the soil solution exceed the solubility products of compounds they can form, then insoluble precipitates of these may be formed that will have an important effect in controlling

Table 15.4 Trace metals normally found co-precipitated with secondary minerals in soils

Mineral	Co-precipitated trace metals
Iron oxides	V, Mn, Ni, Cu, Zn, Mo
Manganese oxides	Fe, Co, Ni, Zn, Pb
Calcite	V, Mn, Fe, Co, Cd
Clay minerals	V, Ni, Co, Cr, Zn, Cu, Pb, Ti, Mn, Fe

From Sposito (1983)

the concentrations of ions in solution in addition to electrostatic adsorption reactions. Examples of some compounds that can be formed and occur in the solid state in soils include:

1. Phosphates of calcium ($Ca_{10}(OH)_2(PO_4)_6$), cadmium ($Cd(PO_4)_3$), and lead ($Pb_5(PO_4)_3Cl$)
2. Carbonates and bicarbonates of calcium ($CaCO_3$, $Ca(HCO_3)_2$), magnesium ($MgCO_3$), sodium (Na_2CO_3), cadmium ($CdCO_3$), and zinc ($ZnCO_3$)
3. Sulfides of iron (FeS_2), cadmium (CdS), and mercury (HgS, Hg_2S)
4. Chlorides of sodium (NaCl) and mercury ($HgCl_2$)
5. Iron (ferrite) forms of copper ($Cu_2Fe_2O_4$), molybdenum ($Fe_2(MoO_4)_3$), and zinc ($ZnFe_2O_4$)

15.4.5.5 Organic Complexation

The solid-phase humic material is involved in the retention of trace elements by forming complexes, such as chelates, in addition to comprising part of the colloidal cation exchange complex. Humic substances with reactive groups, which include hydroxyl, phenoxyl, and carboxyl groups, form coordination complexes with metal and other ions. Carboxyl groups are particularly important in binding by the humic and fulvic acid fractions of humus. The stability constants of chelates with elements tend to be in the following order:

Soluble, low molecular weight organic compounds (DOC) of both humic and non-humic origin can form soluble complexes with many trace elements and thus prevent them from being sorbed onto solid surfaces. This has the effect of making these elements more readily leached down the soil profile and/or more available for plant uptake.

15.4.5.6 Quantitative Description of the Sorption of Ions in Soils

A great deal of research has been conducted on the adsorption of elements by soils, especially trace metals that can pose a potential risk of toxicity in plants, animals, and humans. It has been found that two different adsorption equations are very useful in describing most of the adsorption measured. These are the Freundlich and Langmuir equations.

The Freundlich equation is expressed as : $x/m = \mathrm{KC}^{1/n}$

where x/m is the amount of solute adsorbed per unit mass, C is the concentration of solute in solution at equilibrium, and K and n are equation constants. This equation provides an effective means of summarizing adsorption that follows a hyperbolic relationship with the greatest amount occurring at lower concentrations and gradually decreasing with higher concentrations.

The Langmuir equation is expressed as:

$$\frac{C}{x/m} = \frac{1}{Kb} + \frac{C}{b}$$

where C is the concentration of the ion in the equilibrium solution, x/m is the amount of C adsorbed per unit mass, K is a constant related to the bonding energy, and b is the maximum amount of ions that will be adsorbed by a given sorbent.

These and other applicable adsorption isotherm equations can then be used in models of trace element availability to plants.

15.5 Concentrations of Selected Trace and Major Elements in Rocks, Soils, and Crop Plants

The concentrations of macro and trace elements in soils vary widely as a result of differences in the mineralogy of the soil parent material and, in the case of trace elements, the amount of contamination from external sources can also vary widely. The ranges of concentrations of both macro and trace elements given in Table 15.5 include soils from around the world developed on highly diverse parent materials that have been subject to varying degrees of contamination. Almost all soils in the technologically advanced regions of the world and also in many parts of developing countries are subject to a certain amount of contamination, often from atmospheric deposition.

The data for elements in soils presented in Table 15.5 are for "normal" agricultural soils not considered to be markedly contaminated. Total concentrations of trace elements, such as lead, copper, and zinc can reach very high concentrations in some overtly contaminated soils such as those near metalliferous mines and smelters, other industrial sources of atmospheric emissions, and on land that has received heavy applications of sewage sludge. The concentrations of metals including cadmium, copper, lead, and zinc in sewage sludges have decreased markedly in most industrialized countries as a result of strict pollution controls and structural changes in industry. Nevertheless, soils that received heavy, repeated applications of sewage

Table 15.5 Concentrations of selected trace and major elements in the lithosphere, agricultural soils, and food crops (mgkg^{-1} dry matter)

Element	Content in lithosphere	Common range for agricultural soils	Selected average for soils	Typical range in food crops
Silver (Ag)	0.10	0.03–0.9	0.05	0.03–2.9
Arsenic (As)	5	<1–95	5.8	0.009–1.5
Barium (Ba)	430	19–2,368	500	1–198
Boron (B)	10	1–467	9.5–85	1.3–16
Calcium (Ca)	36,000	7,000–500,000	13,700	1,000–50,000
Cadmium (Cd)	0.2	0.01–2.5	0.06–1.1	0.13–0.28
Cobalt (Co)	40	0.1–70	7.9	8–100
Chromium (Cr)	200	1.4–1,300	54	0.013–4.2
Copper (Cu)	70	1–205	13–24	1–10
Fluorine (F)	625	10–1,360	329	0.2–28.3
Iron (Fe)	51,000	5,000–50,000	38,000	25–130
Mercury (Hg)	0.1	0.05–0.3	0.03	0.0026–0.086
Iodine (I)	0.3	0.1–10	2.8	0.005–10.4
Potassium (K)	26,000	400–30,000	8,300	20,000–50,000
Magnesium (Mg)	21,000	20–10,000	5,000	1,500–3,500
Manganese (Mn)	900	270–525	437	15–133
Molybdenum (Mo)	2.3	0.013–17	1.8	0.07–1.75
Sodium (Na)	28,000	750–7,500	6,300	–
Nickel (Ni)	100	0.2–450	20	0.3–3.8
Phosphorus (P)	1,200	200–5,000	600	3,000–5,000
Lead (Pb)	16	3–189	32	0.05–3.0
Sulfur (S)	600	30–10,000	700	1,000–5,000
Selenium (Se)	0.09	0.005–3.5	0.33	0.001–18.0
Tin (Sn)	40	1–11	–	0.2–7.9
Titanium (Ti)	6,000	1,000–9,000	3,500	0.15–80
Vanadium (V)	150	18–115	58	0.5–280
Zinc (Zn)	80	17–125	64	1.2–73.0

Compiled from Lindsay (1979), Kabata-Pendias and Pendias (1992), Adriano (2001), and Marschner (1995)
Note: The typical plant concentrations for the macronutrients calcium, potassium, phosphorus, and sulfur are those for optimum growth and not the full range that may be actually found in crops around the world

sludge in earlier periods when metal contents in sludges were much higher are still likely to retain relatively high total concentrations of several metals for many years (see also Chap. 4, this volume).

Heavily contaminated soils can sometimes contain hundreds of mg kg^{-1} of cadmium, when the safe maximum limits are considered to be around 1–3 mg kg^{-1}; hundreds or thousands of mg kg^{-1} of lead and zinc when the safe maximum values for human health is considered to be in the range 125–450 mg kg^{-1} for lead, and the safe maximum for plants for zinc is around 200–300 mg kg^{-1}. However, in most cases the sites with very high concentrations of metals arising from contamination are generally of relatively small extent in comparison with the total area of agricultural land. Perhaps the most insidious contamination problem is with inputs of cadmium in phosphatic fertilizers which are used in most parts of the world, at least where high-yielding crops are grown. Some of the phosphate rock used for making these fertilizers can contain relatively high concentrations of cadmium (<100 mg kg^{-1}).

15.6 The Bioavailability of Trace and Major Elements to Plants

Uptake of ions by roots can involve several processes including cation exchange by roots, transport inside cells by chelating agents and other carriers, and rhizosphere effects.

Uptake of ions by roots is controlled by the release of ions and organic compounds which include amino acids (e.g., aspartic, glutamic, and prolinic acids). These exudates vary with plant species, microorganism association, and plant growth conditions (e.g., supply of essential trace elements).

15.6.1 Uptake of Trace Elements by Plants

The uptake of trace elements by plants is a key stage in the soil-plant-animal/human pathway and is second only to intake via drinking water with regard to the link between geochemistry and human health apart from where there is excessive ingestion of soil.

Plants readily take up ionic or soluble complexed forms of trace elements present in the soil solution. The factors affecting the amounts of elements absorbed through the roots are those controlling: (1) the concentration and speciation of the element in the soil solution, (2) movement of the element from the bulk soil to the root surface, (3) transport of the element from the root surface into the root, and (4) its translocation from the root to the shoot. Absorption of mobile ions present in the soil solution is mainly determined by the total quantity of this ion in the soil. However, in the case of strongly adsorbed ions, absorption into the root is more dependent on the amount of root produced and its ability to explore a large volume of soil. Mycorrhizae are symbiotic fungi which effectively increase the absorptive area of the root and can assist in the uptake of nutrient ions such as orthophosphates and micronutrients when concentrations of these are low. Roots also possess a significant CEC, mainly due to the presence of carboxyl groups, and this probably forms part of the mechanism transporting ions through the outer part of the root to the plasmalemma where active absorption takes place. Evapotranspiration (flow of water into roots, up through the plant and out of leaf canopy) is an important factor controlling root uptake of elements because plants that have more rapid evapotranspiration will absorb greater volumes of soil solution containing ions and complexes in solution.

Absorption of ions by plant roots can be both passive and metabolically active processes. Passive uptake involves diffusion of ions in the soil solution to the root endodermis. In contrast, metabolically active absorption takes place against a diffusion gradient but requires the expenditure of energy so it is vulnerable to inhibition by toxins. The type of mechanism appears to differ with different elements; for example, lead is passive whereas the uptake of copper, molybdenum, and zinc is considered to be metabolically active or a combination of both active and passive mechanisms. Ions which are absorbed into the root by the same type of mechanism are likely to have an antagonistic relationship through competition with each other. For example, zinc absorption is inhibited by copper and hydrogen, but not by manganese and iron. Copper absorption is inhibited by zinc, ammonium ions, calcium, and potassium (Barber 1984; Graham 1981).

The rhizosphere is a narrow zone (1–2 mm thick) between a root and the surrounding soil. It is a zone of intense microbiological and biochemical activity because it receives appreciable amounts of organic substances from the roots that provide a substrate for a diverse microbial population. These organic substances include exudates, mucilage, and sloughed-off cells and their lysates (Marschner 1995). As a result of the processes taking place in the rhizosphere which include acidification, redox changes, and organic complex formation, some ions adsorbed onto the soil in the vicinity of the root may be desorbed and become available for absorption into the root. Phenolic compounds and amino acids are known to be involved in the mobilization of oxidized forms of iron and manganese (Fe^{3+} and Mn^{4+}) (Marschner 1995).

Cereal plants experiencing a deficiency of iron and/or zinc appear to have exudates containing substances (usually referred to as phytosiderophores) such as phytosiderophore 2′-deoxymugineic acid, which are effective in mobilizing these and other elements co-precipitated with them, including cadmium and copper, from iron and manganese oxides and other sorption sites in the vicinity of the root (Kabata-Pendias and Pendias 1992). Tobacco plants have root exudates that increase the absorption of cadmium but decrease that of iron (Mench and Martin 1991).

Kabata-Pendias and Pendias (1992) summarized the main points in the literature relating to the absorption of trace elements from solutions as:

- One of the most important factors determining the biological availability of trace elements is the extent to which they are bound to soil constituents.
- Plants take up the species of trace elements that are dissolved in the soil solution in either ionic or chelated and complexed forms.
- Absorption usually operates at very low concentrations.
- Adsorption depends largely on the concentrations in solution, especially at low ranges.
- The rate of absorption depends on the occurrence of H^+ and other ions.
- The intensity of absorption varies with plant species and stage of development.
- The processes of absorption are sensitive to some properties of the soil environment, such as: temperature, aeration, and redox potential.
- Absorption by a plant may be selective for a particular ion.
- The accumulation of some ions can take place against a concentration gradient.
- Mycorrhizae play an important role in cycling between media (e.g., soil) and roots.
- Root absorption can be by both passive (non-metabolic) and active (metabolic) processes.

In addition, Marschner (1995) stressed that there can be marked differences in ion uptake by different plant species (and cultivars within species; see Sect. 15.6.3).

15.6.2 Uptake of Major Elements by Plants

The major elements are present in much higher concentrations in plant tissues than trace elements and are referred to as "macronutrients." Trace elements are usually present at concentrations of <100 mg kg^{-1}, whereas macronutrients are present at levels of 1,000–40,000 mg kg^{-1} ($<4\%$) in the dry matter.

Wild and Jones(1988) and Marschner (1995) summarized the concentrations of macronutrient elements in crop plants as follows:

1. Calcium has a normal concentration range of 0.1–2.5% in plant dry matter but has a low mobility in plants and thus is not redistributed. Therefore the relatively high concentrations do not necessarily reflect the plants' metabolic requirements. Calcium plays a key role in the maintenance and integrity of membranes.
2. Magnesium generally occurs in the dry matter at lower concentrations than calcium (0.2–0.56%), but this element is more mobile than calcium. Magnesium is a specific constituent of chlorophyll. On serpentine soils, there may be a disproportionately large amount of available magnesium compared with calcium and this has lead to the development of specialized flora.
3. Sulfur is usually present at concentrations of 0.1–1% (dry matter) and is a constituent of the amino acids cysteine, cystine, and methionine, and therefore of proteins containing these. Biochemically, sulfur is a very important element because it is a constituent of enzymes and other key proteins. Crops take up 15–50 kg Sha^{-1} a^{-1}.
4. Phosphorus is present in plants at around 0.2% in the dry matter of shoots and is a key component in metabolic processes involving phosphorylation (e.g., ADP-ATP). It is present in the soil solution mainly as HPO_4^{2-} and $H_2PO_4^-$ but is taken up mainly as $H_2PO_4^-$. It is relatively strongly bound in most soils, in the organic matter, and sorbed to iron oxides (acid soils) or calcium carbonate with which it reacts to form apatite in calcareous and heavily limed soils.
5. Potassium occurs in similar concentrations as nitrogen in plants (1.4–5.6% in the dry matter) and is the most abundant cellular cation, but it can often be in short supply in crops. Crops can take up <500 kg Kha^{-1} a^{-1}, (400–500 kg Kha^{-1} a^{-1} in a 15th a^{-1}DM grass crop, 300 kg Kha^{-1} a^{-1} in a normal crop of potatoes, and <300 kg Kha^{-1} a^{-1} in a 10th a^{-1} (grain) crop of cereals).
6. Nitrogen can be present at 1.6–4% in the dry matter and is the fourth most abundant element in plants after carbon, hydrogen, and oxygen. Nitrogen has an essential role as a constituent of proteins, nucleic acids, chlorophyll, and growth hormones. It can be absorbed into the roots as either nitrate or ammonium ions but whatever the source, ammonium is the intermediate for the formation of amino acids, amides, and subsequently proteins.

15.6.3 Differences in Trace Element Accumulation Between and Within Plant Species

The amount of a trace element taken up from any particular soil will depend on plant factors as well as soil properties. Most important of the plant factors is the genotype or genetic makeup of the plant. Differences between plant families and species and also between varieties (cultivars) within a species are clearly shown in the case of cadmium. Grant et al. (1999) reviewed the literature on cadmium accumulation in crops and reported the general trend between plant families as:

1. Low accumulators: Leguminosae
2. Moderate accumulators: Graminae, Liliacae, Cucurbitacae, and Umbelliferae
3. High accumulators: Chenopodiacae, Cruciferae, Solanacae, and Compositae

However, marked differences can occur within families of plants and within species. For example, many varieties of durum wheat (*Triticum durum*) accumulate significantly more cadmium than common spring or winter (bread) wheats (*T. aestivum*).

In a review, Welch and Norvell (1999) reported that median cadmium concentrations in the seeds of seven different grain crops show a 100-fold variation, ranging from low values in rice and maize, to non-durum wheat, with durum wheat containing higher concentrations than non-durum wheat, and sunflower and flax containing the highest cadmium concentrations. Intraspecific variations of 40-fold have been reported for cadmium in 20 inbred lines of maize.

With regard to the risk to the health of consumers, it is not just the amount of a potentially toxic element, such as cadmium, which occurs in the whole plant that is important, but the concentrations found in the edible portions. Many plants (e.g., oat, soybean, timothy grass, alfalfa, maize, and tomato) tend to concentrate cadmium in their roots which implies that fruits, seeds, or leaves will be less enriched. Other species tend to have higher cadmium concentrations in their leaves, and these include the green leaf crops: lettuce, carrot, and tobacco. Lettuce and tobacco have a higher risk to consumers because of this characteristic distribution within the plant. It is considered that differences between varieties of some crops can be due to variations in the proportions of cadmium translocated around the plant from the roots (e.g., durum wheat).

The finding that varieties of key food crops can vary significantly in their accumulation of potentially hazardous elements such as cadmium does open up the possibility that plant breeders could use this to select new varieties with minimal concentrations in their edible parts. This is a distinct possibility for potatoes, rice, wheat, maize, lettuce, sunflower, and soybean. On the other hand, essential trace elements, such as zinc and copper, also show marked genotypic variations in plant species so plant breeders could exploit this trait. By selecting cultivars of food crops which contain higher concentrations of essential trace elements, such as zinc, it would be possible to match crops to marginally deficient soils and thus reduce the amount of trace element fertilizers needed to rectify the deficiency problems. The diets of people in many developing countries contain marginal to deficient concentrations of zinc and this can have marked effects on health.

15.7 Trace Elements in Soils and Crops and Human Health

Soils used for growing food crops in domestic gardens and commercial horticulture and agriculture can become contaminated with potentially toxic chemicals. In cases where the contamination is not sufficiently great to cause phytotoxicity and/or possible crop failure, there is a possible risk that livestock or people consuming large quantities of crop products grown on contaminated soil could suffer illness and even death from the chronic intake of contaminants or by the direct ingestion of the soil itself.

15.7.1 Concentrations of Selected Trace Elements in Soils in Different Parts of the World

Cadmium and lead are the elements in contaminated soils that are generally considered to constitute the widest possible health risk to humans through the plant uptake-dietary route. Surveys of soils in various countries have been carried out to determine the total (and sometimes bioavailable) concentrations of a range of trace elements, some of which are often found to have been elevated from anthropogenic sources of pollution. The concentrations of cadmium, lead, copper, and zinc in England and Wales, the United States, Florida, China, and Poland are shown in Table 15.6. These data show that all four elements are generally present at higher concentrations in England and Wales than in the other countries and state. This can be explained, at least in part, by the grid sampling (5 × 5 km cells) for England and Wales which included sites contaminated from various sources including metalliferous mining and heavy applications of sewage sludge. In the United States, the samples were collected on a more selective basis which avoided obvious sources of contamination either from sewage sludge or atmospheric deposition from nearby industries.

15.7.2 Concentrations of Cadmium and Lead in Food Crop Products

Surveys of the concentrations of potentially toxic elements, such as cadmium and lead, in food crops have also been undertaken in various countries and these help to show the extent to which anthropogenic contamination and/or geochemically enriched soil parent materials have influenced the composition of food crops consumed in the countries or areas considered (see also Chap. 8, this volume).

Table 15.6 Total concentrations of selected elements in topsoils in various countries or states (mg kg^{-1})

Country	Element	Mean	Minimum	Maximum	Median
England & Wales (n = 5,692)[a]	Cadmium	0.8	<0.2	40.9	0.7
	Lead	74.0	3.0	16,338	40
	Copper	23.1	1.2	1,508	18.1
	Zinc	97.1	5.0	3,648	82.0
United States (n = 3,045)[b]	Cadmium	0.265	<0.01	2.0	0.2
	Lead	12.3	7.5	135.0	11.0
	Copper	29.6	<0.6	495.0	18.5
	Zinc	56.5	<3.0	264.0	53.0
Florida (n = 448)[c]	Cadmium	0.07	0.004	2.8	0.004
	Lead	11.2	0.18	290	4.89
	Copper	6.1	0.1	318	1.9
	Zinc	8.35	0.9	169	4.6
China (n = 4,095)[d]	Cadmium	–	0.2	0.33	–
	Lead	–	9.95	56.0	–
	Copper	–	7.26	55.1	–
	Zinc	–	28.5	161.0	–
Poland (n = 127)[e]	Cadmium	–	0.1	1.7	–
	Lead	–	7.1	50.1	–
	Copper	–	2.0	18.0	–
	Zinc	–	10.5	154.7	–

[a]McGrath and Loveland (1992)
[b]Holmgren et al. (1993)
[c]Chen et al. (1998)
[d]Wei et al. (1990)
[e]Dudka (1992)

Consumption of grain, potatoes, and leafy vegetables is considered to account for more than 50% of total cadmium intake by people in most countries (Adriano 2001). In the UK, the largely wheat-based diet gave an intake of 8 μgCd day^{-1} which is only 11% of the WHO limit of 70 μgCd day^{-1}. The concentration of Cd in wheat grain decreased from 0.052 mgCd kg^{-1} in 1982 –0.042 in 1992 and 0.038 in 1993 due to reductions in atmospheric deposition, change to lower cadmium content phosphorus fertilizers, and increased yields giving rise to a dilution effect (Adriano 2001).

The European Union (EU) has legally defined maximum permissible concentrations for cadmium and lead in a range of foodstuffs including wheat (*T. aestivum* L.) and barley (*Hordeum vulgare* L.) grain. For cereals, excluding wheat grain, bran, germ, and rice, the maximum permissible cadmium concentration is 0.1 mg kg^{-1} (fresh weight), and the limit for the excluded crops mentioned above is 0.2 mg kg^{-1} (fresh weight). For lead the maximum permissible concentration in all cereals is 0.2 mg kg^{-1}. Assuming an 85% dry matter content, the effective limiting concentration on a dry matter basis is 0.118 mg kg^{-1} for cadmium in cereals excluding wheat grain, bran, germ, and rice and 0.235 mg kg^{-1} for the excluded categories. The dry matter based limiting concentration for lead is 0.235 mg kg^{-1} in all crop products.

A survey of the cadmium and lead concentrations in 250 samples of wheat grain and 233 samples of barley grain from throughout the UK after the 1998 harvest is reported by Adams et al. (2001). In the samples of wheat, they found overall mean concentrations (in the dry matter) of 0.063 mg kg^{-1} for cadmium and 0.025 mg kg^{-1} for lead. Only one sample of wheat had a cadmium concentration above the maximum permissible cadmium concentration. The 230 barley samples showed an overall mean value of 0.022 mg kg^{-1} for cadmium and 0.039 mg kg^{-1} for lead. Only one sample of barley exceeded the maximum permissible level for cadmium and two samples exceeded the maximum level for lead (Adams et al. 2001).

Both the wheat and barley samples showed small but significant differences between cultivars in mean concentrations of cadmium. The wheat samples also showed significant cultivar differences in lead concentrations (Adams et al. 2001).

Lead concentrations in samples of rice (*Oryza sativa*) from 17 areas of the world were reported by Zhang et al. (1996). The grand means for different countries ranged from 0.002 mg kg^{-1} in Australia to 0.039 mg kg^{-1} in Indonesia. Samples from within China showed a 76-fold variation in mean lead contents (0.016–0.152 mg kg^{-1}) in different parts of the country. A later paper (Zhang et al. 1998) reported lead and cadmium concentrations in 59 samples of several types of cereals and 34 samples of pulses from open markets in northeastern China. Average lead concentrations were 0.031 mg kg^{-1} for the cereals and 0.026 mg kg^{-1} for the pulses. Mean cadmium concentrations were higher in pulses than cereals (0.056 mg kg^{-1} pulses, 0.009 mg kg^{-1} cereals). Foxtail millet (*Stetaria itaica*) was found to contain the highest amounts of lead (0.054 mg kg^{-1}) and cadmium was highest in soya beans (*Glycine max*; 0.074 mg kg^{-1}). There were some possible links between food crop composition and human health effects. It had earlier been found that the consumption of Foxtail millet was a leading determinant of blood lead concentrations in Shandong Province and there was concern about the elevated levels of cadmium in pulses which are an important source of protein and lipids in Asia.

A much larger survey of cadmium and lead in 4,113 samples of rice and other cereal products in Japan was reported by Shimbo et al. (2001). They found a grand geometric mean for cadmium in polished, uncooked rice of 0.05 and 0.019 mg kg^{-1} in wheat flour. Mean lead concentrations were much lower for both rice and wheat flour ranging from 0.002 to 0.003 mg kg^{-1}. Rice was therefore shown to be a more important source of cadmium than wheat flour for Japan as a whole.

These four examples show how varying concentrations of elements, such as cadmium and lead in soil, together with genotypic differences between species and cultivars can give rise to variations in the trace element composition of food crops. The soil concentrations are a result of both the geochemical composition of the parent material and inputs from anthropogenic sources.

Examples of significant human health effects occurring as a result of soil contamination include (1) lead poisoning in children ingesting lead-rich garden soils and house dust, (2) skeletal deformity and death due to excessive cadmium intake by multiparous women living in the Jinzu Valley in the Toyama Province of Japan, and (3) tumors due to arsenic poisoning in people drinking contaminated water and eating vegetables and other crops from soils irrigated with arsenic-rich groundwaters in Bangladesh, India, and Taiwan (see Chap. 12).

15.7.3 Examples of Contaminated Soils Affecting Human Health

15.7.3.1 Exposure of Children to Lead in Contaminated Domestic Garden Soils

Domestic gardens can contain relatively high concentrations of lead from weathered and scraped painted interior and exterior surfaces (especially important around timber houses), deposition of lead from vehicle exhaust emissions

when gasoline used to contain relatively high levels of lead as an anti-knock agent, and other diffuse sources. Prior to the 1950s, some house paints contained more than 50% of lead in the dry matter. It was estimated in the late 1980s that up to 11.7 million children in the United States could have been at risk due to exposure to excessive amounts of lead in contaminated garden soils and house dusts (Millstone 1997). Culbard et al. (1988) conducted a survey of soil in 3,550 urban gardens in the UK and found the following geometric mean concentrations (in mg kg^{-1}): lead 230 (<14,125), cadmium 1.2 (<17), copper 53 (<16,800), and zinc 260 (<14,568). The geometric mean concentrations of metals in 579 gardens in greater London were significantly higher than for the rest of the country: lead 647 mg kg^{-1}, cadmium 1.3 mg kg^{-1}, copper 73 mg kg^{-1}, and zinc 424 mg kg^{-1}. This reveals that urban gardens in large cities, such as London, can be relatively heavily contaminated with potentially toxic elements including lead and zinc. However, zinc in high concentrations is normally regarded as more of a potential toxicity hazard to plants than to humans or animals.

Lead has a relatively low phytotoxicity compared with most other trace metals: cadmium > copper > cobalt, nickel > arsenic, chromium > zinc > manganese, iron > lead (Chino 1981). Therefore, crops grown in lead-contaminated gardens are unlikely to show symptoms of toxicity. However, the greatest risk to children and some adults is not through consumption of vegetables which have accumulated lead by uptake through their roots but by direct ingestion of lead-contaminated soil. This soil may have been ingested accidentally on unwashed vegetables and eating with unwashed hands. However, the greatest risk is to children who intentionally eat soil (*pica*) (see also Chap. 18, this volume).

A model has been proposed by Wixson and Davies (1993) to calculate safe guideline values for lead in soils where there may be a risk of soil ingestion by children. This model is:

$$S = \frac{(T/G^n - B)}{(\delta)} 1,000$$

where S is the soil guideline value, a geometric mean concentration of lead in $gPbg^{-1}$ of soil, T is the blood lead guideline or target concentration in μg $Pbdl^{-1}$ whole blood, G is the geometric standard deviation of the blood lead distribution (typically 1.3–1.5 but may be higher, e.g., in mining areas), B is the background or baseline blood lead concentration in the population from sources other than dust, n is the number of standard deviations corresponding to the degree of protection required, and δ is the slope or response of the blood Pb–soil Pb relationship and has units of μg $Pbdl^{-1}$ blood increase per 1,000 $\mu gPbg^{-1}$.

15.7.3.2 Cadmium Contamination in Japan and Great Britain

The Jinzu Valley in the Toyama Province of Japan has metalliferous mining and smelting industries located near the Jinzu River together with intensive paddy rice cultivation around the urban and industrial areas on the alluvial soils of the valley floor. Over many years leading up to WWII the mining and smelting operations had resulted in the contamination of the paddy soils with cadmium, lead, and zinc. Because of the cycle of reducing and oxidizing conditions associated with the flooding and drying out of the soils in the paddy fields, the contaminants, especially cadmium, underwent marked changes in speciation. In the flooded soils with strong reducing conditions, cadmium is present as an insoluble sulfide precipitate (CdS). As the growing rice approaches maturity, the paddy fields are drained to facilitate harvesting and it is during this period that the cadmium sulfide becomes oxidized and forms Cd^{2+} and SO_4^{2-}, which result in a massive uptake of cadmium ions into the rice plant and translocation into the grain (Asami 1984). As a consequence of this, rice grain grown in Jinzu Valley had significantly elevated concentrations of cadmium, and the people consuming this rice were exposed to a high risk of cadmium poisoning. It was the women who had had several children who were the worst affected by any kidney damage and an acute skeletal disorder known locally as itai-itai disease, which translated from Japanese means "ouch, ouch" due to the pain experienced by the sufferers when their bodies were touched. More than 200 elderly women living in the Valley during the 1940s who had borne several children were found to have been disabled by the disease and a further 65 women died from its effects. At that time, the rice-based subsistence diets of the peasant farmers were generally deficient in protein, calcium, and vitamin A, which all exacerbated the cadmium poisoning. In addition to the intake of cadmium in the rice, water taken from the river for drinking and cooking was also significantly contaminated with cadmium.

The average concentration of cadmium in the rice grown on the contaminated paddy soils was 0.7 mgCd kg^{-1}, which was more than ten times greater than the average cadmium concentration in local control samples of rice (0.07 mgCd kg^{-1}). The maximum content of cadmium found in rice was 3.4 mgCd kg^{-1}.

The mean cadmium intake for the residents of the Jinzu Valley was estimated to be around 600 mg day^{-1}, which is around ten times greater than the maximum tolerable intake. The Japanese government has set 1mgCd kg^{-1} in rice as the maximum allowable limit, and it has been found that 9.5% of paddy soils in Japan have been polluted with cadmium to the extent that the concentrations in rice are greater than this

value. Even in areas considered to be unpolluted with cadmium, rice is estimated to be the source of more than 60% of cadmium in the diet. Much of this cadmium is probably derived from phosphatic fertilizers.

An interesting comparison can be made with the village of Shipham in Somerset, England, where a relatively large number of houses had been built in the 1950s and 1960s on land contaminated with cadmium, lead, zinc, and other elements from historic metalliferous mining (Morgan and Sims 1988). Although the total concentrations of cadmium were in some cases more than 100 times those in the Jinzu Valley soils (<360 mgCd kg^{-1}), the cadmium was less bioavailable due to the mining-contaminated soils in Shipham containing appreciable amounts of free calcium carbonate and, consequently, a higher pH and a significant amount of chemisorption (Alloway et al. 1988). The Shipham soils were also contaminated with zinc and lead (<37,200 mgZn kg^{-1} and 6,540 mgPb kg^{-1}) and antagonistic reactions with these metals may have also played a role in reducing cadmium uptake. Nevertheless, the mean concentration of cadmium in vegetables grown in gardens in Shipham was 0.25 mgCd kg^{-1} DM, which is nearly 17 times higher than the UK national average of 0.015 mgCd kg^{-1} DM. Some samples of leafy vegetables growing on the most highly contaminated soils contained much higher concentrations but these were not common. Health studies on approximately 500 inhabitants of Shipham village revealed some small but significant differences in biochemical parameters, but there was no evidence of adverse health effects in the participants in the investigations.

15.8 Widespread Deficiencies of Essential Trace Elements

Large areas of the world have soils which are unable to supply staple crops, such as rice, maize, and wheat with sufficient zinc. Smaller, but still significant areas of the world are affected by deficiencies of boron, copper, iron, and manganese. In crops, acute deficiencies of essential trace elements result in visible symptoms of stress, such as chlorosis (yellow coloration due to impaired chlorophyll production) and reduced dry matter growth and yield of edible crop products. However, marginally deficient amounts of certain trace elements, such as copper and zinc, can give rise to hidden deficiencies in which yields can be reduced by up to 30% or more without the appearance of obvious visible symptoms of stress. Plants may be smaller but unless normal (sufficient) crops are available nearby for comparison, marginal deficiencies are often difficult to detect. This can result in crop yields which may be very much less than optimal due to a trace element deficiency that could be easily treated by either application of fertilizers to the soil or foliar sprays to the crop.

The impact of deficiencies of essential trace elements on crop yields and quality is much greater overall than that of toxicity due to pollution. In several countries, large proportions of the arable soils are affected by deficiencies, such as in India where around 45% of soils are deficient in zinc, 33% deficient in boron, 8.3% deficient in iron, 4.5% deficient in manganese, and 3.3% deficient in copper (Singh 2001). Apart from lost food production in terms of tonnes of carbohydrates and protein, the trace element composition of the crop products is also affected. The zinc status of many people in developing countries, such as India and Pakistan, is considered to be suboptimal and therefore low zinc concentrations in food products will exacerbate this problem. Cereal staples, such as rice, wheat, and maize contain relatively large amounts of phytate which tends to make zinc in the diet less available. Phytate concentrations tend to be higher in crops receiving high levels of phosphorus fertilizer, such as the new, high-yielding varieties.

Deficiencies of essential trace elements in crops can be caused by low total concentrations of elements in the soil parent material (such as sandstones and sandy drift deposits), by low availability due to high soil pH, high concentrations of calcium (in calcareous soils on limestones or in semi-arid or arid areas), high organic matter contents (peaty soils), and waterlogged (gleyed) conditions. Zinc deficiency is the most widespread essential trace element deficiency problem in the world and many large areas of land in hot arid climates have severe zinc deficiency problems in calcareous soils (such as in the Middle East) and in rice paddy soils (as in India, Philippines, and Bangladesh).

As with other plant species, crops such as wheat, rice, and maize have been found to vary widely in their ability to tolerate deficiencies of essential trace elements such as zinc. "Zinc-efficient" cultivars of wheat are those that are able to grow and yield reasonably well on soils in which less tolerant cultivars would be acutely affected by zinc deficiency. The zinc efficiency character will enable plant breeders to produce new hybrids that are better suited to zinc-deficient soils. Although any zinc-deficient soil can be fertilized with a zinc compound (such as zinc sulfate) to raise its plant-available concentration of zinc, the use of zinc-efficient cultivars would enable plants to be matched to the soils, rather than the soils fertilized to match them to the crop's requirements. This is particularly important in areas where zinc fertilizers are difficult for farmers to obtain.

As discussed in Selenium Deficiency and Toxicity in the Environment (in this volume), widespread deficiencies of selenium are also related to soil parent materials containing low concentrations of this and other essential trace elements.

15.9 Summary

The health of people and animals can be affected by imbalances in their dietary intake of trace and major elements and therefore the rock-soil-plant-animal/human pathway is of major importance in considerations of medical geology. Soils can differ widely in their total concentrations of both macro and trace elements due to variations in the mineralogy of the geological parent material on which the soil has formed, even without inputs from environmental pollution or agricultural husbandry. Concentrations of essential trace elements can be anomalously low on parent materials such as sandstones and sandy or gravelly glacial drift. These can be due to both low concentrations of these elements in the predominant mineral (quartz) and the relatively low adsorptive capacities of sandy soils (arenosols) due to their very low clay contents. On the other hand, anomalously high concentrations of both essential and potentially toxic elements can occur in soils formed on geological materials which themselves have anomalously high concentrations of certain elements. Examples of this are marine black shales enriched in arsenic, cadmium, copper, chromium, mercury, molybdenum, lead, and zinc and serpentinite rocks enriched in magnesium, cobalt, chromium, and nickel.

Given that marked variations can occur in the total concentrations of elements in the soil, the mobility and bioavailability of these elements is going to be controlled by a range of soil factors of which the pH is probably the most important. The adsorptive capacity of a soil is dependent on its mineralogy, especially the amount and type of clays; the amounts of iron, manganese, and aluminum oxides and free carbonates; and the organic matter content. The redox conditions will have a marked control on the oxide content and the free carbonates will have a major effect on the soil pH.

Even after the bioavailable concentrations have been determined by the soil and parent material factors, the plants growing on the soil have both an effect on the soil and also vary markedly in their ability to accumulate trace and major elements due to differences in their genotypes. Therefore, the potential impact of geology on the dietary intake of trace and major elements by humans and livestock is dependent on a range of soil and plant factors. Fortunately many of the mechanisms in soils are reasonably well understood and in due course it will become possible to model their behavior and make more realistic risk assessments of both potential toxicity and deficiency problems. With regard to the plant genotypic factor, with the continual breeding of new crop varieties, there is an ongoing need to screen these for their ability to accumulate both essential and non-essential elements as they come into use. This would enable the most suitable crop varieties to be grown. This could apply to varieties of crops that have a high efficiency of use of essential trace elements such as zinc in areas where there is a major problem of deficiencies both limiting yields and the dietary available concentrations of trace elements essential for humans consuming the crops. Conversely, where soils have relatively high concentrations of potentially harmful elements, such as arsenic and cadmium, either as a result of anomalous geochemical composition or environmental pollution, then crops which accumulate relatively low concentrations of these elements would be most suitable. Amelioration of soils by manipulation of soil chemical properties, such as pH, can be carried out, but selecting the most appropriate plant species and cultivars would also help to mitigate the problems.

Apart from the uptake of elements by plants followed by the consumption of crop products by people, there are other ways (dealt with elsewhere in this book) by which the composition and properties of soils can affect human health. These range from the inhalation and ingestion of atmospheric dusts containing soil minerals eroded by the wind from land, the intake of elements in drinking water which has been affected by the leaching of elements through soil profiles, and, lastly, to the intentional eating of soil (geophagia). The soil is bviously a key consideration in medical geology.

See Also the Following Chapters. Chapter 17 (Soils and Iodine Deficiency) • Chapter 18 (Geophagy and the Involuntary Ingestion of Soil) • Chapter 19 (Natural Aerosolic Mineral Dusts and Human Health) • Chapter 20 (The Ecology of Soil-Borne Human Pathogens)

Further Reading

Adams ML, Zhao FJ, McGrath SP, Nicholson FA, Chalmers A, Chambers BJ, Sinclair AH (2001) Cadmium and lead in British wheat and barley: survey results and the factors affecting their concentration in grain, Project report 265. Home Grown Cereals Authority (HGCA), London

Adriano DC (2001) Trace elements in terrestrial environments, 2nd edn. Springer, New York

Alloway BJ (ed) (1995) Heavy metals in soils, 2nd edn. Blackie Academic & Professional, London

Alloway BJ, Thornton I, Smart GA, Sherlock J, Quinn MJ (1988) Metal availability (in the Shipham report). Sci Total Environ 75:41–69

Asami T (1984) Pollution of soils by cadmium. In: Nriagu JO (ed) Changing metal cycles and human health. Springer, Berlin, pp 95–111

Barber S (1984) Soil nutrient bioavailability: a mechanistic approach. Wiley, New York

Bittel JE, Miller RJ (1974) Lead, cadmium and calcium selectivity coefficients on montmorillonite, illite and kaolinite. J Environ Qual 3:243–254

Brady NC (1984) The nature and properties of soils, 9th edn. Collier Macmillan, New York

Brummer GW (1986) The importance of chemical speciation in environmental processes. Springer, Berlin, pp 169–192
Bunzl K, Schmidt W, Sansomi B (1976) The kinetics of ion exchange in soil organic matter. IV adsorption and desorption of Pb^{2+}, Cu^{2+}, Cd^{2+}, Zn^{2+} and Ca^{2+} by Peat. J Soil Sci 27:154–166
Chen M, Ma LQ, Harris W (1998) Background concentrations of trace elements in Florida surface soils: annual progress report, Soil and Water Science Dept. University of Florida, Gainesville
Chino M (1981) Uptake-transport of toxic metals in rice plants, in Heavy Metal Pollution of soils in Japan. In: Kitagishi K, Yamane I (eds) Japan Scientific Society proceedings, Tokyo
Culbard EB, Thornton I, Watt J, Wheatley M, Moorcroft S, Thompson M (1988) Metal contamination in British suburban dusts and soils. J Environ Qual 17:226–234
Dudka S (1992) Factor analysis of total concentrations of elements in surface soils of Poland. Sci Total Environ 121(1):39–52
Forbes EA, Posner AM, Quirk JP (1976) The specific adsorption of divalent Cd, Co, Cu, Pb and Zn on goethite. J Soil Sci 27: 154–166
Graham RD (1981) Chapter 7: Absorption of copper by plant roots. In: Loneragan JF, Robson AD, Graham RD (eds) Copper in soils and plants. Academic, Sydney, pp 141–164
Grant CA, Bailey LD, McLaughlin MJ, Singh BR (1999) Chapter 7: Management factors which influence cadmium concentrations in crops. In: McLaughlin MJ, Singh BR (eds) Cadmium in soils and plants. Kluwer Academic, Dordrecht
Hayes MHB, Swift RS (1978) Chapter 3: The chemistry of organic colloids. In: Greenland DJ, Hayes MHB (eds) The chemistry of soil constituents. Wiley, Chichester, pp 179–320
Holmgren CGS, Meyer NW, Chaney RL, Daniels RB (1993) Cadmium, lead, zinc, copper and nickel in agricultural soils in the United States. J Environ Qual 22:335–348
ISSS Working Group RB, Bridges EM, Batjes NH, Nachtergaele FO (eds) (1998) World reference base for soil resources: atlas. ISRIC-FAO-ISSS-Acco, Leuven
Jenny H (1941) The factors of soil formation. McGraw-Hill, New York
Kabata-Pendias A, Pendias H (1992) Trace elements in soils and plants, 2nd edn. CRC Press, Boca Raton
Kinniburgh DG, Jackson ML, Syers JK (1976) Adsorption of alkaline earth transition and heavy metal cations by hydrous gels of iron and aluminum. Soil Sci Soc Am Proc 40:769–799
Lindsay WL (1979) Chemical equilibria in soils. Wiley, New York
Marschner H (1995) Mineral nutrition of higher plants, 2nd edn. Academic, San Diego
McGrath SP, Loveland PJ (1992) Soil geochemical atlas of England and Wales. Blackie Academic & Professional, Glasgow
McKenzie RM (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron. Aust J Soil Res 18:61–73
Mench M, Martin E (1991) Mobilisation of cadmium and other metals from two soils by root exudates of *Zea Mays* L., and *Nicotiana rustica* L. Plant Soil 132:187–193
Millstone E (1997) Lead and public health. Earthscan Publications, London
Morgan H, Sims DL (1988) Discussion and conclusions (in "the Shipham report"). Sci Total Environ 75:135–143
Murray K, Lindler PW (1983) Fulvic acids: structure and metal binding. 1. A random molecular model. J Soil Sci 34:511–523
Okazaki M, Takamidoh K, Yamane I (1986) Adsorption of heavy metal cations on hydrated oxides of iron and aluminum with different crystallinities. Soil Sci Plant Nutr (Tokyo) 32:522–533
Papadopoulos P, Rowell DL (1988) The reactions of cadmium with calcium carbonate surfaces. J Soil Sci 39:23–35
Reimann C, Siewers U, Tarvainen T, Bityukova L, Eriksson J, Glucis A, Gregorauskiene V, Lukashev V, Matinian NN, Pasieczna A (2000) Baltic soil survey: total concentrations of major and selected trace elements in arable soils from 10 countries around the Baltic Sea. Sci Total Environ 257:155–170
Ross SA (1989) Soil processes: a systematic approach. Routledge, London
Shimbo S, Zhang ZW, Watanabe T, Nakatsuka H, Matsuda-Inoguchi N, Higashikawa K, Ikeda M (2001) Cadmium and lead contents in rice and other cereal products in Japan in 1998–2000. Sci Total Environ 281:165–175
Singh MV (2001) Evaluation of current micronutrient stocks in different agro-ecological zones of India for sustainable crop production. Fertil News (Delhi) 46(2):25–42
Sposito G (1983) The chemical form of trace elements in soils. In: Thornton I (ed) Applied environmental geochemistry. Academic, London, pp 123–170
Stevenson FJ (1979) Humus. In: Fairbridge RW, Finkl CW (eds) Encyclopedia of soil science, Dowden, Hutchinson & Ross, Stroudsburg, PA, pp 195–205
Stuanes A (1976) Adsorption of Mn^{2+}, Zn^{2+}, Cd^{2+} and Hg^{+} from binary solutions by mineral material. Acta Agrica Scand 26:243–250
Tipping E, Hurley MA (1992) A unifying model of cation binding by humic substances. Geochem Cosmochim 56:3627–3641
Wei FS, Chen JS, Zheng CJ, Jiang DZ (eds) (1990) Elemental background concentrations of soils in China. Chain Environmental Scientific Publishing Ltd., Beijing (Reported in Chen et al., 1998)
Welch RM, Norvell WA (1999) Chapter 6: Mechanisms of cadmium uptake, translocation and deposition in plants. In: McLaughlin MJ, Sing BR (eds) Cadmium in soils and plants. Kluwer Academic Publishers, Dordrecht, pp 125–150
White RE (1997) Principles and practice of soil science, 3rd edn. Blackwell Science, Oxford
Wild A, Jones LHP (1988) Chapter 4: Mineral nutrition of crop plants. In: Wild A (ed) Russell's soil conditions and plant growth, 11th edn. Longman Scientific & Technical, Harlow
Wixson BG, Davies BE (1993) Lead in soil: recommended guidelines. Science Reviews, Northwood
Zhang ZW, Moon CS, Watanabe T, Shimbo S, Ikeda M (1996) Lead content of rice collected from various areas of the world. Sci Total Environ 191:169–175
Zhang ZW, Watanabe T, Shimbo S, Higahikawa K, Ikeda M (1998) Lead and cadmium contents in cereals and pulses in north-eastern China. Sci Total Environ 220:137–145

Selenium Deficiency and Toxicity in the Environment

16

Fiona M. Fordyce

Contents

16.1 Background

Selenium (Se) is a naturally occurring metalloid element, which is essential to human and other animal health in trace amounts but is harmful in excess. Of all the elements, selenium has one of the narrowest ranges between dietary deficiency (<40 μg day^{-1}) and toxic levels (>400 μg day^{-1}) (WHO 1996), which makes it necessary to carefully control intakes by humans and other animals, hence, the importance of understanding the relationships between environmental exposure and health. Geology exerts a fundamental control on the concentrations of selenium in the soils on which we grow the crops and animals that form the human food chain. The selenium status of populations, animals, and crops varies markedly around the world as a result of different geological conditions. Because diet is the most important source of selenium in humans, understanding the biogeochemical controls on the distribution and mobility of environmental selenium is key to the assessment of selenium-related health risks. High selenium concentrations are associated with some phosphatic rocks, organic-rich black shales, coals, and sulfide mineralization, whereas most other rock types contain very low concentrations and selenium-deficient environments are far more widespread than seleniferous ones. However, health outcomes are not only dependent on the total selenium content of rocks and soils but also on the amount of selenium taken up into plants and animals—the bioavailable selenium. This chapter demonstrates that even soils containing adequate total amounts of selenium can still produce selenium-deficient crops if the selenium is not in a form ready for plant uptake.

The links between the environmental biogeochemistry of selenium and health outcomes have been documented for many years. Selenium was first identified in 1817 by the Swedish chemist Jons Jakob Berzelius; however, selenium toxicity problems in livestock had been recorded for hundreds of years previously although the cause was unknown. Hoof disorders were noted in livestock in Colombia in 1560 and in South Dakota (U.S.) in the

F.M. Fordyce (✉)
Geochemical Baselines and Medical Geology Team, British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK
e-mail: fmf@bgs.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_16, British Geological Survey © NERC 2013

mid-nineteenth century where the symptoms were called "alkali disease" (Reilly 2006). In the early 1930s this disease, which is characterized by hair and hoof loss and poor productivity, was identified as selenium toxicosis (selenosis) (Franke 1934; Moxon 1937; Rosenfield and Beath 1964). Since then, seleniferous areas have been reported in Ireland, Israel, Australia, Russia, Venezuela, China, the United States, India and South Africa (Levander 1986; Oldfield 1999; Hira et al. 2004).

Conversely, selenium was identified as an essential trace element during pioneering work into selenium-responsive diseases in animals in the late 1950s and early 1960s (Schwarz and Foltz 1957). Selenium forms a vital constituent of the biologically important enzyme glutathione peroxidase (GPx), which acts as an antioxidant preventing oxidative cell degeneration (Rotruck et al. 1969). In animals, selenium deficiency has been linked to muscular weakness and muscular dystrophy, but it also causes reduced appetite, poor growth and reproductive capacity, and embryonic deformities. These disorders are generally described as white muscle disease (WMD). Following these discoveries, selenium deficiencies in crops and livestock have been reported in all regions of the world including the United States, the UK, Finland, Denmark, Sri Lanka, New Zealand, Australia, India, Canada, Thailand, Africa, and China and selenium supplementation has become common practice in agriculture (Levander 1986; WHO 1987; Oldfield 1999).

Selenium deficiency has also been implicated in the incidence of a heart disorder (Keshan disease) and bone and joint condition (Kashin-Beck disease) in humans in various parts of China (Chen et al. 1980; Yang et al. 1983). More recent research has shown that selenium deficiency also adversely affects thyroid hormone metabolism, which is detrimental to growth and development (Arthur and Beckett 1994). Indeed, 25 essential selenoproteins have now been identified in humans, many of which are involved in catalytic functions in the body (Papp et al. 2007; Rayman 2012). Selenium deficiency has also been implicated in a host of conditions including cancer, heart disease, immune system function, and reproduction. This chapter outlines some of the health problems in humans and animals that can arise as a result of selenium deficiency and toxicity in the natural environment. These links are more obvious in regions of the world where the population is dependent on local foodstuffs in the diet, but studies show that even in countries such as the United States where food is derived from a range of exotic sources, the local environment still determines the selenium status of the population. This fact should not be ignored because medical science continues to discover new essential functions for this biologically important element.

16.2 Selenium in the Environment

The naturally occurring element selenium belongs to group VIA of the periodic table and has chemical and physical properties that are intermediate between metals and nonmetals (Table 16.1). Selenium occurs in nature as six stable isotopes; however, it should be noted that although ^{82}Se is generally regarded as a stable isotope, it is actually a β^- emitter with a very long half-life of 1.4×10^{20} year. The chemical behavior of selenium resembles that of sulfur and like sulfur, selenium can exist in the 2^-, 0, 4^+, and 6^+ oxidation states (Table 16.2). As a result of this complex chemistry, selenium is found in all natural materials on Earth including rocks, soils, waters, air, and plant and animal tissues (Table 16.3) (see also Chap. 2, this volume).

At the global scale, selenium is constantly recycled in the environment via the atmospheric, marine, and terrestrial systems. Estimates of selenium flux indicate that anthropogenic activity is a major source of selenium release in the cycle, whereas the marine system constitutes the main natural pathway (Table 16.4) (Haygarth 1994). Selenium cycling through the atmosphere is significant because of the rapidity

Table 16.1 Physical properties of selenium

Element name	Selenium
Chemical symbol:	Se
Atomic number:	34
Periodic table group:	VIA
Atomic mass:	78.96
Density:	4,808 kg m^{-3}
Melting point:	220°C
Boiling point:	685°C
Vapor pressure:	1 mmHg @ 356°C
Natural isotopes:	Abundance:
^{74}Se	0.87%
^{76}Se	9.02%
^{77}Se	7.58%
^{78}Se	23.52%
^{80}Se	49.82%
^{82}Se	9.19%

From Jacobs (1989) and ATSDR (2003)

Table 16.2 Chemical forms of selenium in the environment

Oxidative state	Chemical forms
Se^{2-}	Selenide (Se^{2-}, HSe^-, H_2Se_{aq})
Se^0	Elemental selenium (Se^0)
Se^{4+}	Selenite (SeO_3^{2-}, $HSeO_3^-$, H_2SeO_{3aq})
Se^{6+}	Selenate (SeO_4^{2-}, $HSeO_4^{2-}$, H_2SeO_{4aq})
Organic Se	Selenomethionine, selenocysteine

From Jacobs (1989) and Neal (1995)

Table 16.3 Selenium concentrations in selected natural materials

Material	Total Se (mg kg^{-1})	Water-soluble Se (ng g^{-1})	Material	Total Se (mg kg^{-1})
Earth's crust:	0.05		Water (μg L^{-1}): World freshwater	0.02
Igneous rocks (general):	0.35		Brazil River Amazon	0.21
Ultramafic (general)	0.05		U. S. (general)	<1
Mafic (general)	0.05		U. S. seleniferous	50–300
Granite (general)	0.01–0.05		U. S. Kesterson	<4,200
			U. S. River Mississippi	0.14
Volcanic rocks (general):	0.35		U. S. River Colorado	10–400
United States	<0.1		U. S. River Gunnison	10
Hawaii	<2.0		U. S. Lake Michigan	0.8–10
Tuffs (general)	9.15		U. S. seleniferous gw	2–1,400
			U. S. drinking water	0.0–0.01
Sedimentary rocks:			Spain freshwater	0.001–0.202
Limestone (general)	0.03–0.08		China Se-deficient sw	0.005–0.44
Sandstone (general)	<0.05		China Se-adequate sw	1.72
Shale (general)	0.05–0.06		China seleniferous sw	0.46–275
W. USA shales	1–675		Finland stream water	0.035–0.153
Wyoming shales	2.3–52		Canada stream water	1–5
South Korea shales	0.1–41		Norway groundwater	0.01–4.82
China Carbon-shale	206–280		Slovakia groundwater	0.5–45
Mudstones (general)	0.1–1500		Bulgaria drinking water	<2
Carbonates (general)	0.08		Sweden drinking water	0.06
Marine carbonates	0.17		Germany drinking water	1.6–5.3
Phosphates (general)	1–300		Ukraine surface water	0.09–3
U. S. Coal	0.46–10.65		Ukraine groundwater	0.07–4
Australia Coal	0.21–2.5		Argentina surface water	2–19
China stone-coal	<6471		Reggio, Italy dw	7–9
Oil (general)	0.01–1.4		Sri Lanka drinking water	0.056–0.235
			Greece drinking water	0.05–0.700
Soil:			Polar ice (general)	0.02
World (general)	0.4		Seawater (general)	0.09
World seleniferous	1–5000			
U. S. (general)	<0.1–4.3		Plants:	
U. S. seleniferous	1–10		U. S. grasses	0.01–0.04
England/Wales (general)	<0.01–16	50–390	U. S. clover and alfalfa	0.03–0.88
Scotland (general)	0.115–0.877	6.69–26.78		
Northern Ireland (general)	< 0.02–7.8			
Ireland seleniferous	1–1200		Norway moss	0.8–1.23
China (general)	0.02–3.81		Canada tree bark	2–16
China Se-deficient	0.004–0.48	0.03–5	Norway grain	0.006–0.042
China Se-adequate	0.73–5.66		Norway forage	0.05–0.042
China seleniferous	1.49–59.4	1–254	Finland hay	0–0.04
Finland (general)	0.005–1.241		Finland grain	0.007
India Se-deficient	0.025–0.71	19–66		
India seleniferous	1–19.5	50–620	Algae:	
Sri Lanka Se-deficient	0.112–5.24	4.9–43.3	Marine (general)	0.04–0.24
Norway (general)	3–6		Freshwater (general)	<2
Greece Se-deficient	0.05–0.10			
Greece Se-adequate	>0.2		Fish:	
New Zealand (general)	0.1–4		Marine (general)	0.3–2
Malawi (general)	0.05–0.62		Freshwater (general)	0.42–0.64
Urban Soils:				
London	<0.2–20		Animal tissue (general):	0.4–4

(continued)

Table 16.3 (continued)

Material	Total Se (mg kg^{-1})	Water-soluble Se (ng g^{-1})	Material	Total Se (mg kg^{-1})
Glasgow	1–15			
Air (ng m^{-3}) (general):	0.00006–30		Stream sediments: Wales	0.4–83
Atmospheric dust (general):	0.05–10			

From Fleming (1980), Thornton et al. (1983), Levander (1986), WHO (1987), Jacobs (1989), Nriagu (1989), Tan (1989), Fergusson (1990), Hem (1992), Haygarth (1994), Neal (1995), Rapant et al. (1996), Fordyce et al. (1998), Reimann and Caritat (1998), Vinceti et al. (1998), Oldfield (1999), British Geological Survey (2000), Fordyce et al. (2000a), Fordyce et al. (2010), British Geological Survey (2011), Chilimba et al. (2011), Fordyce et al. (2012), GSNI (2012), Rawlins et al. (2012)
Note: gw groundwater, *sw* surface water, *dw* drinking water

Table 16.4 Global selenium fluxes

Cycle	Selenium flux (tonnes per year)
Anthropogenic	76,000–88,000
Marine	38,250
Terrestrial	15,380
Atmospheric	15,300

Modified from Haygarth (1994)

of transport, but the terrestrial system is most important in terms of animal and human health because of the direct links with agricultural activities and the food chain.

Although the element is derived from both natural and man-made sources, an understanding of the links between environmental geochemistry and health is particularly important for selenium as rocks are the primary source of the element in the terrestrial system (Table 16.5) (Fleming 1980; Neal 1995). Selenium is dispersed from the rocks through the food chain via complex biogeochemical cycling processes including weathering to form soils, rock-water interactions, and biological activity (Fig. 16.1). As a result, selenium is not distributed evenly across the planet, rather concentrations differ markedly depending on local conditions and an understanding of these variations is essential to aid the amelioration of health problems associated with selenium deficiency and toxicity. The following sections of this chapter provide a brief summary of anthropogenic sources of the element before going on to discuss the important aspects of selenium in the natural biogeochemical cycle and impacts on health.

16.2.1 Man-Made Sources of Selenium

Following its discovery in 1817, little industrial application was made of selenium until the early twentieth century when it began to be used as a red pigment and improver in glass and ceramic manufacture; however, it was not until the invention of the photocopier in the 1930s that demand for the element significantly increased due to its photoelectric and semi-conductor properties. Today the use of selenium in photocopiers is declining as it has been replaced by organic photoreceptor compounds, which give better performance at lower cost (USGS 2007). However, selenium is widely used in a number of industries (Table 16.6); most commonly selenium dioxide is employed as a catalyst in metallurgy and organic synthesis and as an antioxidant in inks, mineral, vegetable, and lubricating oils. Selenium mono- and disulfide are also used in anti-dandruff and antifungal pharmaceuticals (WHO 1987; Haygarth 1994; ATSDR 2003).

The world industrial output of selenium was estimated at 3,000–3,500 tonnes in 2010, and the largest producers are Japan, Germany, and the United States. It is used mainly in the metallurgy and glass industries (Table 16.6). It is not economical to mine mineral deposits specifically for selenium; rather the element is recovered from the electrolytic refining of copper and lead and from the sludge accumulated in sulfuric acid plants (ATSDR 2003).

Selenium compounds are released to the environment during the combustion of coal and petroleum fuels; during the extraction and processing of copper, lead, zinc, uranium, and phosphate; and during the manufacture of selenium-based products. According to monitoring data in the United States, in 2002 over 2,700 tonnes of selenium compounds were released to the environment, primarily from power and sanitation industries (Table 16.7) (U.S.-EPA 2002).

It is estimated that 76,000–88,000 tonnes $year^{-1}$ of selenium are released globally from anthropogenic activity, compared to natural releases of 4,500 tonnes $year^{-1}$, which gives a biospheric enrichment factor value of 17. This value is significantly higher than 1 indicating the important influence of man in the cycling of selenium (Nriagu 1991). For example, long-term monitoring data from the Rothamstead Agricultural Experimental Station in the UK demonstrate the impact of anthropogenic activity on selenium concentrations in herbage. Samples collected between 1861 and 1990 bulked at 5-year intervals reveal that the highest concentrations occurred between 1940 and 1970 which coincided with a period of intensive coal use. Due to the move to fuel sources such as nuclear, oil, and gas in more recent decades, selenium concentrations in herbage are declining (Haygarth 1994).

Table 16.5 Main sources of selenium in the environment

	Comments
Natural sources	
Volcanic activity	Important source
Weathering of rocks	Important source
Sea spray	Concentrations in ocean water are only an order of magnitude lower than those in rocks
Atmospheric flux	From the ocean surface to the atmosphere
Volatilization and recycling from biota	
Aerial deposition	For example, in the UK annual selenium deposition = 2.2–6.5 g ha^{-1}
Man-made sources	
Selenium-based industries	
Metal processing industries	Important source
Burning of fossil fuels	Important source
Disposal of sewage sludge to land	Typical selenium contents 1–17 mg kg^{-1}
Agricultural use of pesticides	Potassium ammonium sulfide ($[K(NH_4)S]_5Se$)
Agricultural use of lime	Typical selenium contents 0.08 mg kg^{-1}
Agricultural use of manure	Typical selenium contents 2.4 mg kg^{-1}
Agricultural use of phosphate fertilizers	Typical selenium contents 0.08–25 mg kg^{-1}

From Fleming (1980), Haygarth (1994), and Neal (1995)

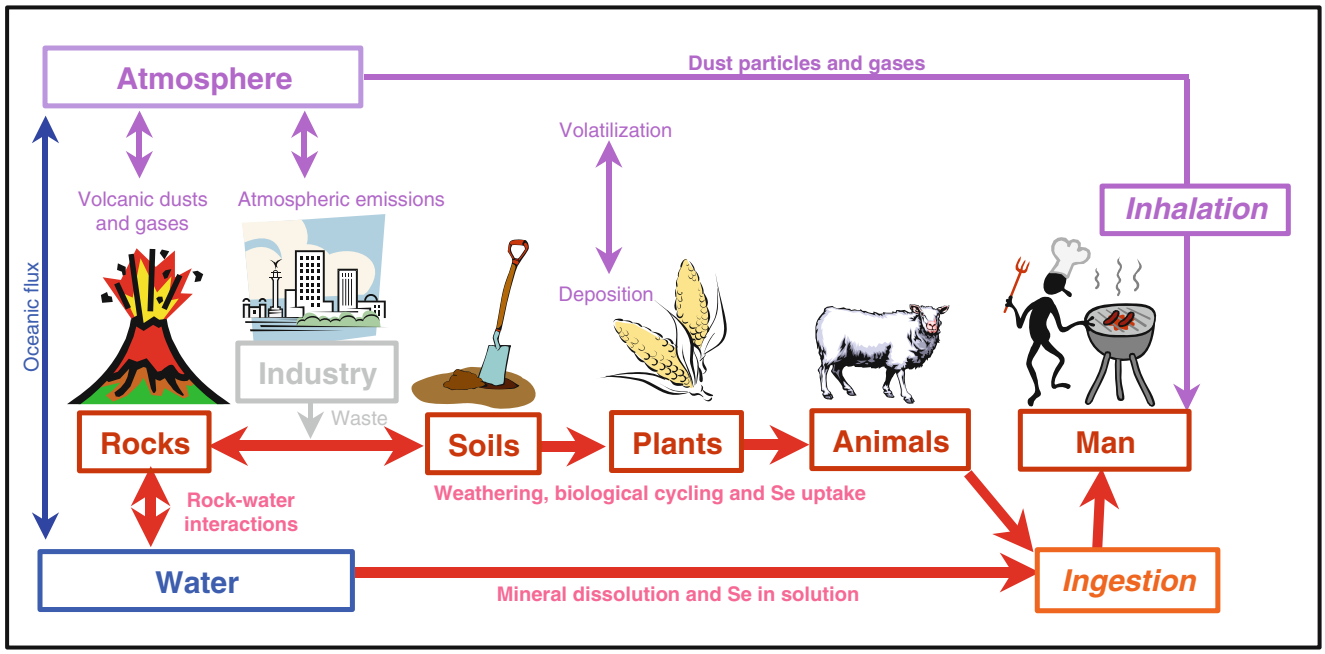


Fig. 16.1 Simplified schematic diagram of the cycling of selenium from the environment to man. The main geochemistry and health pathways are shown in *red/thicker arrows*

Selenium is also released inadvertently into the environment from the agricultural use of phosphate fertilizers, from the application of sewage sludge and manure to land, and from the use of selenium-containing pesticides and fungicides (Table 16.5). For example, in the European Union (EU) it is no longer permissible to dump sewage sludge at sea; consequently the application to land has increased in recent years. To help avoid potential environmental problems, a maximum permissible concentration (MAC) of selenium in sewage sludge in the UK is set at 25 mg kg^{-1} and in France at 200 mg kg^{-1}, whereas the MAC in soil after application is 3 mg kg^{-1} in the UK and 10 mg kg^{-1} in France (Haygarth 1994; DOE 1996). Clearly, the application of sewage sludge to land increases the selenium content of the soil; however, there are relatively few studies of uptake into plants. Those that have been conducted suggest soil-to-plant concentration factors of $1.16 - 9.36 \times 10^{-3}$ mg kg^{-1} for sludge amended soils (EA 2009). The application of selenium-bearing fertilizers to land has been used to remediate selenium

Table 16.6 Industrial uses of selenium

Industry	Percentage of total world production used by various industries in 2010
Metal alloys/metallurgy	30%
Glass manufacture	30%
Agriculture	10%
Fertilizer	
Nutritional additive to livestock feed	
Medical use, e.g., dietary supplements	
Electrical components—semi-conductors, cables, and contacts	10%
Photocopier components	
Pharmaceutical catalyst	10%
Fungicide	
Pesticide	
Anti-dandruff shampoo	
Pigments in plastics, paints, enamels, inks, rubber, textiles	
Additive to petroleum fuel and lubricant products	
Accelerator and vulcanizing agent in rubber manufacture	
Photographic emulsions, printing, and graphics	
Other	10%

From USGS (2011)

Table 16.7 Industrial dispersion of selenium compounds in the U.S. in 2002

	Selenium compounds released to land and water in 2002	
Source (tonnes)	Water	Land
Top five states:		
Alabama	308	301
Nevada	196	195
Utah	166	164
Oregon	82	82
Texas	157	81
Main industries:		
Electric, gas, sanitary services	13	767
Metal/Coal mining	0.1	217
Primary metal industries	0.8	63
Chemicals and allied products	0.002	0.1
Petroleum and coal products	1.2	0.1
Totals	925	1,869

U.S.-EPA (2002)

deficiency in a number of countries and is discussed in Sect. 16.5 of this chapter. Environmental problems related to selenium emissions may also arise in areas surrounding selenium processing or fossil-fuel burning industries. Selenium concentrations in the air within 0.5–10 km of copper-sulfide ore processing plants have been reported to reach 0.15–6.5 μg m^{-3} (WHO 1987).

It is clear that man-made sources of selenium have a major impact upon the selenium cycle; despite this, the natural environment is still a very important source and pathway of selenium in animal and human exposure and requires careful consideration in selenium-related health studies (see also Chap. 4, this volume).

16.2.2 Selenium in Rocks

The most important natural source of selenium in the environment is the rock that makes up the surface of the planet. Selenium is classed as a trace element as average crustal abundances are generally very low (0.05–0.09 mg kg^{-1}; Taylor and McLennan 1985). Average concentrations in magmatic rocks such as granites rarely exceed these values (Table 16.3). Relationships with volcanic rocks are more complicated. Volcanoes are a major source of selenium in the environment and it is estimated that over the history of the Earth, volcanic eruptions account for 0.1 g of selenium for every cm^2 of the Earth's surface. Ash and gas associated with volcanic activity can contain significant quantities of selenium and, for example, values of 6–15 mg kg^{-1} have been reported in volcanic soils on Hawaii and high concentrations are associated with volcanic tuffs in the Prairies of North America. Conversely, because selenium escapes as high-temperature gases during volcanic activity,

Table 16.8 Most common mineral forms of selenium in natural rocks

Selenium–mineral	Chemical formula
Crookesite	$(Cu, Tl, Ag)_2Se$
Clausthalite	PbSe
Berzelianite	Cu_2Se
Tiemannite	HgSe
Elemental selenium	Se
Selenium is also commonly found in the sulfide host minerals	
Pyrite	FeS_2
Chalcopyrite	$CuFeS_2$
Pyrrhotite	FeS
Sphalerite	ZnS
Typical mineral associations with selenium:	
Polymetallic sulfide ores	Se–Hg–As–Sb–Ag–Cu–Zn–Cd–Pb
Copper-pyrite ores	Cu–Ni–Se–Ag–Co
Sandstone-uranium deposits	U–V–Se–Cu–Mo
Gold-silver selenide deposits	Au–Ag–Se

From Fleming (1980), Neal (1995), and Reimann and Caritat (1998)

selenium concentrations left behind in volcanic rocks such as basalts and rhyolites are usually very low (Fleming 1980; Jacobs 1989; Nriagu 1989; Neal 1995). In general terms, sedimentary rocks contain greater concentrations of selenium than igneous rocks, but even so, levels in most limestones and sandstones rarely exceed 0.1 mg kg^{-1} (Neal 1995). Since these major rock types account for most of the Earth's surface, a picture should begin to emerge that selenium-deficient environments are far more widespread than selenium-adequate or selenium-toxic ones. Exceptions to the generally low concentrations occur in particular types of sedimentary rocks and deposits. Selenium is often associated with the clay fraction in sediments and is found in greater concentrations in rocks such as shales (0.06 mg kg^{-1}) than limestones or sandstones. Very high concentrations ($\leq$300 mg kg^{-1}) of selenium have also been reported in some phosphatic rocks, probably reflecting similarities between organically derived PO_4^{3-} and SeO_4^{2-} anions (Fleming 1980; Jacobs 1989; Nriagu 1989; Neal 1995). Selenium concentrations in coal and other organic-rich deposits can be high relative to other rock types and typically range from 1 to 20 mg kg^{-1} (although values of over 600 mg kg^{-1} have been reported in some black shales) with selenium present as organo-selenium compounds, chelated species, or adsorbed element (Jacobs 1989). Selenium is often found in sulfide mineral deposits and has been used as a pathfinder for gold and other precious metals in mineral exploration (Boyle 1979). In most situations, selenium substitutes for sulfur in sulfide minerals due to similarities in crystallography, however, elemental Se^0 is occasionally reported (Fleming 1980; Neal 1995; Tokunaga et al. 1996). The main mineral forms and common mineral associations of selenium are outlined in Table 16.8.

Therefore, the distribution of selenium in the geological environment is highly variable depending on different rock types. An illustration of the relationships between geology and selenium distribution is shown in the map of Wales (Fig. 16.2). The highest selenium concentrations in stream sediment are associated with the mineralized areas of Parys Mountain, the Harlech Dome, and Snowdon in north Wales. In South Wales they are seen in the Forest of Dean and South Wales Coalfields and in the Permian Mercia Mudstone Group of the Welsh borderlands. In contrast concentrations over Devonian age sandstones in mid-Wales are extremely low (<0.4 mg kg^{-1}) (see also Chap. 2, this volume).

16.2.3 Selenium in Soil

In many circumstances there is a strong correlation between the concentration of selenium in geological parent materials and the soils derived from them. The selenium content of most soils is very low at 0.01–2 mg kg^{-1} (world mean is 0.4 mg kg^{-1}), but high concentrations of up to 1,200 mg kg^{-1} have been reported in some seleniferous areas (Table 16.9) (Fleming 1980; Jacobs 1989; Mayland 1994; Neal 1995). The relationships between geology, soil selenium concentrations, uptake into plants, and health outcomes in animals were first examined in detail in pioneering work carried out during the 1930s by Moxon (1937). Soils capable of producing selenium-rich vegetation toxic to livestock were reported over black shale, volcanic tuff and sandstone deposits of the Great Plains in the United States. Subsequent studies into selenium-deficiency-related

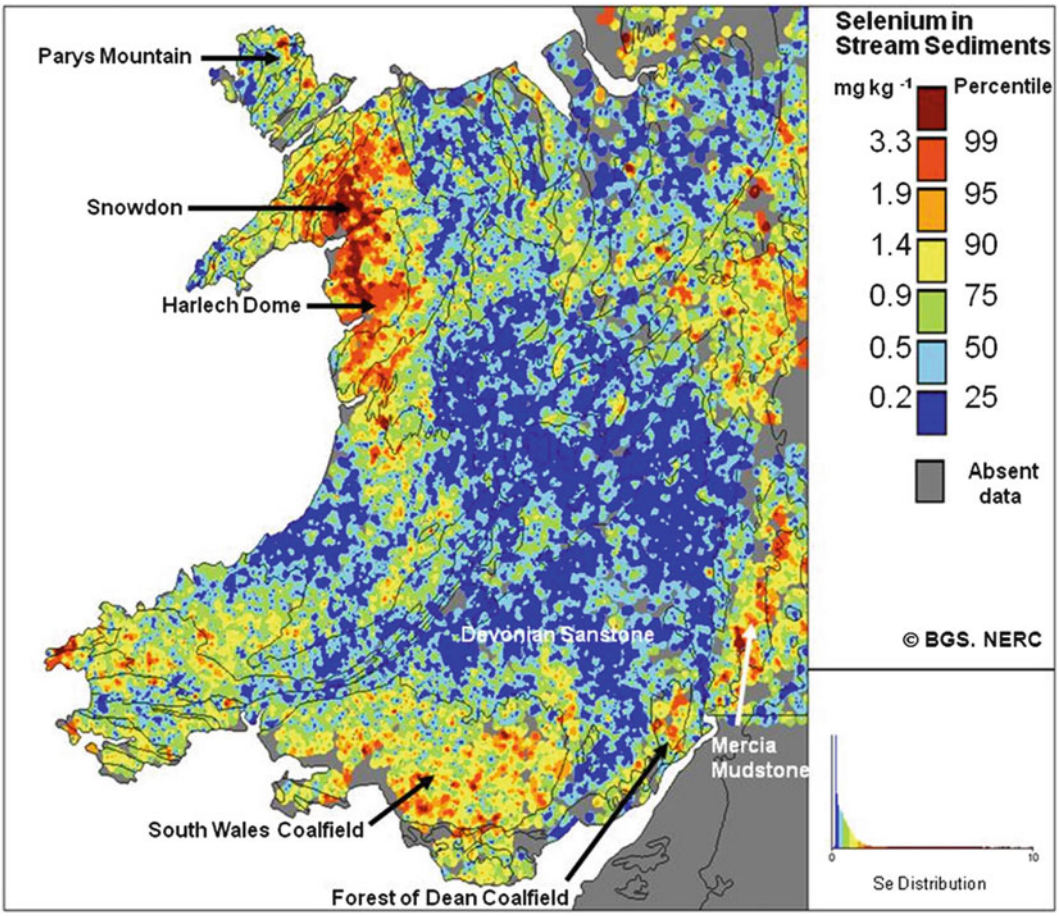


Fig. 16.2 Distribution of selenium in stream sediments of Wales, showing high concentrations over the coalfields of south Wales, the mineralized areas (Parys Mountain, Snowdon, Harlech Dome) of north Wales, and the Mercia Mudstone in the Welsh Borderlands. Contrasting low values occur over the sandstones and siltstones of mid-Wales demonstrating a very strong relationship between selenium in the environment and geology (Adapted with permission from British Geological Survey 2000)

Table 16.9 Some examples of seleniferous soils and geological parent materials

Country	Parent material
U. S.	Cretaceous shale, tuff, Jurassic shales and sandstones, Triassic sandstones
Canada	Cretaceous shales
Colombia	Black slates
Puerto Rico	Volcanic soils
Ireland	Carboniferous shales and limestones
UK	Carboniferous and Ordovician shales and slates
Israel	Cretaceous limestone
South Africa	Cretaceous shales and sandstones
Australia	Cretaceous shales and limestones
Russia	Jurassic sandstones
China	Permian coal and shales

Modified from Fleming (1980)

diseases in animals lead to one of the first maps of the selenium status of soils, vegetation, and animals and the establishment of the classic Great Plain seleniferous soil types (Fig. 16.3) (Muth and Allaway 1963). The organic matter content of soils also determines soil selenium concentrations due to the propensity for selenium to adsorb to organic materials. Several studies have demonstrated that in low selenium environments organic matter is the dominant control on soil Se composition (e.g. Johnson et al. 2000; Ander et al. 2010; Fordyce et al. 2010).

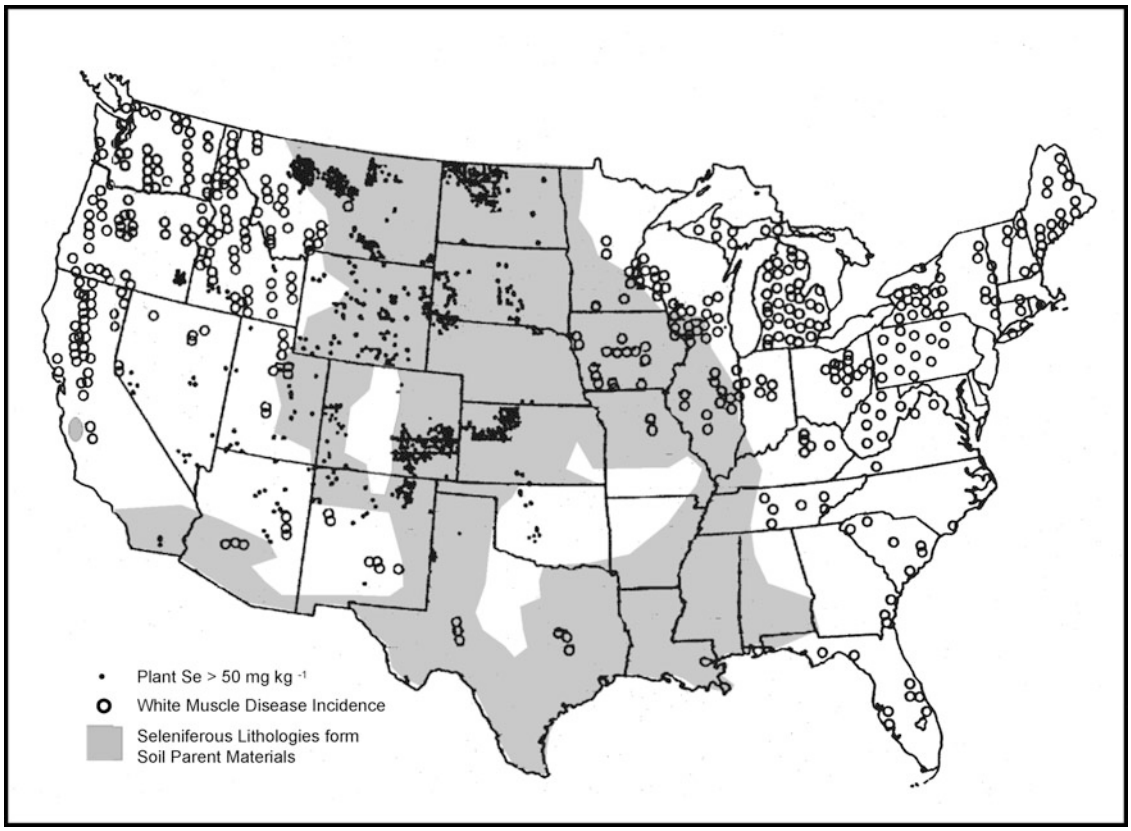


Fig. 16.3 The geographic distribution of selenium-rich soils (*shading*), localities where plant selenium concentrations are known to exceed 50 mg kg^{-1}, and reported incidences of the selenium-deficiency-related disorder WMD in animals in the United States (Adapted with permission from Muth and Allaway 1963)

Although the underlying geology and organic matter content are the primary controls on selenium concentrations in soils, the mobility and uptake of selenium into plants and animals, known as the bioavailability, is determined by a number of bio-physiochemical parameters. These include the prevailing pH and redox conditions, the chemical form or speciation of selenium, soil texture and mineralogy, and the presence of competitive ions. An understanding of these controls is essential to the prediction and remediation of health risks from selenium as even soils that contain adequate total selenium concentrations can result in selenium deficiency if the element is not in readily bioavailable form.

The principal controls on the chemical form of selenium in soils are the pH and redox conditions (Fig. 16.4). Under most natural redox conditions, selenite (Se^{4+}) and selenate (Se^{6+}) are the predominant inorganic phases with selenite the more stable form. Selenite is adsorbed by ligand exchange onto soil particle surfaces with greater affinity than selenate. This process is pH dependent and adsorption increases with decreasing pH. In acid and neutral soils, selenite forms very insoluble iron oxide and oxyhydroxide complexes such as $Fe_2(OH)_4$ SeO_3. The low solubility coupled with stronger adsorption makes selenite less bioavailable than selenate. In contrast, selenate, the most common oxidation state in neutral and alkaline soils, is generally soluble, mobile, and readily available for plant uptake. For example, experiments have shown that addition of selenate to soils results in ten times more plant uptake than addition of the same amount of selenium as selenite (Jacobs 1989; Neal 1995).

Elemental selenium (Se^0), selenides (Se^{2-}), and selenium sulfide salts tend to exist in reducing, acid and organic-rich environments only. The low solubility and oxidation potential of these element species make them largely unavailable to plants and animals. However, the oxidation and reduction of selenium is closely linked to microbial activity, for example, the bacterium *Bacillus megaterium* is known to oxidize elemental selenium to selenite. It is estimated that perhaps 50% of the selenium in some soils may be held in organic compounds; however, few have been isolated and identified. To date, selenomethionine has been extracted from soils and is two to four times more bioavailable to plants than inorganic selenite whereas selenocysteine is less bioavailable than selenomethionine (Jacobs 1989; Mayland 1994; Neal 1995).

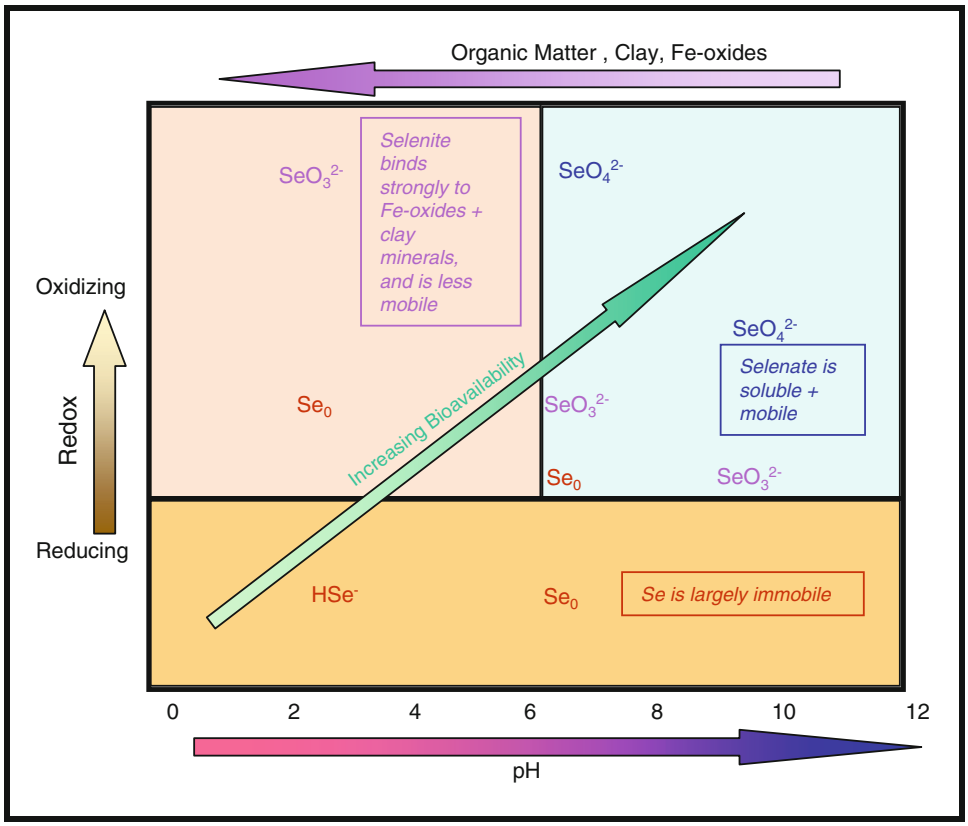


Fig. 16.4 Schematic diagram showing the main controls on the chemical speciation and bioavailability of selenium in soils

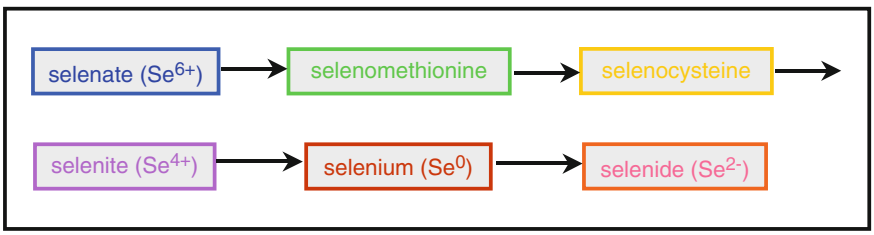


Fig. 16.5 Decreasing order of bioavailability of the different forms of selenium found in soils (Sources: Jacobs 1989; Neal 1995)

The bioavailability of the different selenium species in soils is summarized in Fig. 16.5. In summary, selenate is more mobile, soluble, and less well adsorbed than selenite, thus, selenium is much more bioavailable under oxidizing alkaline conditions and much less bioavailable in reducing acid conditions (Fig. 16.4) (Fleming 1980; Jacobs 1989; Neal 1995) (see also Chap. 15, this volume).

In addition to the speciation of selenium in soils, other soil properties affect mobility. The bioavailability of selenium in soil generally correlates negatively with clay content due to increased adsorption on fine particles; indeed, the selenium uptake in plants grown on clay-loamy soils can be half that of plants grown on sandy soils. Iron also exerts a major control on selenium mobility as both elements are affiliated under oxidizing and reducing conditions and adsorption of selenium by iron oxides exceeds that of clay minerals. As mentioned above, the capacity of clays and iron oxides to adsorb selenium is strongly influenced by pH, reaching a maximum between pH 3–5 and decreasing with increasing pH (Jacobs 1989; Neal 1995). Soil organic matter also has a large capacity to remove selenium from soil solution possibly as a result of fixation by organometallic complexes. For example, plant uptake of selenate added to organic-rich soils can be ten times less than from mineral soils (Jacobs 1989; Neal 1995).

The presence of ions such as SO_4^{2-} and PO_4^{3-} can influence selenium uptake in plants by competing for fixation sites in the soil and plants. SO_4^{2-} inhibits the uptake of selenium by plants and has a greater effect on selenate than selenite. The addition of PO_4^{3-} to soils has been shown to increase selenium uptake by plants as the PO_4^{3-} ion is readily adsorbed in soils and displaces selenite from fixation

Table 16.10 Examples of the three types of selenium accumulating vegetation

Type	Examples (genus, family, or species)
Primary accumulator	G. *Astragalus* (e.g., milk vetch)
	G. *Machaeranthera* (woody aster, U. S.)
	G. *Haplopappus* (North and South American goldenweed)
	G. *Stanleya* (Prince's Plume)
	G. *Morinda* (rubiaceous trees and shrubs, Asia/Australia)
	F. Lecythidaceae (South American trees)
	Sp. *Neptunia* (Legume Asia/Australia)
Secondary accumulator	G. *Aster*
	G. *Astragalus*
	G. *Atriplex* (Saltbush)
	G. *Castilleja* (North and South American perennials)
	G. *Grindelia* (gummy herbs of western North and Central America)
	G. *Gutierrezia* (perennial herbs of western North and South America)
	G. *Machaeranthera*
	G. *Mentzelia* (bristly herbs of western America)
	Sp. *Brassica* (mustard, cabbage, broccoli, cauliflower)
Non-accumulator	Sp. *Pascopyrum* (wheat grass)
	Sp. *Poasecunda* (blue grass)
	Sp. *Xylorhiza* (Woody Aster)
	Sp. *Trifolium* (clover)
	Sp. *Buchloe* (buffalo grass)
	Sp. *Bouteloua* (North and South American tuft grass)
	Sp. *Beta* (sugar beet)
	Sp. *Horedeum* (barley)
	Sp. *Triticum* (wheat)
	Sp. *Avena* (oats)

From Rosenfield and Beath (1964), Jacobs (1989), and Neal (1995)

sites making it more bioavailable. Conversely, increasing the levels of PO_4^{3-} in soils can dilute the selenium content of vegetation by inducing increased plant growth (Jacobs 1989; Mayland 1994; Neal 1995).

Therefore, in any study of the selenium status of soil, consideration of the likely bioavailability is important. Several different chemical techniques are available to assess bioavailability but one of the most widely accepted indicators is the water-soluble selenium content (Jacobs 1989; Tan 1989; Fordyce et al. 2000b). In most soils, only a small proportion of the total selenium is dissolved in solution (0.3–7%) and water-soluble selenium contents are generally <0.1 mg kg^{-1} (Table 16.3) (Jacobs 1989).

The importance of soil selenium bioavailability and health outcomes is exemplified by seleniferous soils in the United States. Toxicity problems in plants and livestock have been reported in soils developed over the Cretaceous shales of the northern mid-West which contain 1–10 mg kg^{-1} total selenium because up to 60% of the element is in water-soluble readily bioavailable form in the semi-arid alkaline environment. In contrast, soils in Hawaii with up to 20 mg kg^{-1} total selenium do not cause problems in vegetation and livestock, because the element is held in iron and aluminum complexes in the humid lateritic soils of that region (Oldfield 1999).

16.2.4 Selenium in Plants

Although there is little evidence that selenium is essential for vegetation growth, it is incorporated into the plant structure. Selenium concentrations in plants generally reflect the levels of selenium in the environment such that the same plant species grown over high and low selenium-available soils will contain concentrations reflecting the soil composition. However, an important factor that may determine whether or not selenium-related health problems manifest in animals and humans is the very wide-ranging ability of different plant species to accumulate selenium (WHO 1987; Jacobs 1989; Neal 1995).

Rosenfield and Beath (1964) were the first to classify plants into three groups on the basis of selenium uptake when grown on seleniferous soils. Some examples of this scheme are outlined in Table 16.10. Selenium accumulator plants grow well on high-selenium soils and can absorb $>1{,}000$ mg kg^{-1} of the element, whereas secondary selenium absorbers rarely concentrate more than 50–100 mg kg^{-1}. The third group, which includes grains and grasses, usually accumulates less than 50 mg kg^{-1} of selenium. Selenium concentration in plants can range from 0.005 mg kg^{-1} in deficient crops to 5,500 mg kg^{-1} in selenium accumulators, but most plants contain <10 mg kg^{-1} selenium. Some

Table 16.11 Relative uptake of selenium in agricultural crops

Selenium accumulation	Plant species
Better accumulators	Cruciferae (broccoli, radish, cress, cabbage, turnip, rape, and mustard)
	Liliaceae (onion)
	Leguminosae (red and white clover, peas)
	Helianthus (sunflower)
	Beta (Swiss chard)
Poorer Accumulators	Compositae (lettuce, daisy, artichoke)
	Gramineae (cocksfoot, ryegrass, wheat, oats, barely)
	Umbelliferae (parsnip, carrot)
Average selenium mg kg^{-1} *dry weight*	*U. S. crop type*
0.407	Roots and bulbs
0.297	Grains
0.110	Leafy vegetables
0.066	Seed vegetables
0.054	Vegetable fruits
0.015	Tree fruits

Jacobs (1989)

species of the plant genera *Astragalus*, *Haplopappus*, and *Stanleya* are characteristic of seleniferous semi-arid environments in the western United States and other parts of the world and are often used as indicators of high-selenium environments. It should be noted, however, that other species in these genera are non-accumulators (WHO 1987; Jacobs 1989; Neal 1995).

The reason why some plants are better at accumulating selenium than others depends upon selenium metabolism. Plants contain many different selenium compounds and the main form in non-accumulator species is protein-bound selenomethionine; however, selenocysteine and selenonium have also been reported (Jacobs 1989; Neal 1995). In contrast, the selenium metabolism in accumulator plants is primarily based on water-soluble, non-protein forms such as Se-methylselenomethionine. The exclusion of selenium from the proteins of accumulator plants is thought to be the basis of selenium tolerance (Jacobs 1989; Neal 1995). Plants also reduce selenate to elemental Se^0 and selenide Se^{2-} forming the volatile organic compounds dimethylselenide and dimethyldiselenide, which are released to the air during respiration giving rise to a "garlic" odor characteristic of selenium-accumulating plants (Mayland 1994).

Despite these coping mechanisms, plants can suffer selenium toxicity via the following processes (Jacobs 1989; Fergusson 1990; Mayland 1994; Wu 1994):

- Selenium competes with essential metabolites for sites in the plant biochemical structure.
- Selenium may replace essential ions, mainly the major cations (for example, iron, manganese, copper, and zinc).
- Selenate can occupy the sites for essential groups such as phosphate and nitrate.
- Selenium can be incorporated into analogues of essential sulfur compounds in plant tissues.

No phytotoxicity symptoms have been reported in nature in the United States, but experimental evidence has shown a negative correlation between increased selenium contents in soil and growth (plant dry weight, root length, and shoot height all decrease). For example, alfalfa yields have been shown to decline when extractable selenium exceeds 500 mg kg^{-1} in soil. Other symptoms include yellowing, black spots, and chlorosis of plant leaves and pink root tissue (Jacobs 1989; Wu 1994). However, phytotoxicity has been reported in nature in China, where high concentrations in soil caused pink discoloration of maize corn-head embryos; the pink color was attributed to the presence of elemental selenium. Levels of >2 mg kg^{-1} and >1.25 mg kg^{-1} selenium were detrimental to the growth and yield of wheat and pea crops, respectively (Yang et al. 1983). In addition to disturbances to the plant metabolism, a more recent study has shown that at low concentrations, selenium acts as an antioxidant in plants inhibiting lipid peroxidation but at high concentrations (additions of >10 mg kg^{-1}), it acts as a pro-oxidant encouraging the accumulation of lipid peroxidation products, which results in marked yield losses (Hartikainen et al. 2000).

Food crops tend to have relatively low tolerance to selenium toxicity, and most crops have the potential to accumulate the element in quantities that are toxic to animals and humans. In general, root crops contain higher selenium concentrations than other plants (Table 16.11) and plant leaves often contain higher concentrations than the tuber. For example, Yang et al. (1983) noted that selenium concentrations in vegetables (0.3–81.4 mg kg^{-1}) were higher than in cereal crops (0.3–28.5 mg kg^{-1} in rice and

maize) in seleniferous regions of China. Turnip greens were particularly high in selenium, with an average of 457 and ranged up to 24,891 mg kg^{-1} compared to an average of 12 mg kg^{-1} in the tuber. In moderate to low selenium environments, alfalfa (Sp. *Medicago*) has been shown to take up more selenium than other forage crops, which may be due to deeper rooting accessing more alkaline conditions, hence more bioavailable selenium at depth. However, in general, crop species grown in very low-selenium soils show little difference in take up and changing the type of plants makes little impact on the selenium content of crops (Jacobs 1989). An exception is reported in New Zealand (Sect. 16.5 of this chapter).

16.2.5 Selenium in Water

It is estimated that the annual global flux of selenium from land to the oceans is 14,000 tonnes $year^{-1}$ via surface and groundwaters, which represent a major pathway of selenium loss from land in the selenium cycle (Nriagu 1989). Approximately 85% of the selenium in most rivers is thought to be in particulate rather than aqueous form, however, the cycling of selenium from the land to the aqueous environment is poorly understood and requires further investigation (Haygarth 1994).

The average concentration of selenium in seawater is estimated at 0.09 μg L^{-1} (Cutter and Bruland 1984), but the mean residence time for selenium is thought to be 70 years in the mixed layer and 1,100 years in the deep ocean, hence, the oceans constitute an important environmental sink for selenium (Haygarth 1994). Biogenic volatilization from seawater to the atmosphere is estimated at 5,000–8,000 tonnes annually (Nriagu 1989) and Amouroux et al. (2001) have demonstrated that the biotransformation of dissolved selenium in seawater during spring blooms of phytoplankton is a major pathway for the production of gaseous selenium emission into the atmosphere. This makes the oceans an important component of the selenium cycle.

Although the oceans via seafood do play a role in human selenium exposure, water used for drinking is more important. Selenium forms a very minor component of most natural waters and rarely exceeds 10 μg L^{-1}. Typically ranges are <0.1–100 μg L^{-1} with most concentrations below 3 μg L^{-1} (Plant et al. 2004). A garlic odor has been noted in waters containing 10–25 μg L^{-1}, whereas waters containing 100–200 μg L^{-1} selenium have an acerbic taste (WHO 1987; Jacobs 1989). In general, groundwaters contain higher selenium concentrations than surface waters due to greater contact times for rock-water interactions (Hem 1992). Groundwaters containing 1,000 μg L^{-1} selenium have been noted in seleniferous aquifers of Montana in the United States and up to 275 μg L^{-1} in China (Jacobs 1989; Fordyce et al. 2000b) (Table 16.3). Although rare in nature, concentrations of up to 2,000 μg L^{-1} selenium have also been reported in saline lake waters in the United States, Venezuela, and Pakistan (Afzal et al. 2000). Anthropogenic sources of selenium can impact surface water quality as a result of atmospheric deposition from fossil fuel combustion, industrial processes, and sewage disposal. For example, concentrations of 400 μg L^{-1} in surface waters have been reported around the nickel-copper smelter at the Sudbury ore deposit in Ontario, Canada (Nriagu 1989) and sewage effluents are known to contain 45–50 μg L^{-1} selenium (Jacobs 1989). Irrigation practices can also affect the amount of selenium in water such as at Kesterson Reservoir in the San Joaquin Valley, California (Jacobs 1989) (See Sect. 16.6 of this chapter).

16.2.6 Atmospheric Selenium

The volatilization of selenium from volcanoes, soil, sediments, the oceans, microorganisms, plants, animals, and industrial activity all contribute to the selenium content of the atmosphere. It is estimated that natural background levels of selenium in non-volcanic areas are very low, around 0.01–1 ng m^{-3}; however, the residency time of selenium can be a matter of weeks, which makes the atmosphere a rapid transport route for selenium in the environment. Volatilization of selenium from the surface of the planet to the atmosphere results from microbial methylation of selenium from soil, plant, and water surfaces and is affected by the availability of selenium, carbon source, oxygen availability, and temperature (Haygarth 1994).

The majority of gaseous selenium is thought to be in dimethylselenide form and it is estimated that terrestrial biogenic sources contribute 1,200 tonnes of selenium per year to the atmosphere. Atmospheric dusts derived from volcanoes and wind erosion of the Earth's surface (180 tonnes $year^{-1}$) and suspended sea salts (550 tonnes $year^{-1}$) from the oceans also constitute significant sources of atmospheric selenium (Nriagu 1989). It is suggested that particle-bound selenium can be transported several thousand kilometers before deposition back to the Earth's surface in both wet and dry forms. Wet deposition is thought to contribute 5,610 tonnes $year^{-1}$ to land (Haygarth 1994). For example, in the UK it has been demonstrated that wet deposition (rain, snow, etc.) accounts for 76–93% of total deposition with >70% of selenium in soluble form. In the proximity of selenium sources (such as industrial emissions), atmospheric deposition can account for 33–82% of uptake in the leaves of plants (Haygarth 1994).

Table 16.12 U.S. inhalation permissible exposure limit for selenium-bearing compounds

Compound	Value
Hydrogen selenide	200 μ gm^{-3}
Se-hexafluoride	400 μ gm^{-3}
Se-compounds	200 μ gm^{-3}

From ATSDR (2003)

16.2.7 Selenium Is All Around Us

From the descriptions above, it is clear that selenium is present in varying quantities in the environment all around us as a result of natural and man-made processes. Animals and humans are exposed to environmental selenium via dermal contact, the inhalation of air, and via ingestion of water and of plants and animals in the diet produced on soils containing selenium.

16.3 Selenium in Animals and Humans

16.3.1 Selenium Exposure

In most non-occupational circumstances atmospheric exposure is insignificant as concentrations of selenium are so low (<10 ng m^{-3}). However, occupational inhalation exposure may occur in the metal, selenium-recovery, and paint industries. In these circumstances, acute (short-term) exposure of humans to hydrogen selenide, the most toxic selenium compound, which exists as a gas at room temperature, results in irritation of the mucous membranes, pulmonary edema, severe bronchitis, and bronchial ammonia whereas inhalation of selenium dust can cause irritation of the membranes in the nose and throat, bronchial spasms, and chemical pneumonia. Selenium dioxide gas is the main source of problems in industrial situations as selenious acid is formed on contact with water or sweat causing irritation. Indigestion and nausea, cardiovascular effects, headaches, dizziness, malaise, and irritation of the eyes have also been reported in occupational selenium exposure (WHO 1987, 1996) (see also Chap. 23 this volume).

As a result of these effects, hydrogen selenide gas is classed as a highly toxic substance and the common selenium-bearing compounds sodium selenite and sodium selenate are considered pollutants of concern (ATSDR 2003). Some regulatory values for selenium compounds in air are presented in Table 16.12. Little information on the long-term (chronic) effects of selenium inhalation is available. In seleniferous areas of China, there is some evidence to suggest that the selenium-loading of the population is enhanced by inhalation of coal smoke from open fires used for cooking as concentrations have been known to rise to 160,000 ng m^{-3} in air during combustion (Yang et al. 1983). However, it is difficult to assess the amount of exposure via this route compared to other sources (Fordyce et al. 1998). Smoking is an inadvertent inhalation exposure route to selenium as tobacco commonly contains 0.03–0.13 mg kg^{-1} (WHO 1987). Assuming a cigarette contains 1 g of tobacco and that all the selenium is inhaled, a person smoking 20 cigarettes could intake 1.6 μg Se day^{-1}. The inhalation of locally grown selenium-rich tobacco in seleniferous regions of China may contribute to the loading of the local population (selenium concentration is 9.05 mg kg^{-1}, Fordyce et al. 1998). In general, however, inhalation is a less important exposure route than ingestion. For example, studies carried out on dogs found that only 52 and 73% of selenium in the form of metal and selenious acid aerosols were adsorbed in the lungs compared to 73 and 96% absorption in the gut (Levander 1986; WHO 1987).

Very few studies have examined the effects of dermal exposure to selenium although sodium selenite and selenium oxychloride solutions have been proved to absorb into the skin of experimental animals. The insoluble compound selenium sulfide is used in anti-dandruff shampoos and is not normally absorbed through the skin, but elevated selenium concentrations in urine have been noted in people with open skin lesions who use these products. In an occupational setting, selenium dioxide gas can result in burns and dermatitis and an allergic body rash. In most normal circumstances, however, dermal contact is not an important exposure route (Levander 1986; WHO 1987, 1996) (see also Chap. 25, this volume).

In the majority of cases, water selenium concentrations are extremely low (<10 μg L^{-1}) and do not constitute a major exposure pathway; however, aquatic life-forms are sensitive to selenium intoxication as soluble forms of selenate and selenite are highly bioavailable and cause reduced reproduction and growth in fish (Jacobs 1989). For this reason, the U.S.-EPA has set acute ecotoxicity thresholds of 258 μg L^{-1} selenite and 417 μg L^{-1} selenate in surface fresh-water (US-EPA 2004). Selenium is a bioaccumulator, which means that plants and animals retain the element in greater concentrations than are present in the environment (Table 16.13) and the element can be bioconcentrated by 200–6,000 times. For example, concentrations in most waters are approximately 1 μg L^{-1} whereas freshwater invertebrates generally contain up to 4 mg kg^{-1} of selenium (Jacobs 1989). Phytoplanktons are efficient accumulators of dissolved selenomethionine and incorporate inorganic selenium into amino acids and proteins (estimated bioconcentration factors range from 100 to 2,600) (Jacobs 1989). However, the reported lethal doses of selenium in water for invertebrates (0.34–42 mg L^{-1}) and fish (0.62–28.5 mg L^{-1}) indicate that in most circumstances water alone is not a major environmental problem. It should be noted, however, that inorganic and organic selenium enter

Table 16.13 Average concentrations of selenium in selected animals

Animal	Selenium mg kg^{-1}
Fish, U. S.	0.5 wet weight
Terrestrial arthropods	1 fresh weight
Earthworms	2.2 (normal soil)–22 (sewage sludge amended soil) fresh weight
Bird livers	4–10 dry weight
Bird eggs	0.4–0.8 wet weight
Bird kidneys	1–3 dry weight
Mammal livers	<2 dry weight

From Jacobs (1989)

the food chain almost entirely via plants and algae and bioconcentration from high-selenium waters could cause problems, because selenium passes up the food chain from algae and larval fish to large fish, birds, and humans (WHO 1987; Jacobs 1989). A chronic aquatic life criterion of 7.9 mg kg^{-1} (dry weight) selenium in fish tissue has been proposed in the United States to prevent toxicity and the uptake of too much selenium into the food chain (U.S.-EPA 2004). At concentrations >50 μg L^{-1} in water, selenium intake can contribute significantly to overall dietary intake in animals and humans and the U.S.-EPA (2012) currently recommends this as the MAC for selenium in drinking water. The World Health Organization currently sets a MAC of 40 μg L^{-1} selenium for drinking water (WHO 2011)

However, the most important exposure route to selenium for animals and humans is the food we eat, as concentrations are orders of magnitude greater than in water and air in most circumstances (WHO 1996). An excellent review of the selenium content in foodstuffs is provided by Reilly (2006). In summary, in terms of the human diet, organ meats such as liver and kidney are good sources of selenium and some seafoods contain almost as much. Muscle meats are also a significant source and garlic and mushrooms contain more than most other vegetables. Cereals are another important source, however, white bread and flour contain less selenium than whole meal by about 10–30% (Table 16.14) (WHO 1987; Reilly 2006; Rayman 2008). Brazil nuts sold in the UK are high in selenium, indeed, cases of selenium poisoning in Amazon peoples following consumption of nuts of the Lecythidaceae family have been reported in Brazil. These incidents resulted in nausea, vomiting, chills, diarrhea, hair and nail loss, painful joints, and death in some cases (see Sect. 16.3.6 of this chapter). Cooking reduces the selenium contents of most foods, and studies have shown that vegetables that are normally high in selenium such as asparagus and mushrooms lose 40% during boiling. Other studies estimate 50% of the selenium content is lost from vegetables and dairy products during cooking especially if salt and low pH foods such as vinegar are added, whereas frying foods results in much smaller losses (Levander 1986; WHO 1987, 1996) (see also Chap. 8, this volume).

Levels of dietary selenium intake show huge geographic variation and are dependent upon the geochemical conditions of the food source environments as well as differences in dietary composition. For example in 1995, cereals accounted for 75% of the total 149 μg day^{-1} selenium intake in Canada but only 10% of the 30 μg day^{-1} intake in Finland (WHO 1996). In general terms, cereals grown in North America contain more selenium than European crops and concern is growing in Europe over declining selenium intakes. The UK traditionally imported large quantities of wheat from North America but since the advent of the EU, most cereals are now more locally derived and as a consequence, daily intakes of selenium in the UK have been falling (Broadley et al. 2006). Marked declines are evident particularly over a 4-year period from intakes of 60 μg day^{-1} in 1991 to 43 μg day^{-1} in 1995 (Fig. 16.6) (FSA 2009). This downward trend is also attributed to a reduction in cereal consumption in the UK, which fell from 1,080 g $person^{-1}$ $week^{-1}$ in 1970 to 756 g $person^{-1}$ $week^{-1}$ in 1995 (MAFF 1997) and changes in the bread making process (Johnson et al. 2010). The selenium content of Irish bread is also significantly lower than in the United States and only marginally higher than the UK (Table 16.14) (Murphy and Cashman 2001). Other cereal crops such as rice generally contain low selenium contents (Table 16.14) and can have a significant influence on overall dietary intake when consumed as the staple food as in most of Asia. Conversely Japanese diets can be very high in selenium (up to 500 μg day^{-1}) in areas where a large amount of seafood is consumed (WHO 1987).

Some examples of daily dietary selenium intakes from around the world are listed in Table 16.15. On a global scale it is estimated that dietary intakes in adults range from 3 to 7,000 μg day^{-1} and for infants in the first month of life from 5 to 55 μg day^{-1}. The wide ranges are attributed to selenium contents in the environment (WHO 1996; Rayman 2008). The greatest variations in dietary intake are reported from selenium-deficient and seleniferous regions of China (Tan 1989), but contrasts also occur in South America between high daily intakes (100–1,200 μg) associated with foodstuffs grown on selenium-rich shales in the Andes and Orinoco

Table 16.14 Concentrations of selenium in selected foodstuffs from around the world

Food type	Source	Selenium mg kg^{-1}
Whole meal flour	Ireland	0.077–0.099
White flour	Ireland	0.060–0.069
Wheat flour	Russia	0.044–0.557
Whole meal bread	Ireland	0.086–0.129
White bread	Ireland	0.066
Wheat	World	0.1–1.9
Wheat	Greece	0.019–0.528
Wheat	Colombia	180
Wheat	China	Deficient 0.001–0.105
Wheat	Scotland	0.003–0.006 fresh weight
Wheat	India	Seleniferous 0.742
Wheat	India	Non-seleniferous 0.010
Barley	U. S.	0.2–1.8
Oats	U. S.	0.15–1
Corn	China	Deficient 0.005–0.089
Corn	China	Seleniferous 0.5–28.5
Corn	Venezuela	14
Maize	China	Deficient 0.001–0.105
Maize	China	Seleniferous 0.017–9.175
Maize	China	Adequate 0.021–2.324
Maize	U. S.	0.136
Maize	Malawi	0.005–0.533 fresh weight
Rice	Venezuela	18
Rice	China	Deficient 0.007–0.022
Rice	China	Seleniferous 0.3–20.2
Rice	Sri Lanka	0.0001–0.777
Cereals	World	0.1–0.8 wet weight
Cereals	Finland/New Zealand	0.01–0.07
Cereals	UK	0.070 fresh weight
Bread	UK	0.060 fresh weight
Liver, kidney, seafood	World	0.4–1.5 wet weight
Liver, kidney, seafood	Finland/New Zealand	0.09–0.92 wet weight
Fish	UK	0.420 fresh weight
Muscle meat	World	0.1–0.4 wet weight
Muscle meat	Finland/New Zealand	0.01–0.06 wet weight
Beef steak	Scotland	0.081–1.51 fresh weight
Offal	UK	0.770 fresh weight
Meat	UK	0.140 fresh weight
Dairy products	World	0.1–0.3 wet weight
Dairy products	Finland/New Zealand	0.01 wet weight
Dairy products	UK	0.030 fresh weight
Cow's milk	Turkey	11.28–36.05mgL^{-1}
Cow's milk	Scotland	0.001–0.022 fresh weight
Cow's milk	UK	0.014 fresh weight
Cow's milk	India	Seleniferous 0.050 mg L^{-1}
Cow's milk	India	Non-seleniferous 0.006 mg L^{-1}
Human milk	World	0.013–0.018 mg L^{-1}
Human milk	New Zealand	0.005 mg L^{-1}
Dried milk	Russia	0.038–0.115
Egg whites	Chile	0.55–1.10
Eggs	UK	0.019 fresh weight
Fruit and vegetables	World	0.1 wet weight
Fruit and vegetables	Finland/New Zealand	0.01–0.07 wet weight

(continued)

Table 16.14 (continued)

Food type	Source	Selenium mg kg^{-1}
Fruit	UK	<0.005 fresh weight
Broccoli	Scotland	0.001–0.007 fresh weight
Vegetables	China	Seleniferous 2.0–475
Vegetables	UK	<0.010–0.018 fresh weight
Brazil nuts	UK	22.3–53
Nuts	UK	0.300 fresh weight
Soyabeans	China	Deficient 0.010
Soyabeans	China	Seleniferous 0.34–22.2
Field peas	Canada	<0.05–79.6

From Yang et al. (1983), Levander (1986), WHO (1987, 1996), Jacobs (1989), Tan (1989), Fordyce et al. (1998, 2000a), Oldfield (1999), Murphy and Cashman (2001), Hira et al. (2004), FSA (2009), Gawalko et al. (2009), Fordyce et al. (2010), Chilimba et al. (2011)

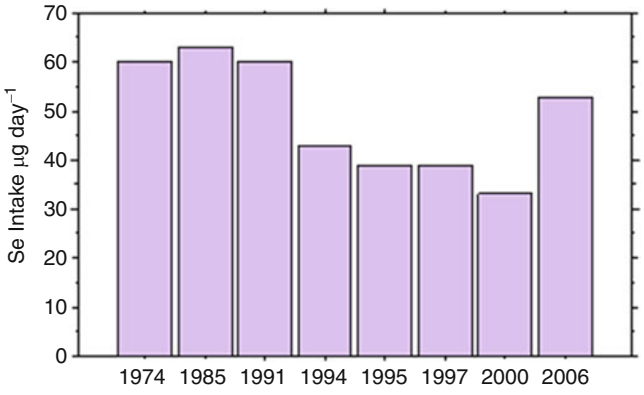


Fig. 16.6 Bar chart showing the decline in the average daily dietary intake of selenium in UK adults 1974–2006 (Data Source: FSA 2009)

River of Venezuela and widespread selenium deficiency in Argentina (WHO 1987; Oldfield 1999; Rayman 2008). Dietary intakes in countries such as New Zealand, Finland, and Turkey are also poor as a consequence of low-selenium soils, whereas intakes in Greece, Canada, and the United States are generally adequate (WHO 1987). On the basis of selenium requirement studies, a range of 50–200 µg day^{-1} has been recommended by the U.S. National Research Council (NRC) for adults depending on various factors such as physiological status. Balance studies to more precisely determine the ratio of selenium inputs and outputs in human beings were attempted, however, these were not successful as humans have the ability to modify fecal and urinary excretion of selenium depending on levels of intake (WHO 1996). Current recommended daily allowances (RDAs) of dietary selenium range from approximately 55 µg in women to 75 µg day^{-1} in men and 8.7–10 µg day^{-1} in infants (Table 16.15) (WHO 1996; ATSDR 2003; MAFF 1997).

Just as bioavailability is an important factor in terms of plant uptake of selenium, it is also an important factor in the diets of animals and humans. Dietary studies have shown that selenomethionine is more readily absorbed in the guts of animals and humans than selenate, selenite, or selenocysteine. More than 90% of ingested selenomethionine and selenate is absorbed, whereas the rate for sodium selenite is slightly lower (>80%). Selenides and elemental selenium are poorly absorbed and in general organic forms of selenium are more readily bioavailable than inorganic forms (Lyons et al. 2007; Rayman et al. 2008). As an indication of the diet in general, studies carried out in New Zealand have shown that 79% of selenium present in natural foods is bioavailable. In addition to foodstuffs, mineral supplements are a source of dietary selenium to humans and animals. Chemical supplement tests show 97% adsorption of selenomethionine, 94% of selenate, and 60% of sodium selenite in this dietary form (Levander 1986; WHO 1987).

In animals, 85–100% of dietary plant selenium is absorbed whereas only 20–50% of the selenium present in meat and fish is taken up by birds and mammals. In general terms, selenium in plant forms is more readily bioavailable than selenium in animal forms (Levander 1986; WHO 1987, 1996) (see also Chap. 21, this volume).

Until recently, very few studies had examined the chemical form of selenium in foodstuffs, but advances in analytical techniques have revealed more information. Previous evidence suggested that 7.6–44% of selenium in tuna was in the form of selenate – with the remainder present as selenite and selenide (WHO 1987). However, it has now been demonstrated that the main organo-selenium compound in fish such as tuna and mackerel is selenoneine (Yamashita and Yamashita 2010). In contrast, 50% of the selenium in wheat and 15% in cabbage had been shown to be in the form of selenomethionine (WHO 1987). More recent studies confirm that selenomethionine is the main compound in plant food sources particularly cereal crops and yeast and it is the main selenium constituent in Brazil nuts and some selenium supplements. Se-methylselenocysteine and γ-glutamyl-Se-methylselenocysteine are found in yeast, garlic, onions and broccoli; whereas selenocysteine is the main form in animal foodstuffs. Some selenate is found in fish and plant

Table 16.15 Some daily dietary intakes of selenium from around the world (μg day^{-1})

Country	Year	Vegetables and fruit	Cereals	Dairy	Meat and fish	Men	Women	Total
New Zealand (general)	1981–1982	1–2	3–4	11	12–16			28–32
New Zealand infants	1987							0.5–2.1
Finland	1975–1979	1	3–25	7–13	19			30–60
UK	1978	3	30	5	22			60
UK	1991							43
UK	1995							29–39
UK	2000							32–34
UK	2006							48–58
Japan (general)	1975	6	24	2	56			88
Japan high seafood	1987							500
Canada	1975	1–9	62–133	5–28	25–90			98–224
Canada	1987							149
U.S. (general)	1974–1976	5	45	13	69			132
U.S. (general)	2002							71–152
U.S. South Dakota	1976	10	57	48	101			216
U.S. Maryland	1987							81
China (general)	1987			2–212				
China seleniferous	1983							240–6,690
China Se-adequate	1983					19.1	13.3	42–232
China Se-deficient	1983					7.7	6.6	3–22
India, Mumbai	2001							61.9
India seleniferous	1998							475–632
India non-seleniferous	1998							52–65
Turkey	2001							20–53
Venezuela	1999			58				100–1,200
Swedish Pensioners	1987							8.7–96.3
Greece	1999							110

Country	RDA range	RDA W	RDA M	RDA I
U.S.	50–200	55	55	15–20
UK	60–200	60	75	10
China	40–600			

From Yang et al. (1983), WHO (1987, 1996), Tan (1989), MAFF (1997), Oldfield (1999), Aras et al. (2001), Mahapatra et al. (2001), ATSDR (2003), Hira et al. (2004), FSA (2009)

Note: RDA recommended daily allowance, *W* women, *M* men, *I* infants

sources such as cabbage (Lyons et al. 2007; Rayman 2012). These differences in the chemical forms of selenium are reflected in the rate of absorption and bioavailability of the element in foodstuffs. For example, it is estimated that over 90% of the selenium in Brazil nuts and beef kidney is bioavailable, compared to only 20–60% in tuna. However, other seafood, such as shrimp, crab, and Baltic herring, have higher bioavailability (Levander 1986; WHO 1987).

The bioavailability of selenium to humans and animals is not only dependent on the amount of absorption but also on the conversion of the ingested selenium to metabolically active forms. In humans, studies based on the activity of the selenium-dependent enzyme GPx have shown that the bioavailability of selenium in wheat is >80% whereas the bioavailability of selenium in mushrooms is very low (WHO 1987). Conversely, other studies report good bioavailability in selenium-enhanced mushrooms (Rayman et al. 2008). In a comparison between wheat and mineral selenate supplements, while the latter were shown to enhance GPx activity, patients fed wheat demonstrated greater increases and better long-term retention of selenium (WHO 1987).

Fairweather-Tait et al. (2010) conclude that the bioavailability of selenium is difficult to quantify due to the complexity of selenium compounds in foodstuffs. Therefore, much has still to be learned about the uptake of selenium in humans and animals; however, it is clear that in most normal circumstances food forms the major exposure route as selenium accumulates from the environment via plants and algae through the food chain to animals and man. Selenium in the form of selenomethionine and other organic compounds is highly bioavailable to animals and humans and foodstuffs that contain high proportions of these forms, such as organ and muscle meats, Brazil nuts, and wheat, are good sources of the element in the diet (WHO 1987; Reilly 2006; Rayman 2012).

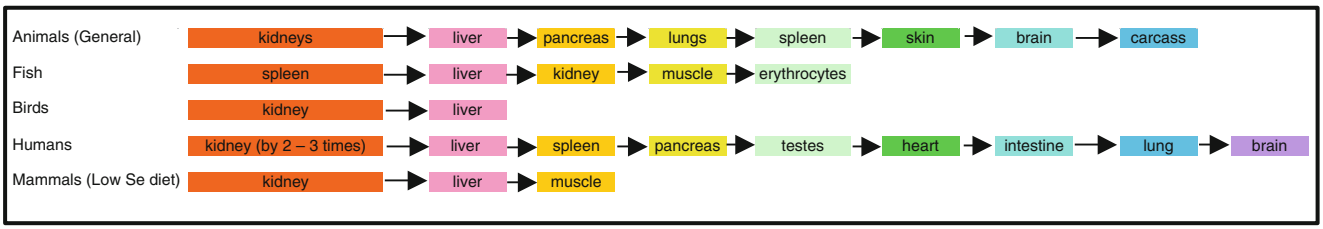


Fig. 16.7 Relative concentrations of selenium in the tissues of different animals, decreasing in concentration from *left* to *right* (Data Sources: WHO 1987; 1996; Levander 1986; Jacobs 1989)

16.3.2 Selenium in the Body of Animals and Humans

Once ingested into the body, most selenium is absorbed in the small intestine of animals and humans, but the rates and mechanisms of selenium metabolism vary between different animal species. In general, single-stomached animals absorb more selenium than ruminants due to the reduction of selenite to insoluble forms by rumen microorganisms. Experiments on rats indicate very little difference in the process of absorption of different selenium forms; 92% of selenite, 91% of selenomethionine, and 81% of selenocysteine were absorbed primarily in the small intestine and none in the stomach. Approximately 95% of the total selenium intake was absorbed regardless of whether the rats were fed a low- or high-selenium diet indicating that selenium intake is not under homeostatic control. This is true in general for intake in animals and humans. However, other studies have shown that oral doses of selenomethionine are retained more readily and turned over more slowly than selenite in humans, therefore unlike rats there is a difference in the metabolism of different forms of selenium. In fact, selenomethionine, the main form of uptake from plants to animals, becomes associated with protein tissues in the body whereas inorganic selenium is absorbed into other tissues (Levander 1986; WHO 1987, 1996; Rayman et al. 2008). Selenite is taken up by red blood cells, bound to albumin and transferred to the liver. Selenate is taken up directly by the liver or excreted into the urine (Lyons et al. 2007).

Indeed, most of the ingested selenium is quickly excreted in the urine, breath, perspiration, and bile and the remainder becomes bound or incorporated into blood and proteins. Urine is the primary route of excretion (70–80%) in single-stomached animals, however, in ruminants selenium is mostly excreted in the feces and studies have shown that the majority of this selenium is in unavailable elemental form. Chemical selenium tracer experiments in humans suggest that the main extraction pathway is via urine, however, in studies using natural foods, excretion in feces was equal to that of urine; whereas minimal amounts of selenium were exuded in sweat and respiration and expulsion of volatile forms of selenium only occurred at very high exposures. Unlike selenium absorption, which is not homeostatically regulated, selenium excretion in animals and humans is directly influenced by nutritional status: excretion rises as intake increases and decreases when selenium intakes are low (Levander 1986; WHO 1987, 1996) (see also Chaps. 8, 9, and 21, this volume).

The remaining selenium is transported rapidly around the body and concentrates in the internal organs, which are rich in protein. This pattern is present in a number of animal species (Fig. 16.7). Hence, in normal conditions in humans, selenium levels are highest in the liver and kidneys and lower in muscle tissues. However, the largest total amount of selenium in the human body is in the muscles as these form the main body mass. It is estimated that muscles account for 46.9% whereas kidneys contain only 4% of the total selenium in the human body (Lyons et al. 2007). Total human body selenium contents are estimated at 3–14.6 mg (WHO 1987).

In rats fed selenium-deficient diets, however, the pattern of selenium distribution is different with selenium reserved in the testes, brain, thymus, and spleen. Also in humans the supply to the testes has priority over the other tissues during selenium deficiency, because the element is found in the mitochondrial capsule protein (MCP) and is involved in biosynthesis of testosterone. Consequently, the selenium content of the testes increases considerably during puberty (Levander 1986; WHO 1987).

Both inorganic and organic selenium are converted by animals and humans to mono-, di-, or trimethylated forms by the main metabolic pathway, rarely reduction. Trimethylselenonium, was thought to be the main urinary excretion form, but recent research suggests the selenosugar one metabolite may predominate following increased selenium intake (Lyons et al. 2007). However, in cases of selenium toxicity, the urinary pathway becomes overloaded and the volatile selenium metabolite dimethylselenide is produced and exhaled via the lungs, which results in the characteristic "garlic breath" symptom in animals and humans suffering selenosis. There is much debate over the form of selenium held in protein tissues. Non-ruminant animals and humans

cannot synthesize selenite into selenomethionine, but there is evidence to suggest that selenomethionine from food sources can be incorporated into protein tissues directly in place of methionine (Lyons et al. 2007; Rayman et al. 2008). However, in the case of rats it is then converted to selenite or selenate. Rabbits and rats can also convert selenite into selenocysteine tissue proteins. Selenium may also be present in proteins in the selenotrisulfide and acid-labile form. Early work suggested that selenium intake in all naturally occurring organic forms was retained in tissues to a greater extent than inorganic forms, however, experiments with mice using selenite, Se-methylselenocysteine, and selenomethionine showed that mice fed selenomethionine had greater quantities and better long-term retention of selenium than those fed selenite or Se-methylselenocysteine. There is evidence for metabolic pools of selenium in animals and humans. For example, studies with ewes fed selenium-adequate and then selenium-deficient diets showed that they were able to pass on adequate levels of selenium to their lambs even though the lambs were born 10 months into the selenium-deficient diet. Two likely metabolic pools have been proposed. The main exchangeable pool includes all forms of Se derived from inorganic selenite and selenate and is an active pool for selenoprotein synthesis. The second pool comprises the sequestration of selenomethionine or other selenoamino acids incorporated into protein structures that are then released during protein turnover (Levander 1986; WHO 1987; Lyons et al. 2007; Rayman et al. 2008).

In most circumstances there is a close correlation between the levels of selenium in the diet of humans and animals and blood selenium content. On average, plasma levels vary from 0.079 to 0.252 mg L^{-1} depending on selenium intake, whereas the mean concentration of selenium in human whole blood is 0.2 mg L^{-1} (WHO 1987, 1996). Human whole blood selenium levels show marked geographic variation depending on dietary intake. Ranges of 0.021–3.2 mg L^{-1} have been reported worldwide with highest concentrations in seleniferous areas of China and Venezuela and lowest concentrations in the selenium-deficient regions of Scandinavia, New Zealand, and China (Table 16.16) (WHO 1996; Oldfield 1999). Similarly, concentrations in hair, nails, and urine vary according to differences in dietary intake, and some examples of the selenium composition of these tissues are given in Table 16.16. Selenium levels in human milk are affected by maternal intake and infants and young children have a high requirement for the element during the rapid growth periods of early life. However, the age of mother and the concentration of selenium during pregnancy do not affect the weight of baby or the length of pregnancy. Wide ranges of 2.6–283 mg L^{-1} in human milk have been reported from selenium-deficient and seleniferous regions in China, compared to ranges of 7–33 mg L^{-1} in the United States (Levander 1986).

In terms of biological function, 25 essential selenoproteins containing selenocysteine have now been identified in humans, many of which are involved in redox reactions acting as components of the catalytic cycle (WHO 1996; Papp et al. 2007; Rayman 2012). A selection of selenoproteins with known health functions is listed in Table 16.17 (Papp et al. 2007; Rayman 2012). Enzyme activity is attributed to the glutathione peroxidase, thioredoxin reductase, iodothyronine deiodinase, and selenophosphate synthetase groups. In complex interactions with vitamin E and polyunsaturated fatty acids, selenium plays an essential biological role as part of the enzyme GPx, which protects tissues against peroxidative damage by catalyzing the reduction of lipid hydrogen peroxide or organic hydroperoxides. Together, GPx, vitamin E and superoxide dismuthases form one of the main antioxidant defense systems in humans and animals. Selenium also appears to enhance T-cell production and natural killer cell activity (Rayman 2012). As such, selenium has been linked to enzyme activation, immune system function, pancreatic function, DNA repair, and the detoxification of xenobiotic agents such as paraquat, however, the exact mechanisms of detoxification are still being established (Combs and Combs 1986; Levander 1986; WHO 1987, 1996; Papp et al. 2007). Selenium is found in the prosthetic groups of several metalloenzymes and appears to protect animals against the toxic effects of arsenic, cadmium, copper, mercury, tellurium, and thallium in most circumstances, but this is not always the case and the biological response depends on the ratio of selenium/metal involved (WHO 1987; Fergusson 1990). Selenium behaves antagonistically with copper and sulfur in humans and animals inhibiting the uptake and function of these elements. Selenium also appears to affect the P-450 cytochrome enzyme system in humans and animals. However, the exact biochemical function of selenium is still being established. Several studies have shown no reduction in P-450 activity under selenium-deficient conditions whereas recent work suggests that selenium inhibits P-450 enzymes. Since these enzymes activate carcinogenic substances such as polynuclear aromatic hydrocarbons (PAHs) and phenobarbital, this may explain why selenium has been shown to prevent cancers induced by these substances in laboratory animals (Shimada et al. 2011). Other important developments in more recent years have shown that selenium is beneficial to the thyroid hormone metabolism. There are three iodothyronine deiodinase (Dio) selenoenzymes. Types 1 and 2 are involved in the synthesis of active 3, 3′ and 5-triiodothyronine (T3) hormones, whereas type 3 Dio catalyzes the conversion of thyroxine (T4) to inactive T3(rT3). These hormones exert a major influence on cellular differentiation, growth, and development, especially in the fetus and child (Arthur and Beckett 1994). Selenium also appears to be important in reproduction. In addition to aiding the biosynthesis of testosterone

Table 16.16 Examples of selenium concentrations in human tissues from around the world

	Selenium (mg L^{-1})		Selenium (mg L^{-1})	
Country	Whole blood	Year	Serum	Year
Average (humans)	0.2			
Normal (humans)			0.06–0.105	
Canada, Ontario	0.182	1967		
China, high Se	1.3–7.5	1983		
China, high Se, no disease	0.44	1983		
China, mod Se	0.095	1983		
China, low Se, no disease	0.027	1983		
China, low Se, disease	0.021	1983		
Tibet, low Se			<0.005	1998
Egypt	0.068	1972		
Finland	0.056–0.081	1977		
Guatemala	0.23	1967		
New Zealand	0.083–0.059	1979		
Sweden		1987	0.86	
UK	0.32	1963		
UK		2001	0.087–0.088	2004 (Plasma)
U.S.	0.256–0.157	1968		
Russia	0.11–0.442	1976		
Venezuela seleniferous	0.355–0.813	1972		
Bulgaria			0.0548	1998
Hungary			0.0558	1998
Slovenia			0.0570	1998
Croatia			0.0642	1998
Russia			0.0718	1999
Italy, Lombardy	0.04–0.19	1986	0.033–0.121	1986
Spain, Barcelona			0.060–0.106	1995
Canary Islands			0.008–0.182	2001

	Selenium (mg kg^{-1})		Selenium (μ g L^{-1})
	Hair	Year	Urine
China, Se deficient	0.074	1983	0.007
China, Se deficient	0.170–0.853	1998	
China, Se deficient	0.094–0.359	1996	
China, low Se	0.16	1983	
China, Se adequate	0.343	1983	0.026
China, high Se	1.9–100	1983	0.04–6.63
China, high Se	0.566–141	1998	
Italy, Lombardy		1986	0.0002–0.068
Sri Lanka	0.104–2.551	1998	
India seleniferous	2.31–2.55	1998	0.170–0.267
India non seleniferous	0.048–0.050	1998	0.009–0.012

Yang et al. (1983), Levander (1986), Akesson and Steen (1987), WHO (1987, 1996), Oldfield (1999), Fordyce et al. (2000a, b), Vinceti et al. (2000), Romero et al. (2001), Hira et al. (2004), Ruston et al. (2004)

(see above), the selenium contents of avian eggs are high whereas morphological deformities, immotility, and reduced fertility have been reported in sperm in selenium-deficient experimental animals (WHO 1987, 1996; Rayman 2012). Although many of the *in vivo* functions of selenium are still poorly understood, deficient and excessive dietary intakes of selenium have a marked effect on animal and human health, some of which are discussed below.

16.3.3 Selenium Deficiency—Effects in Animals

Due to the complementary role of selenium and Vitamin E, all selenium deficiency diseases in animals are concordant with vitamin E deficiency with the exception of neutrophil microbicidal activity reduction and the 5-deiodinase enzymes responsible for the production of triiodothyronine from thyroxine. Selenium is necessary for growth and

Table 16.17 List of main known human selenoproteins

Name	Function
Glutathionine peroxidase GPx	Antioxidant enzymes
Sperm nuclei selenoprotein GPx4	Essential for male fertility and sperm maturation
Mitochondrial capsule selenoprotein PHGPx	Protects sperm cells from oxidative damage
Spermatid selenoprotein 34 kDa	May protect developing sperm
Iodothyronine deiodinases (Dio)	Regulation and production of active thyroid hormones
Thioredoxin reductases (TrxR)	Reduction of nucleotides and binding of transcription factors in DNA
Selenophosphate synthetase SPS2	Required for selenoprotein synthesis
Selenoprotein P	Protects endothelial cells against perioxynitrite; transport of selenium; possible heavy metal chelator
Selenoprotein W	Skeletal and heart muscle metabolism
Selenoprotein 15	May affect glycoprotein folding and regulate cell apoptosis
Selenoprotein 18-kDa	Found in the kidney
Selenoprotein R: Methionine sulfoxide reductase	May regulate lifespan
Selenoprotein N	May regulate calcium in early muscle development; if deficient may be linked to muscular dystrophy and multiminicore disease
Selenoprotein M	May have a role in Alzheimer's disease
Selenoprotein S (SEPS1)	Anti-inflammatory; linked to glucose metabolism
Selenoprotein K	May be an antioxidant in the heart

From Papp et al. (2007), Rayman (2012)

fertility in animals and clinical signs of deficiency include dietary hepatic apoptosis in rats and pigs; exudative diathesis, embryonic mortality, and pancreatic fibrosis in birds; nutritional muscular dystrophy, known as white muscle disease, and retained placenta in ruminants and other species; and mulberry heart disease in pigs. Clinical signs of selenium deficiency in animals include reduced appetite, growth, production and reproductive fertility, unthriftyness, and muscle weakness (Levander 1986; WHO 1987, 1996; Oldfield 1999).

White muscle disease is a complex condition that is multifactorial in origin and causes degeneration and apoptosis of the muscles in a host of animal species. This disease rarely affects adult animals but can affect young animals from birth. In lambs born with the disease, death can result after a few days. If the disease manifests slightly later in life, animals have a stiff and stilted gait, arched back, are not inclined to move about, lose condition, become prostrate and die. The disease responds to a combination of vitamin E and selenium supplementation (Levander 1986; WHO 1987, 1996; Oldfield 1999).

Exudative diathesis in birds leads to massive hemorrhages beneath the skin as a result of abnormal permeability of the capillary walls and accumulation of fluid throughout the body. Chicks are most commonly affected between 3 and 6 weeks of age and become dejected, lose condition, show leg weakness, and may become prostrate and die. The disease responds to either vitamin E or selenium supplementation, but it will not respond to vitamin E alone if selenium is deficient (WHO 1987).

Hepatic apoptosis in pigs generally occurs at 3–15 weeks of age and is characterized by necrotic liver lesions. Supplements of alpha-tocopherol and selenium can protect against death (Levander 1986).

Low-selenium pastures containing 0.008–0.030 mg kg^{-1} are associated with a condition called "ill thrift" in lambs and cattle from New Zealand. The disease is characterized by subclinical growth deficits, clinical unthriftiness, rapid weight loss, and sometimes death but can be prevented by selenium supplementation with marked increases in growth and wool yields (Levander 1986; WHO 1987).

The level of dietary selenium needed to prevent deficiency depends on the vitamin E status and species of the host. For example, chicks receiving 100 mg of vitamin E require 0.01 mg kg^{-1} of selenium to protect against deficiency, whereas chicks deficient in vitamin E require 0.05 mg kg^{-1} of selenium. Under normal vitamin E status, concentrations of 0.04–0.1 mg kg^{-1} (dry weight) in feedstuffs are generally adequate for most animals with a range of 0.15–0.20 mg kg^{-1} for poultry and 0.03–0.05 mg kg^{-1} for ruminants and pigs (Levander 1986; WHO 1987).

Selenium deficiency and WMD are known to occur in sheep when blood selenium levels fall below 50 μg L^{-1} and kidney concentrations below 0.21 mg kg^{-1} (dry weight). Blood levels of 100 μg L^{-1} selenium are needed in sheep and cattle and 180–230 μg L^{-1} in pigs to maintain the immunoresponse systems. Studies have shown that most farmland grazing in the UK is not able to provide enough selenium to support 0.075 mg L^{-1} in blood in cattle. Indeed, selenium deficiency in animals is very common and

Fig. 16.8 Distribution of the incidence of KD in China (Adapted with permission from Tan 1989)

widespread around the globe affecting much of South America, North America, Africa, Europe, Asia, Australia, and New Zealand. Many western countries now adopt selenium supplementation programs in agriculture, but these are often not available in South America, Africa, and Asia and livestock productivity is significantly impaired by selenium deficiency in these regions (Levander 1986; WHO 1987, 1996; Oldfield 1999) (see also Chap. 21, this volume).

16.3.4 Selenium Deficiency—Effects in Humans

No clear-cut pathological condition resulting from selenium deficiency alone has been identified in humans, however, the element has been implicated in a number of diseases (WHO 1987, Rayman 2012) (see also Chaps. 9 and 25, this volume).

16.3.4.1 Keshan Disease

Keshan disease (KD) is an endemic cardiomyopathy (heart disease) that mainly affects children and women of childbearing age in China. The disease has been documented for over 100 years, but the name is derived from a serious outbreak in Keshan County, northeast China in 1935. Outbreaks have been reported in a broad belt stretching from Heilongjiang Province in the northeast of China to Yunnan province in the southwest that transcends topography, soil types, climatic zones, and population types (Fig. 16.8). This disease manifests as an acute insufficiency of the heart function or as a chronic moderate-to-severe heart enlargement and can result in death. Seasonal variations in outbreak were noted with peaks in the winter in the south and in the summer in the north. The worst affected years on record were 1959, 1964, and 1970 when the annual prevalence exceeded 40 per 100,000 with more than 8,000 cases and 1,400–3,000 deaths each year (Tan 1989).

Although the disease occurred in a broad belt across China, all of the affected areas were characterized by remoteness and a high proportion of subsistence farmers who were very dependent on their local environment for their food supply. Investigators noticed that WMD in animals occurred in the same areas and further studies demonstrated that the soils and crops were very low in selenium. KD occurred in areas where grain crops contained

<0.04 mg kg^{-1} of selenium and dietary selenium intakes were extremely low, between 10 and 15 μg day^{-1}. Affected populations were characterized by very low selenium status indicated by hair contents of <0.12 mg kg^{-1} (Xu and Jiang 1986; Tan 1989; Yang and Xia 1995). On the basis of these findings, large-scale mineral supplementation was carried out on 1- to 9-year-old children who were at high risk of the disease. In a trial carried out in Mianning County, Sichuan Province, from 1974 to 1977, 36,603 children were given 0.5–1.0-mg sodium selenite tablets per week whereas 9,642 children were given placebo tablets. During the 4 years of investigation, 21 cases of the disease and 3 deaths occurred in the selenium-supplemented group whereas 107 cases and 53 deaths occurred in the control group. By 1977 all the children were supplemented with selenium and the disease was no longer prevalent in either group. The results showed that supplements of 50-μg day^{-1} selenite could prevent the disease but if the disease was already manifest, selenium was of no therapeutic value (Anonymous 2001).

Although the disease proved to be selenium-responsive, the exact biological function of the element in the pathogenesis was less clear and the seasonal variation in disease prevalence suggested a viral connection. Subsequent studies have demonstrated a high prevalence of the Coxsackie B virus in KD patients (see, for example, Li et al. 2000) and studies have proved increased cardiotoxicity of this virus in mice suffering from selenium and vitamin E deficiency. For a number of years it was thought that selenium deficiency impaired the immune function lowering viral resistance, however, more recent work by Beck (1999) has shown that a normally benign strain of Coxsackie B3 (CVB3/0) alters and becomes virulent in either selenium-deficient or vitamin E-deficient mice. Once the mutations are completed, even mice with normal nutritional status become susceptible to KD. These changes in the virus are thought to occur as a result of oxidative stress due to low vitamin E and low selenium status. This work demonstrates not only the importance of selenium deficiency in immunosuppression of the host but in the toxicity of the viral pathogen as well. Other studies have implicated moniliformin mycotoxins produced by the fungi *Fusarium proliferatum* and *F. subglutinans* in corn as a possible cause of KD (Pineda-Valdes and Bullerman 2000). As with many environmental conditions, KD is likely to be multifactorial but even if selenium deficiency is not the main cause of the disease, it is clearly an important factor.

During the 1980s the prevalence of KD dropped to less than 5 per 100,000 with less than 1,000 cases reported annually. The reason for this is twofold: first, widespread selenium supplementation programs have been carried out on the affected populations and secondly, economic and communication improvements in China as with the rest of the world mean that the population is increasingly less dependent on locally grown foodstuffs in the diet. In recent years the incidence of the disease has dropped still further so that it is no longer considered a public health problem in China (Burk 1994).

Fig. 16.9 Kashin-Beck disease patient (*left*) and woman of the same age (Reproduced with permission from Tan 1989)

16.3.4.2 Kashin-Beck Disease

Kashin-Beck disease (KBD), an endemic osteoarthropathy (stunting of feet and hands) causing deformity of the affected joints, occurs in Siberia, China, North Korea, and possibly parts of Africa. The disease is named after the Russian scientists who first described it between 1861 and 1899. It is characterized by chronic disabling degenerative osteoarthrosis affecting the peripheral joints and the spine with apoptosis of the hyaline cartilage tissues. Impairment of movement in the extremities is commonly followed by bone development disturbances such as shortened fingers and toes and in more extreme cases, dwarfism (Fig. 16.9) (Levander 1986; Tan 1989; WHO 1996). Indeed the main feature of KBD is short stature caused by multiple focal apoptosis in the growth plate of the tubular bones. In China, the pattern of

disease incidence is concordant with KD in the north of the country, but the links with selenium-deficient environments are less clear (Tan 1989).

Initial studies revealed that rats fed grain and drinking water from the affected areas in China suffered acute massive liver apoptosis and foodstuffs from the affected areas were found to be low in selenium. Children and nursing mothers were supplemented with 0.5–2.0 mg sodium selenite per week for a period of 6 years and the disease prevalence dropped from 42 to 4% in children aged 3–10 years as a result (Tan 1989). More recent studies carried out since the early 1990s demonstrated KBD-like cartilage changes and bone mineral density reduction in selenium-deficient rats (Sasaki et al. 1994; Moreno-Reyes et al. 2001). However, other factors have been implicated in the pathogenesis of KBD. The main theory proposed by Russian investigators was that the disease was a result of mycotoxins in the diet, and other work carried out in China has suggested ingestion of contaminated drinking water as a possible cause. In China, higher fungal contamination of grain in KBD areas has been known for a number of years. Other work suggests that the presence of humic substances in drinking water is a factor, and the mechanism of action is free radical generation from the oxy and hydroxyl groups of fulvic acid. Nonetheless, selenium was confirmed as a preventative factor in KBD in these studies (Peng et al. 1999).

There are similarities between KBD and the iodine-deficiency disorder cretinism. Several studies have considered the relationships between KBD and selenium and iodine deficiency. Work in the Yulin District of China (Zhang et al. 2001) carried out on 353 rural school children aged 5–14 years compared data between three endemic KBD villages (prevalence rates 30–45%) and a non-endemic village. Higher fungal contamination was recorded in cereal grain stores in KBD areas than in the non-endemic village, and hair selenium and urinary iodine concentrations were lower in families suffering from the disease than in control groups. However, iodine deficiency did not correlate significantly with increased KBD risk. More recent work into the disease has focused on Tibet and it does implicate iodine in the pathogenesis. Among 575 5- to 15-year-old children examined in 12 villages, 49% had KBD, 46% had the iodine deficiency disorder goiter, and 1% had cretinism. Of the examined population, 66% had urinary iodine contents of <0.02 μg L^{-1} and the content was lower in KBD patients (0.12 μg L^{-1}) than in control subjects (0.18 μg L^{-1}). Hypothyroidism was more frequent in the KBD group (23%) compared to 4% in the controls. Severe selenium deficiency was present in all groups with 38% of subjects with serum concentrations of <5 mg L^{-1} (normal range 60–105 mg L^{-1}). Statistical analyses revealed an increased risk of KBD in groups with low urinary iodine in the severe selenium-deficient areas. Here also, mesophilic fungal contamination in barley (*Alternaria* sp.) was higher in KBD areas than non-endemic areas and disease prevalence correlated positively with the humic content of drinking water. The results suggest that KBD is multifactorial and occurs as a consequence of oxidative damage to cartilage and bone cells when associated with decreased antioxidant defense. Another mechanism that may coexist is bone remodeling stimulated by thyroid hormones whose actions are blocked by certain mycotoxins (Suetens et al. 2001).

16.3.4.3 Iodine Deficiency Disorders and Thyroid Function

In addition to the links between selenium and iodine deficiency in KBD, the establishment of the role of the selenoenzyme, iodothyronine deiodinase (Dio), in thyroid function means that selenium deficiency is now being examined in relation to the iodine deficiency disorder (IDD) goiter and cretinism. Many areas around the world where IDD is prevalent are deficient in selenium including China, Sri Lanka, India, Africa, and South America (WHO 1987, 1996). Concordant selenium and iodine deficiency are thought to account for the high incidence of cretinism in Central Africa, in Zaire and Burundi in particular (Kohrle 1999), and selenium deficiency has been demonstrated in populations suffering IDD in Sri Lanka (Fordyce et al. 2000a). Relationships between low selenium status and thyroid tissue damage and goiter in France and with thyroid cancer in Norway have been reported. Selenium supplementation has also been shown to protect against Hashimoto's thyroiditis and Graves' disease (Rayman 2012).

16.3.4.4 Cancer

Following studies that revealed an inverse relationship between selenium in crops and human blood versus cancer incidence in the United States and Canada (Shamberger and Frost 1969), the potential anti-carcinogenic effect of selenium has generated a great deal of interest in medical science. Many studies to examine the links between selenium and cancer in animal experiments and humans have been carried out; however, to date, the results are equivocal. There is some evidence to suggest that selenium is protective against bladder, colorectal, lung and prostate cancer due to its antioxidant properties, the ability to counteract heavy metal toxicity, the ability to induce cell death, the ability to inhibit cell growth, and the ability to inhibit nucleic acids and protein synthesis, but trial results are mixed (WHO 1987, 1996; Clark et al. 1996; Varo et al. 1998; Rayman 2012). However, other studies have shown that selenium may promote cancer based on the pro-oxidant mutagenic and immunosuppressive actions of some selenium compounds. For example, the supplementation of sodium selenate, sodium selenite, and organic selenium have been

shown to reduce the incidence of several tumor types in laboratory animals, but selenium sulfide has been shown to be carcinogenic in animals and has been classified as a Group B2 compound—a possible human carcinogen (WHO 1987, 1996; ATSDR 2003). Human studies have demonstrated low levels of selenium in the blood of patients suffering gastrointestinal cancer, prostate cancer, or non-Hodgkin's lymphoma, but there is some evidence to suggest that selenium increases the risks of pancreatic and skin cancer (WHO 1987, 1996; Birt et al. 1989).

Excellent reviews of the work into selenium and cancer are presented by Vinceti et al. (2000) and Rayman (2012) and are summarized as follows. In a Nutritional Prevention of Cancer study in 1996 of patients with a history of basal or squamous cell skin cancer, selenium intakes of 200 μg day^{-1} appeared to reduce mortality from all cancers and the incidence of lung, colorectal, and prostate cancers. However, it did not prevent the appearance of skin cancer. Indeed, some studies have shown an inverse relationship with melanoma risk but other studies have shown no relationship with non-melanoma skin cancer. However, Vinceti et al. (1998) carried out assessments of populations inadvertently exposed to high selenium in drinking water and reported higher mortality from lung cancer, melanoma, and urinary cancer among men and lymphoid neoplasm in women in the exposed group compared to controls. Other studies have shown an increased risk of colon and prostate cancer in populations taking selenium supplements in Iowa in the United States and Finland than in control populations. However, a study in Montreal carried out between 1989 and 1993 found no association between selenium status (measured by toenail selenium) and breast or prostate cancer but showed an inverse relationship with colon cancer (Vinceti et al. 2000). The Selenium and Vitamin E Cancer Trial (SELECT) showed that supplementation with 200 μg selenium per day did not reduce the risk of localized prostate cancer in 35,533 American men but the men were not of low selenium status to begin with (Lippman et al. 2009). More recent studies have shown no significant association between selenium status and lung cancer and a complex relationship with prostate cancer whereby selenium appears to protect against risk of advanced rather than localized or low grade disease particularly in smokers (Rayman 2012).

In a study of stomach cancer in Finland and The Netherlands, an inverse relationship between selenium status and disease prevalence was found in Finnish men but not in Finnish women or in men or women from The Netherlands. No relationship between selenium status and stomach cancer was evident in studies carried out in Japan. A link between low-selenium status and pancreatic cancer was observed in Maryland in the United States and in Finland, but a similar relationship with bladder and oropharyngeal cancer was evident in Maryland only (Vinceti et al. 2000). Indeed Finland provides an interesting case because the government was so concerned about the low level of selenium intake in the Finnish diet that in 1984 a national program was initiated to increase the selenium content of Finnish foodstuffs by adding sodium selenate fertilizers to crops. Mean daily intakes rose from 45 μg day^{-1} in 1980 to 110–120 μg day^{-1} between 1987 and 1990 and 90 μg day^{-1} in 1992 (Varo et al. 1998). Studies of cancer incidence over this time carried out in Finland, Sweden, and Norway showed no reduction in colon cancer, non-Hodgkin's lymphoma, or melanoma in Finland whereas breast and prostate cancer rates increased compared to the other two countries. Populations in Finland and New Zealand are known to have much lower selenium status than many other countries and yet no excess incidence of breast and colon cancer is evident. Similarly no relationship between cancer prevalence and selenium intakes as low as 14 μg day^{-1} was identified among rural farmers in China. Conversely, work by Finley et al. (2001) has demonstrated a link between consumption of high-selenium broccoli and reduced colon and mammary cancer prevalence. Based on the evidence presented above, it is fair to say that "the jury is still out" in terms of the beneficial effects or otherwise of selenium and cancer.

16.3.4.5 Cardiovascular Disease

Selenium deficiency has also been implicated in cardiovascular health and it is suggested that serum concentrations of less than 45 μg L^{-1} increase the risk of ischemic heart disease. Animal studies have demonstrated that selenium could play a protective role by influencing platelet aggregation and increasing production of thromboxane A2 while reducing prostacyclin activity as well as several cardiometabolic effects that have been linked to selenium dependant enzymes such as GPx and selenoprotein S. However, the epidemiological evidence and studies into selenium status and disease risk provide contradictory results (Levander 1986; WHO 1987, 1996). Randomized trials based on selenium supplementation have shown no protective effect on cardiovascular disease but there is some evidence of an inverse association between coronary heart disease and low selenium status (Rayman 2012).

16.3.4.6 Reproduction

The full role of selenium in reproduction has yet to be established; however, selenium deficiency has been shown to cause immotile and deformed sperm in rats (Wu et al. 1979; Hawkes and Turek 2001). In men the testis takes up selenoprotein P1 required for adequate GPx4 activity, which is essential for sperm motility. Trials have shown reduced GPx4 in infertile men and selenium supplementation has been shown to increase sperm motility. There is some evidence of lower selenium status in women suffering miscarriages and low selenium intake has also been linked to pre-eclampsia (Rayman 2012).

16.3.4.7 Other Diseases

Selenium deficiency has been linked to a number of other conditions in man as the concentration of the element is decreased in the serum/plasma or erythrocytes of patients with AIDS, trisomy-21, Crohn's and Down syndrome, and phenylketonuria. The evidence of viral mutogeny under selenium deficiency established by Beck (1999) in the case of the Coxsackie B virus has major implications in terms of the toxicity and immunoresponse to many viral infections, particularly AIDS, in light of the widespread selenium-deficient environments of central and southern Africa where the disease has reached epidemic proportions (Longombe et al. 1994). Two randomized trials in the U.S. have shown benefits from selenium supplementation in HIV infection whereas one in Tanzania did not (Rayman 2012). Selenium supplementation is now being used alongside other treatments to help the fight against HIV in Africa.

Selenium deficiency has also been linked to muscular dystrophy (a similar human disease to WMD in animals) and muscular sclerosis, but again the medical evidence for the role of the element is equivocal. Selenium supplementation has proved beneficial to patients suffering renal disease and finally an inverse relationship between selenium status and asthma incidence has also been postulated (WHO 1987).

Dietary intakes of 0.1–0.2 mg kg^{-1} selenium are considered nutritionally generous converting to 50–100 μg day^{-1} for a typical person and 0.7–2.8 μg kg body $weight^{-1}$. Even in New Zealand and Finland, where selenium intake is 30–50 μg day^{-1}, compared with 100–250 μg day^{-1} in the United States and Canada, overt clinical signs of selenium deficiency are rare among humans (WHO 1987, 1996). Nonetheless, research increasingly shows the essential nature of selenium to human health and the potential for subclinical effects should not be underestimated. Concern is growing in many regions of the world over low levels of dietary selenium intake in human populations.

16.3.5 Selenium Toxicity—Effects in Animals

Experiments on laboratory animals have demonstrated that hydrogen selenide is the most toxic selenium compound by inhalation, sodium selenite the most toxic via ingestion, and elemental selenium in the diet has low toxicity as it is largely insoluble (ATSDR 2003; WHO 1987). Sodium selenite or seleniferous wheat containing 6.4 mg kg^{-1} selenium causes growth inhibition and hair loss in animals and at concentrations of 8 mg kg^{-1}, pancreatic enlargement, anemia, elevated serum bilirubin levels, and death follow (Levander 1986; WHO 1987). In addition to food intake, the application of sodium selenate in drinking water has been shown to cause fetal deaths and reduced fertility in mice. Selenium sulfide is the only compound proven to be carcinogenic in animal studies which results in increased liver tumors in rats. Although it is used in anti-dandruff shampoos it is not normally found in food and water. The oral lethal dose for sodium selenite in laboratory animals has been shown to range from 2.3 to 13 mg kg^{-1} body weight. Methylation of selenium is used as a detoxification mechanism by animals, and inorganic and organic forms of selenium are metabolized to form mono-, di-, or trimethylated selenium, of which monomethylated forms are most toxic. For example, dimethylselenide is 500–1,000 times less toxic than selenide (Se^{2-}) (WHO 1987) (see also Chap. 21, this volume).

In natural conditions, acute selenium intoxication is uncommon as animals are not normally exposed to high selenium forage and tend to avoid eating selenium accumulator plants. Abnormal posture and movement, diarrhea, labored respiration, abdominal pains, prostration, and death, often as a result of respiratory failure, characterize toxicity. The characteristic symptoms of selenium poisoning are the garlic odor due to exhalation of dimethylselenide, vomiting, shortness of breath, and tetanic spasms. Pathological changes include congestion of the liver and kidneys, and swelling and hemorrhages of the heart (Levander 1986; WHO 1987, 1996).

Chronic selenium intoxication is more common and leads to two conditions known as alkali disease and blind staggers in grazing animals. Alkali disease occurs after ingestion of plants containing 5–40 mg kg^{-1} over weeks or months and is characterized by dullness, lack of vitality, emaciation, rough coat, sloughing of the hooves, erosion of the joints and bones, anemia, lameness, liver cirrhosis, and reduced reproductive performance. Although much of the work on alkali disease has focused on cattle, consumption of feeds containing 2 mg kg^{-1} of selenium has also been shown to cause hoof deformation, hair loss, hypochromic anemia, and increased alkali and acid phosphatase activities in sheep (Levander 1986; WHO 1987).

Blind staggers occurs in cattle and sheep but not in horses and dogs and occurs in three stages:

- Stage 1: The animal wanders in circles, has impaired vision, and is anorexic.
- Stage 2: The stage 1 effects get worse and front legs weaken.
- Stage 3: The tongue becomes partially paralyzed and the animal cannot swallow and suffers blindness, labored respiration, abdominal pain, emaciation, and death.

Pathological changes include liver apoptosis, cirrhosis, kidney inflammation, and impaction of the digestive tract. Treatment of the condition involves drenching with large amounts of water and ingestion of strychnine sulfate.

However, selenium may not be the main cause of blind staggers, which has similarities to thiamine deficiency. High sulfate intake has been implicated in the disease and may enhance the destruction of thiamine (WHO 1987).

In addition to alkali disease and blind staggers, high selenium intakes in pigs, sheep, and cattle have been shown to interfere with normal fetal development. Selenosis has been known to cause congenital malformation in sheep and horses and reproductive problems in rats, mice, dogs, pigs, and cattle whereby females with high selenium intakes had fewer smaller young that were often infertile. Blood selenium levels of >2 mg L^{-1} in cattle and >0.6–0.7 mg L^{-1} in sheep are associated with selenosis with borderline toxicity at 1–2 mg L^{-1} in cattle (Levander 1986; WHO 1987).

Although much of the work into selenium toxicity has focused on agricultural species, selenosis has also been reported in wild aquatic species and birds. Selenium concentrations of 47–53 μg L^{-1} in surface waters results in anemia and reduced hatchability of trout whereas concentrations of 70–760 μg L^{-1} in water are toxic to most aquatic invertebrates. Cranial and vertebral deformities occur in frogs exposed to 2,000 μg L^{-1} in surface waters. Selenium toxicity is also associated with embryonic deformities in birds; indeed, the hatchability of fertile eggs is a sensitive indicator of selenium intoxication. At concentrations of 6–9 mg kg^{-1} in the diet, embryos suffer brain tissue, spinal cord, and limb bud deformities whereas >7 mg kg^{-1} causes reduced egg production and growth (WHO 1987; Jacobs 1989).

16.3.6 Selenium Toxicity—Effects in Humans

The toxicity of selenium compounds to humans depends on the chemical form, concentration, and on a number of compounding factors. The ingestion of selenious acid is fatal to humans, preceded by stupor, hypertension, and respiratory depression whereas the toxicity of methylated selenium compounds depends not only on the dose administered but also on the previous level of selenium intake. Higher selenium intake prior to dosing with methylated compounds has been shown to be protective against toxicity in animal experiments. Poor vitamin E status increases the toxicity of selenium and the nutritional need for the element, whereas sulfate counteracts the toxicity of selenate but not of selenite or organic selenium and increases selenium urinary excretion. Methylmercury enhances selenium deficiency but inorganic mercury increases methylated selenium toxicity. At intakes of 4–8 mg kg^{-1}, selenium increases the copper contents of the heart, liver, and kidneys but has a detoxifying or protective effect against cadmium and mercury (WHO 1987; Bedwal et al. 1993). High selenium intake has also been shown to decrease sperm motility in healthy men (Hawkes and Turek 2001) and has been related to increased incidence of some forms of cancer including pancreatic and skin cancer (see Sect. 16.3.4 in this chapter). Possible increased risk of Type-2 diabetes as a result of selenium supplementation has also been reported in recent cancer trials. This increased risk may relate to the effect of high selenium on insulin signaling (Rayman 2012).

Overt selenium toxicity in humans is far less widespread than selenium deficiency. Following the discovery of seleniferous environments and the incidence of alkali disease in animals in the Great Plains in the United States during the 1950s, concern about potential adverse affects on the human population were raised. The health status of rural populations in seleniferous areas was examined. Results showed elevated urinary selenium levels in the population but no definite links to clinical symptoms of selenosis. However, a higher incidence of gastrointestinal problems, poor dental health, diseased nails, and skin discoloration were reported (Smith and Westfall 1937). In similar studies in a seleniferous region of Venezuela, the prevalence of dermatitis, hair loss, and deformed nails among children was higher than in non-seleniferous areas. The hemoglobin and hematocrit values in children from the seleniferous areas were lower than in controls but did not correlate with blood or urine selenium levels and evidence of selenium toxicity effects was rather inconclusive. Nine cases of acute selenium intoxication due to the intake of nuts of the *Lecythis ollaria* tree in a seleniferous area of Venezuela have been reported to result in vomiting and diarrhea followed by hair and nail loss and the death of a 2-year-old boy (WHO 1987) (see also Chap. 25, this volume).

In China, an outbreak of endemic human selenosis was reported in Enshi District, Hubei Province, and in Ziyang County, Shanxi Province, during the 1960s. This condition was associated with consumption of high-selenium crops grown on soils derived from coal containing >300 mg kg^{-1} selenium. In the peak prevalence years (1961–1964) morbidity rates reached 50% in the worst affected villages, which were all located in remote areas among populations of subsistence farmers. Hair and nail loss were the prime symptoms of the disease but disorders of the nervous system, skin, poor dental heath, garlic breath, and paralysis were also reported. Although no health investigations were carried out at the time, subsequent studies in these areas carried out in the 1970s revealed very high dietary intakes of 3.2–6.8 mg with a range of selenium in the blood of 1.3–7.5 mg L^{-1} and hair selenium levels of 4.1–100 mg kg^{-1} (Yang et al. 1983; Tan 1989).

Selenium toxicity related to mineral supplement intake has also been reported in the United States. In 1984, 12 cases of selenosis due to intakes over 77 days of tablets labeled to contain 0.15–0.17 mg selenium, but which actually were found to contain 27–31 mg selenium were reported. Patients

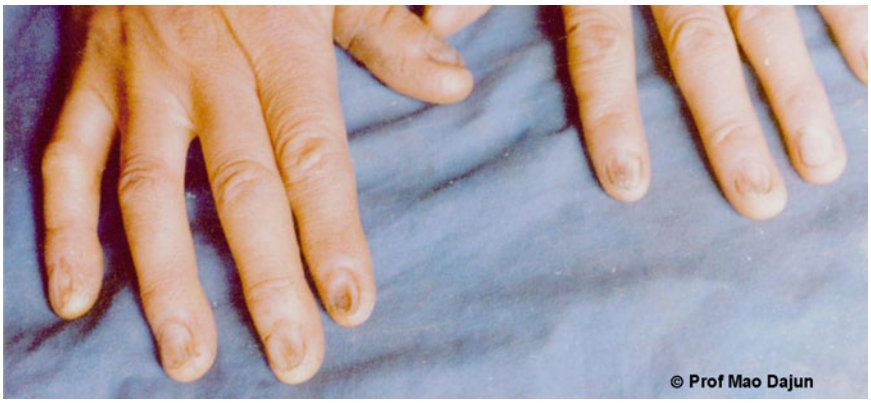


Fig. 16.10 Nail deformities as a result of selenium poisoning in Enshi District, China (Photo: Professor Mao Dajun. Reproduced with permission from the British Geological Survey, Keyworth, Nottingham)

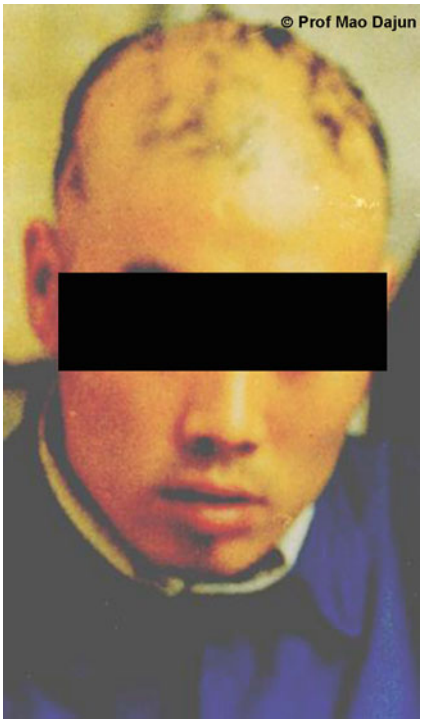


Fig. 16.11 Hair loss as a result of selenium toxicity in Enshi District, China (Photo: Professor Mao Dajun. Reproduced with permission from the British Geological Survey, Keyworth, Nottingham)

suffered nausea, vomiting, nail damage, hair loss, fatigue, irritability, abdominal cramps, watery diarrhea, skin irritation, and garlic breath and had blood serum levels of 0.528 mg L^{-1} (WHO 1987). The U.S.-EPA recommends an upper limit of mineral supplementation of selenium of 0.1 mg kg^{-1}.

Indeed, a whole list of symptoms has been implicated in elevated selenium exposure including severe irritations of the respiratory system, metallic taste in the mouth, tingling and inflammation of the nose, fluid in the lungs, pneumonia, the typical garlic odor of breath and sweat due to dimethylselenide excretion, discoloration of the skin, dermatitis, pathological deformation, and loss of nails (Fig. 16.10), loss of hair (Fig. 16.11), excessive tooth decay and discoloration, lack of mental alertness and listlessness, peripheral neuropathy, and gastric disorders. The links with dental health are somewhat equivocal and many of the studies indicating a possible link with selenium failed to take into account other factors such as the fluoride status of the areas of study (WHO 1987).

Part of the problem in assessing high selenium exposure is that there is some evidence to suggest populations can adapt to or tolerate high selenium intakes without showing major clinical symptoms. Investigations are also hampered by the lack of a sensitive biochemical marker of selenium overexposure (WHO 1996). Hair loss and nail damage are the most common and consistent clinical indications of the condition. Chinese studies carried out in the seleniferous areas of Hubei and Ziyang have demonstrated that these effects are evident above dietary intakes of 900 μg day^{-1}, blood plasma levels of 1 mg L^{-1}, and whole blood concentrations of 0.813 mg L^{-1} (Yang et al. 1983). Interestingly, further work in China has shown a marked reduction in the ratio of selenium in plasma compared to that in erythrocytes at dietary intakes of 750 mg kg^{-1}. This is the first indication of a biochemical response to high selenium intakes prior to the development of clinical symptoms (Yang et al. 1989). There is still a great deal of uncertainty about harmful doses of selenium, but a maximum recommended dietary intake of 400 μg day^{-1} has been proposed based on half the level of intake found in the Chinese studies (WHO 1996).

In summary, there is a U-shaped association between selenium intake and health and this may explain the contradictory results from the clinical trials into selenium impacts on disease. Populations of low-selenium status may benefit from selenium supplementation in terms of disease/health outcomes; whereas populations with pre-existing adequate or high status may not benefit and may in fact suffer detrimental health outcomes (Rayman 2012).

16.4 Measuring Selenium Status

Thus far in this chapter selenium deficiency and toxicity in the environment, plants, animals, and human beings has been discussed, but in order to assess selenium status, it is important to consider how it is measured. Information about the chemical composition of the terrestrial environment is generally collated by national survey organizations concerned with geology, soil, water, agriculture, and vegetation. In terms of geology, for example, over 100 countries around the world carry out national geochemical mapping programs. These programs are based on systematic collection of materials such as soil, sediment, water, rock, and vegetation, which are then analyzed for a range of element compositions and used to produce maps of element distributions in the environment. This type of approach was pioneered in the 1950s by Russian geochemists and the wide application of these methods has been made possible by improvements in rapid multi-element analysis techniques over the last 60 years.

However, selenium is not an easy element to analyze, partly because concentrations in natural materials are so low. Therefore, in many multi-element geochemical surveys, selenium was not included in the analysis. It was not until the last 30 years that analytical advances have allowed the detection of selenium at low enough concentrations to be of real interest to environmental studies but because these techniques are more expensive than routine analytical programs, selenium is still often missing from the group of determinants despite its environmental importance (Darnley et al. 1995). A summary of some of the selenium data available around the world is provided by Oldfield (1999).

Analytical methods that give good limits of detection (<1 mg kg^{-1}) include colorimetry, neuron activation analysis (NAA), x-ray fluorescence spectrometry (XRF), atomic fluorescence spectrometry (AFS), gas chromatography (GC), inductively coupled mass spectrometry (ICP-MS), and inductively coupled atomic emission spectrometry (ICP-AES). Of these, AFS is the most widely used for natural materials such as foods, plants, and soils. NAA is often used to determine different selenium isotopes, especially in tracer studies using ^{75}Se, etc., and gives good detection limits, but it is a more specialized form of analysis. For studying stable isotopes of selenium (for example, ^{78}Se and ^{82}Se) ICP-MS may be used, but it requires enriched and expensive isotope materials. In more recent years, hydride-generation techniques have improved the detection limits of spectrometric methods such as ICP-AES. Ion exchange chromatography has been extensively used to determine selenium compounds in plants whereas gas chromatography is employed to determine volatile selenium compounds. Ion exchange or solvent extraction methods are used to distinguish selenate and selenite species in solution. Recently developed anion exchange high-performance liquid chromatography (HPLC) and ICP-dynamic-reaction–cell MS methods can be used to measure selenium isotopes and selenoamino acids including selenocysteine and selenomethionine. Using these techniques it is now possible to measure relatively low selenium concentrations and selenium element species in a wide variety of environmental and biological materials (WHO 1987).

In animals and humans a variety of bio-indicators of selenium status have been employed. Due to the close association between the level of dietary selenium intake and GPx activity, the fact that the enzyme activity represents functional selenium and that assessments of this enzyme are easier to perform than selenium tests, GPx activity has been used extensively to measure selenium status, especially in animals. However, this method requires caution because GPx activity is influenced by other physiological factors and a non-selenium-dependent GPx enzyme is also present in animals and humans. Furthermore, the enzyme activity may provide an indication of selenium status at lower levels of intake, but at higher concentrations of selenium the GPx activity becomes saturated and the enzyme cannot be used to indicate toxic selenium status (WHO 1996) (see also Chaps. 25 and 32, this volume).

Other indicators of selenium status include whole blood, plasma, or serum; hair; toenail; and urine content. Of these, hair has been used extensively as it is easy to collect. However, caution is required to ensure that samples are not contaminated with residues from selenium-containing shampoos. It should also be noted that urinary selenium cannot be used to measure inhalation exposure to hydrogen selenide gas, selenium oxychloride, or organic-selenium compounds as severe damage to the lungs occurs before elevated selenium contents are evident in the urine (WHO 1987). Dietary surveys are also commonly used as an indication of selenium intake. Single-day dietary surveys can give errors of up to 90% when used to estimate the real long-term exposure to selenium, because wide ranges in daily intake are commonplace (0.6–221 μg day^{-1}). Comparisons of different methods have shown that 3-week dietary observations give estimates of overall intake to within 20% and are a much more reliable indication of likely selenium status (WHO 1987).

Regardless of the material sampled, whether it is soil, food, blood, water, hair, etc., selenium status is determined by comparison to a set of thresholds and normative values that have been determined by examining the levels at which physiological effects occur in plants, animals, and humans. Some of these thresholds are listed in Table 16.18. In general, total soil selenium contents of 0.1–0.6 mg kg^{-1} are considered deficient as these are the concentrations of selenium found in regions where selenium-deficient livestock are commonplace such as New Zealand, Denmark, and the

Table 16.18 Deficiency and toxicity thresholds and recommended upper limits for selenium in various media

	Deficient mg kg^{-1}	Marginal mg kg^{-1}	Moderate mg kg^{-1}	Adequate mg kg^{-1}	Toxic mg kg^{-1}	Reference
Soils						
World total Se	0.1–0.6					Various
China total Se	0.125	0.175	0.400		>3	Tan (1989)
China water-soluble Se	0.003	0.006	0.008		0.020	Tan (1989)
Plants						
World plants	0.1			0.1–1.0	3–5	Jacobs (1989)
China cereal crops	0.025	0.040	0.070		>1	Tan (1989)
Animals						
Fodder, animals chronic	<0.04			0.1–3	3–15	Jacobs (1989)
Cattle and sheep liver	0.21					WHO (1987)
Cattle/sheep blood mg L^{-1}	<0.04	0.05–0.06		0.07–0.1		Mayland (1994)
Humans						
China hair	0.200	0.250	0.500		>3	Tan (1989)
Urine excretion μg day^{-1}				10–200		Oldfield (1999)
Ref. dose mg kg^{-1} day^{-1}					0.005	ATSDR (2003)
Food	<0.05				2–5	WHO (1996)
Diet μg day^{-1}	<40			55–75	>400	WHO (1996)
	Maximum admissible concentration mg kg^{-1}					Reference
UK soil	120–13,000					EA (2009)
France soil	10					Haygarth (1994)
USA sewage sludge	100					ATSDR (2003)
UK sewage sludge	25					Haygarth (1994)
USA air mg m^{-3}	0.2					ATSDR (2003)
World water μg L^{-1}	40					WHO (2011)
USA water μg L^{-1}	50					US-EPA (2012)
World human urine mg L^{-1}	0.1					WHO (1987)

Atlantic Region of Canada. Work regarding Keshan disease in China suggests levels of 0.125 mg kg^{-1} total selenium in soil cause selenium deficiency in the food chain (Yang et al. 1983; Yang and Xia 1995). However, it should be kept in mind that the amount of total selenium in the soil is not necessarily the critical factor determining selenium status. Several studies have demonstrated that the total selenium content of the soil can be considered "adequate" but if the selenium is not in bioavailable form, it is not taken up into plants and animals and selenium deficiency can result (see, for example, Fordyce et al. 2000a). Total selenium concentrations in soil can give an indication of likely selenium status but do not necessarily tell the whole story, and a selenium deficient environment is not necessarily one in which total concentrations of selenium in soil are the lowest. In more recent years soil with water-soluble selenium has been used as an indicator of the bioavailable fraction and Chinese scientists have recommended soil deficiency and toxicity thresholds on this basis, 0.003 and 0.020 mg kg^{-1}, respectively (Tan 1989).

Due to the many different factors that can influence the uptake of selenium from soil into plants, vegetation often provides a better estimate than soil of likely environmental status with regard to health problems in animals and humans. Feed crops containing more than 0.1 mg kg^{-1} of selenium will protect livestock from selenium deficiency disorders, whereas levels of >3–5 mg kg^{-1} in plants have been shown to induce selenium toxicity in animals (Levander 1986; Jacobs 1989). The current MAC for selenium additives in animal feedstuffs in the United States is 0.3 mg kg^{-1}. In terms of cereal crops for human consumption, Chinese workers suggest deficiency and toxicity thresholds of 0.02 and 0.10 mg kg^{-1}, respectively, based on epidemiological studies in Keshan disease and areas affected with selenosis (Tan 1989). Determination of the selenium status of water is usually made using comparisons to the WHO maximum admissible concentration of 40 μg L^{-1}.

In veterinary science, concentrations of <0.04 mg L^{-1} selenium in animal whole-blood are considered deficient and are related to WMD in ruminant species whereas 0.07–0.10 mg L^{-1} is considered adequate, which highlights the extremely narrow range in selenium status between clinical and non-clinical outcomes (Levander 1986; WHO 1987). Human selenium status is rather more difficult to categorize because of the lack of overt clinical symptoms in many populations exposed to supposedly deficient or

toxic intakes, but based on work in China, deficiency and toxicity thresholds in human hair of 0.2 and >3 mg kg^{-1} (Yang et al. 1983; Tan 1989; Yang and Xia 1995), respectively, have been suggested whereas dietary limits of 40 and 400 μg day^{-1} are proposed by the WHO as an indication of human selenium status (WHO 1996).

16.5 Remediation

A variety of methods have been used to try and counteract the impacts of selenium deficiency and toxicity in environments and within animals and humans as follows.

16.5.1 Remediating Selenium Deficiency

Methods to enhance selenium in the environment and uptake into agricultural crops and animals have been developed over a number of years. One approach is to alter the species of crops grown on deficient soil to plant types that take up more selenium. Switching from white clover production to certain grasses to increase the selenium content of fodder crops has been used successfully in New Zealand (Davies and Watkinson 1966). Another approach is to apply selenium-rich fertilizers to the soil to increase the amount of selenium taken up by plants, animals, and humans. Some rock phosphate fertilizers are rich in selenium and can be used to enhance uptake; however, there is some risk associated with application of selenates to alkali and neutral soils because of high bioavailability. Use of selenate fertilizer results in much higher selenium contents of first cuts of crops or forage, which decrease sharply with subsequent cropping. Addition of selenite to acid-neutral soils can result in some loss of selenium to soil adsorption, which decreases the effectiveness of the application, but in some cases this mechanism can ensure that levels of uptake are not toxic (Fleming 1980; Jacobs 1989). The selenium concentration of foods can also be increased by supplementing ordinary fertilizers with soluble selenium compounds. Finland, New Zealand, and parts of Canada and China allow selenium-enhanced fertilizers to be used for the cultivation of food crops and trials have recently been conducted in the UK (Broadley et al. 2010). These countries mainly use fertilizers based on sodium selenate (Oldfield 1999). For example, in New Zealand 1% granular selenium is mixed with granulated fertilizer and is applied at a rate of 10 g Se ha^{-1} over about a quarter of the agricultural land in the country (in 1998 1.2 million of 4.5 million ha underwent selenium fertilization) (Jacobs 1989; Oldfield 1999).

Problems of uptake associated with retention of selenium in the soil can be circumvented by direct application of the fertilizer to the plants themselves. Foliar application of selenite to plants has been successfully used to increase the selenium content of crops and animals. Spraying selenium at 3–5 g ha^{-1} has been shown to increase the content in grain whereas sodium selenite applied at 50–200 g Se ha^{-1} maintained >0.1 mg kg^{-1} contents in crops through three harvests. Studies have shown that the selenium content of crops is enhanced by mid-tillering spraying with selenium fertilizer but it cannot be applied successfully to seeds (Jacobs 1989).

For example, Chinese workers have reported much better uptake of selenium in maize crops grown on aerated oxygenated soils than in rice grown in the same soils under waterlogged conditions due to reduction of selenium to insoluble forms. To avoid poor uptake of selenium from soils as a consequence of the water-logged conditions, foliar spraying of sodium selenite at an early shooting stage of the rice plant growth was found to improve the selenium content of the grain and hull (Cao et al. 2000). In another study, the average wheat selenium contents in Kashin-Beck endemic areas were 0.009 mg kg^{-1} dry weight resulting in daily intakes of 12 μg in the local population. Following foliar selenium fertilizer application, wheat contents increased to 0.081 mg kg^{-1} dry weight and human daily dietary intakes rose to 47 μg (Tan et al. 1999).

In Finland, the bioavailability of soil selenium for plants is generally poor due to the relatively low selenium concentration, low pH, and high iron content of the soil as much of the country comprises very ancient hard crystalline granite and gneiss rocks. This is very similar to eastern Canada where selenium supplementation is also practiced (Jacobs 1989). In 1984, the Finnish government approved a program of selenium supplementation in fodder and food crops. The program initially involved spraying a 1% selenium solution onto fertilizer granules giving an application rate of 6 g Se ha^{-1} for silage and 16 g ha^{-1} for cereal crops. Within two years a threefold increase of mean selenium intake in the human population was observed and human serum contents increased by 70%. The supplementation affected the selenium content of all major food groups with the exception of fish. In 1990 the amount of selenium that was supplemented was reduced to 6 g Se ha^{-1} for all crops and the mean human selenium intake fell by 30% and the serum selenium concentration decreased by 25% from the highest levels observed in 1989. According to data obtained, supplementation of fertilizers with selenium is a safe and effective means of increasing the selenium intake of both animals and humans and is feasible in countries like Finland with relatively uniform geochemical conditions (Aro et al. 1995). In other countries where the low level of selenium intake is currently of concern, such as in the UK, this kind of intervention would require careful planning and monitoring of the effects on both animal and human nutrition and the environment, because geochemical conditions vary markedly across the country as a result of a diverse geological environment.

In addition to attempts to enhance selenium in fodder crops and animal feeds supplemented with sodium selenite or selenate, selenium deficiency in animals is also prevented by veterinary interventions such as selenium injections to females during late gestation and/or to the young stock shortly after birth, dietary supplements, salt licks, and drenches (Levander 1986).

In humans also, direct dietary supplementation methods have been used successfully to counteract selenium deficiency. Pills containing selenium alone or in combination with vitamins and/or minerals are available in several countries. Selenium supplements contain selenium in different chemical forms. In the majority of supplements, the selenium is present as 35 selenomethionine; however, in multivitamin preparations, infant formulas, protein mixes, and weight-loss products sodium selenite and sodium selenate are predominantly used. In other products, selenium is present in protein-chelated or amino acid chelated forms. Current animal studies and epidemiological evidence favors selenomethionine as the most bioavailable and readily taken up form of selenium in mineral supplements. A dosage of 200 μg day^{-1} is generally considered safe and adequate for adults of average weight consuming a North American diet (WHO 1987). Studies carried out in KD areas of China have shown that both selenite and selenium-yeast supplements were effective in raising GPx activity of selenium-deficient populations, but selenium-yeast provided a longer lasting body pool of selenium (Alfthan et al. 2000). Altering the diets of humans to include selenium-rich foods has also proved successful in preventing selenium deficiency. In China, selenium-rich tea, mineral water, and cereal crops are now marketed in selenium-deficient areas.

16.5.2 Remediating Selenium Toxicity

One of the most common methods to reduce the effects of selenium in soil is phytoremediation. This practice is carried out by growing plant species, which accumulate selenium from the soil and volatilize it to the air to reduce levels in soil. For example, the hybrid poplar trees *Populus tremula x alba* can transfer significant quantities of selenium by volatilization from soil to air; the rate for selenomethionine is 230 times that of selenite and 1.5 times higher for selenite than selenate. These trees have been used successfully to reduce selenium contents in soil in the western areas of the United States (Oldfield 1999) and have been tested in seleniferous regions of India (Dhillon et al. 2008).

There is some evidence to suggest that the presence of phosphate and sulfate in soils can inhibit the uptake of selenium in plants and application of these minerals as soil treatments could be beneficial against selenium toxicity in agricultural crops. Studies have shown a tenfold increase in sulfate content reduced uptake from selenate by >90% in ryegrass and clover, whereas a similar increase in phosphate content caused 30–50% decreases in selenium accumulation from selenite in ryegrass, but in clover such decreases only occurred in the roots. Therefore, sulfate-selenate antagonisms were much stronger than phosphate-selenite antagonisms. The addition of sulfur or calcium sulfate (gypsum) to seleniferous soils in North America was not successful in reducing uptake into plants probably because these soils already contain high quantities of gypsum. However, additions of calcium sulfate and barium chloride have been shown to markedly reduce the uptake of selenium in alfalfa in the United States (90–100%) probably due to the formation of $BaSeO_4$, which is barely soluble. The practicalities of this type of selenium remediation method are rather limited (Jacobs 1989).

It is more common to counteract selenium toxicity with veterinary and medical interventions. Sodium sulfate and high protein intakes have been shown to reduce the toxicity of selenate to rats but not of selenite or selenomethionine in wheat. Arsenic, silver, mercury, copper, and cadmium have all been shown to decrease the toxicity of selenium to laboratory animals and they have been used to alleviate selenium poisoning in dogs, pigs, chicks, and cattle (Moxon 1938; Levander 1986; WHO 1987). The protective effect of arsenic is thought to be a consequence of increased biliary selenium excretion. Laboratory evidence suggests that mercury, copper, and cadmium exert a beneficial effect due to reactions with selenium in the intestinal tract to form insoluble selenium compounds. However, consideration must be given to the toxic effects of these elements before they are applied as selenium prophylaxis. Linseed meal has also been found to counter selenium toxicity in animals by the formation of selenocyanates, which are excreted (WHO 1987).

In terms of human diets, dietary diversification can also help reduce selenium toxicity. In China, high-selenium cereal crops are banned from local consumption and exported out of the seleniferous regions where they are mixed with grains from elsewhere before they are sold in selenium-deficient parts of the country.

16.6 Case Histories

16.6.1 Selenium Toxicity in Animals—Kesterson Reservoir, United States

One of the best known and most studied incidences of selenium toxicity in animals has been recorded at Kesterson Reservoir, California, in the United States. The information summarized here is taken from Jacobs (1989), Wu et al. (2000), Wu (1994), and Tokunaga et al. (1996). These publications should be referred to for further details.

Table 16.19 Comparison of selenium toxicity effects in biota from the seleniferous kesterson reservoir and the selenium-normal volta wildlife area, California, United States

	Kesterson	
Sample type	mg kg^{-1}	Volta mg kg^{-1}
Algae and rooted aquatic plants	18–390	0.17–0.87
Emergent aquatic plants leaves	17–160	<2.0
Terrestrial plants leaves	0.5–27	<4.7
Plankton (geomean)	85.4	2.03
Aquatic insects	58.9–102	1.1–2.1
Mosquito fish	149–380	1.1–1.4
Reptiles (frogs, snakes) liver	11.1–45	2.05–6.22
Birds (coot, duck, stilt, grebe) liver	19.9–43.1	4.41–8.82
Voles liver (geomean)	119	0.228
No. of dead or deformed chick/embryo	22%	1%

From Jacobs (1989)

Due to a scarcity of wetlands in California, wildlife resource managers tried using irrigation runoff from subsurface agricultural drains to create and maintain wetland habitats at the Kesterson Reservoir. The reservoir comprises 12 shallow ponds acting as evaporation and storage basins for agricultural drain waters from the San Joaquin Valley. During part of the year, the water from the reservoir was to be discharged via the San Luis drain back into the Sacramento-San Joaquin River delta when river flows were high enough to dilute the contaminants present in the agricultural water. However, construction of the San Luis drain was halted in 1975 due to increased environmental concerns about the impact of the drain water on the river delta. During the 1970s surface water flow into the reservoir predominated, but into the 1980s almost all the flow was shallow subsurface agricultural drainage water. Selenium concentrations in agricultural drainage water entering the Kesterson Reservoir area between 1983 and 1985 were 300 mg L^{-1} as a result of contact with seleniferous soils in the catchment area. In this arid alkaline environment, 98% of the selenium was in the most readily bioavailable selenate form with only 2% present as selenite. The effects of this water on plants and animals were relatively unknown prior to studies carried out between 1983 and 1985 by the U. S. Wildlife Service comparing Kesterson to the adjacent Volta Wildlife area, which was supplied with clean irrigation water with normal concentrations of selenium. The mortality of embryos, young and adult birds, survival of chicks, and embryonic deformities were compared between the two sites. The selenium content of the livers of snakes and frogs from the two areas were also examined in addition to tissues from 332 mammals of 10 species, primarily moles. Results of some of the comparisons between biota from the two sites are shown in Table 16.19. In all cases, the levels of selenium in biological materials at Kesterson Reservoir exceeded those of the Volta Wildlife area several-fold. Concentrations of selenium in water were compared to those in biota collected from the same site and bioaccumulation factors of more than 1,000 for animals were found at Kesterson. Although no overt adverse health effects were noted in reptile or mammal species such as voles and raccoons in the area, the levels of selenium present were of concern in terms of bioaccumulation in the food chain. In contrast, the overt health effects on birds were very marked with 22% of eggs containing dead or deformed embryos as a result of selenium toxicity. The developmental deformities included missing or abnormal eyes, beaks, wings, legs and feet, and hydrocephaly, and were fatal. It is estimated that at least 1,000 adult and juvenile birds died at Kesterson from 1983 to 1985 as a result of consuming plants and fish with 12–120 times the normal amount of selenium (Jacobs 1989).

Following these revelations, Kesterson Reservoir was closed and a series of remedial measures were tested by a team of scientists who were able to provide a more thorough understanding of processes leading to selenium transport and biologic exposure in this environment. Some of the schemes proposed included the development of an *in situ* chemical treatment to immobilize soluble selenium in drained evaporation pond sediments by amendment with ferrous iron, which occludes selenate and selenite in ferric oxyhydroxide (FeOOH). Phytoremediation techniques were also tested. These included the growing of barley (*Hordeum vulgare* L.) and addition of straw to the soil, which contained 0.68 mg kg^{-1} soluble selenium and 6.15 mg kg^{-1} total selenium. Four treatments were evaluated: soil only, soil + straw, soil + barley, and soil + straw + barley. At the end of the experiments, selenium in barley represented 0.1–0.7% of the total selenium in the system, and volatilized selenium accounted for 0.2–0.5% of total selenium. In contrast, straw amendments were found to greatly reduce the amount of selenium in soil solution by 92–97% of the initial soluble selenium and represented a possible remediation strategy for the reservoir. The planting of canola (*Brassica napus*) was also evaluated but accumulated 50 mg kg^{-1} (dry weight), which accounted for less than 10% of total selenium lost in the soil solution during the post-harvest period.

Bioremediation through the microbial reduction of toxic oxyanions selenite and selenate into insoluble Se^0 or methylation of these species to dimethylselenide was proposed as a potential bioremediation cleanup strategy. Field trials demonstrated that microorganisms, particularly *Enterobacter cloacae*, were very active in the reduction of selenium oxyanions in irrigation drainage water, into insoluble Se^0 and, by monitoring various environmental conditions and the addition of organic amendments, the process could be stimulated many times. Based upon the promising results of these studies, a biotechnology prototype was developed for the cleanup of polluted sediments and water at Kesterson.

A soil excavation plan had been proposed to remove selenium-contaminated material from the site; however, extensive monitoring of pore water in the vadose zone demonstrated that this plan would be ineffective in reducing the elevated selenium concentration in ephemeral pools present during the winter at Kesterson. Furthermore, extensive biological monitoring demonstrated that selenium concentrations in the dominant species of upland vegetation at Kesterson were near or equal to "safe" levels.

On the basis of these studies, a cost-effective remediation strategy was devised. First, the groundwater under Kesterson was protected from selenium contamination by naturally occurring biogeochemical immobilization. Secondly the contaminated soil and sediment was left in place but low-lying areas were infilled to prevent the formation of the ephemeral pools that attracted wildlife. The area was then planted over with upland grassland species. Monitoring studies carried out on soil and vegetation between 1989 and 1999 showed that selenium losses from soil via volatilization were approximately 1.1% per year. Soil selenium concentrations in the fresh soil fill sites increased in the top 15 cm, which indicated that the plants were able to effectively take up soluble soil selenium from the lower soil profile and deposit it at the land surface thus reducing the rate of leaching of soil selenium. In general plant tissue concentrations reflected the amount of soil water-soluble selenium present, which was low. In 1999 plant tissue concentrations averaged 10 mg kg^{-1} (dry weight) and soil water-soluble selenium contents 110 mg kg^{-1} giving an estimated bioaccumulation value for the upland grassland of less than 10% of the previous wetland habitat. It was concluded that the new Kesterson grassland did not pose a risk to the environment (Wu et al. 2000).

16.6.2 Selenium Toxicity and Drinking Water—Reggio, Italy

Examples of high selenium exposure related to intakes in water are very scarce. An exception is reported by Vinceti et al. (1998) and occurred in the town of Reggio, Italy, between 1972 and 1988 where the population in the Rivalta neighborhood was inadvertently exposed to wells containing 3–13 μg L^{-1} selenium as selenate and resultant tap water containing 7–9 μg L^{-1} compared to selenium contents in the drinking water of adjacent neighborhoods of <1 μg L^{-1}. The wells were closed off in 1989 and the population was no longer exposed to water from this source. Apart from the selenium content, water quality between Rivalta and the other neighborhoods was the same. Using residency and water supply records, 2,065 people (1,021 men and 1,044 women) were identified as having been exposed for at least 11 years to the elevated selenium content in the water between 1975 and 1988. This cohort was compared to a control population of non-exposed individuals from the same town. To examine the effects of this exposure on cancer incidence in the local population, all cases of pathologically confirmed primary invasive melanoma occurring during 1996 were collated for the entire town of Reggio as well as records on age, sex, educational level, and occupation. The exposed and non-exposed populations had similar educational and occupational profiles and once the data were corrected for age and sex, a higher prevalence of skin cancer was noted in the exposed population. On the basis of melanoma rates in the unexposed population, 2.06 cases would be expected in the exposed population whereas 8 cases were reported. Although other confounding factors could not be taken into account in this study, there is some evidence to suggest that the skin is a target organism in chronic selenium toxicity and that inorganic selenium can act as a pro-oxidant and mutagen and cell apoptosis suppressant, which may account for the higher prevalence of cancer in the exposed group. It should be noted that selenium is ineffective against melanoma although beneficial for other forms of cancer (Clark et al. 1996).

16.6.3 Selenium Deficiency in Humans—Zhangjiakou District, Hebei Province, China

Zhangjiakou District, Hebei Province, in China lies between Inner Mongolia to the north and Beijing to the south and is one of the remotest regions of China lying within the northeast-southwest KD belt (Fig. 16.8). The area is underlain by Archaen metamorphic and Jurassic volcanic rocks, which are overlain by Quaternary loess and alluvial deposits, all of which contain low amounts of selenium. Within Zhangjiakou District, the KD belt follows the mountainous watershed between the two rock types, which reflects the fact that villages in the remotest locations where populations are most dependent on locally grown foodstuffs are most at risk from the disease. However, within the KD belt, prevalence rates show marked variability between villages ranging from 0 to 10.8% between 1992 and 1996. In a study to examine why this variation may occur and to pinpoint the relationships between environmental selenium and disease, Johnson et al. (2000) examined soil, staple crop (wheat and oats), water, and human hair selenium levels in 15 villages in the region classified according to disease prevalence into three groups: (1) no KD 0% prevalence; (2) moderate KD, 0–3% prevalence; and high KD, >3% prevalence. Results showed that hair, grain, and water selenium concentrations showed an inverse relationship with disease prevalence as expected; the highest selenium contents were reported in villages with lowest prevalence of the disease. However, contrary to expectations, soil total selenium contents showed the opposite relationship and were highest in the villages with greatest disease prevalence

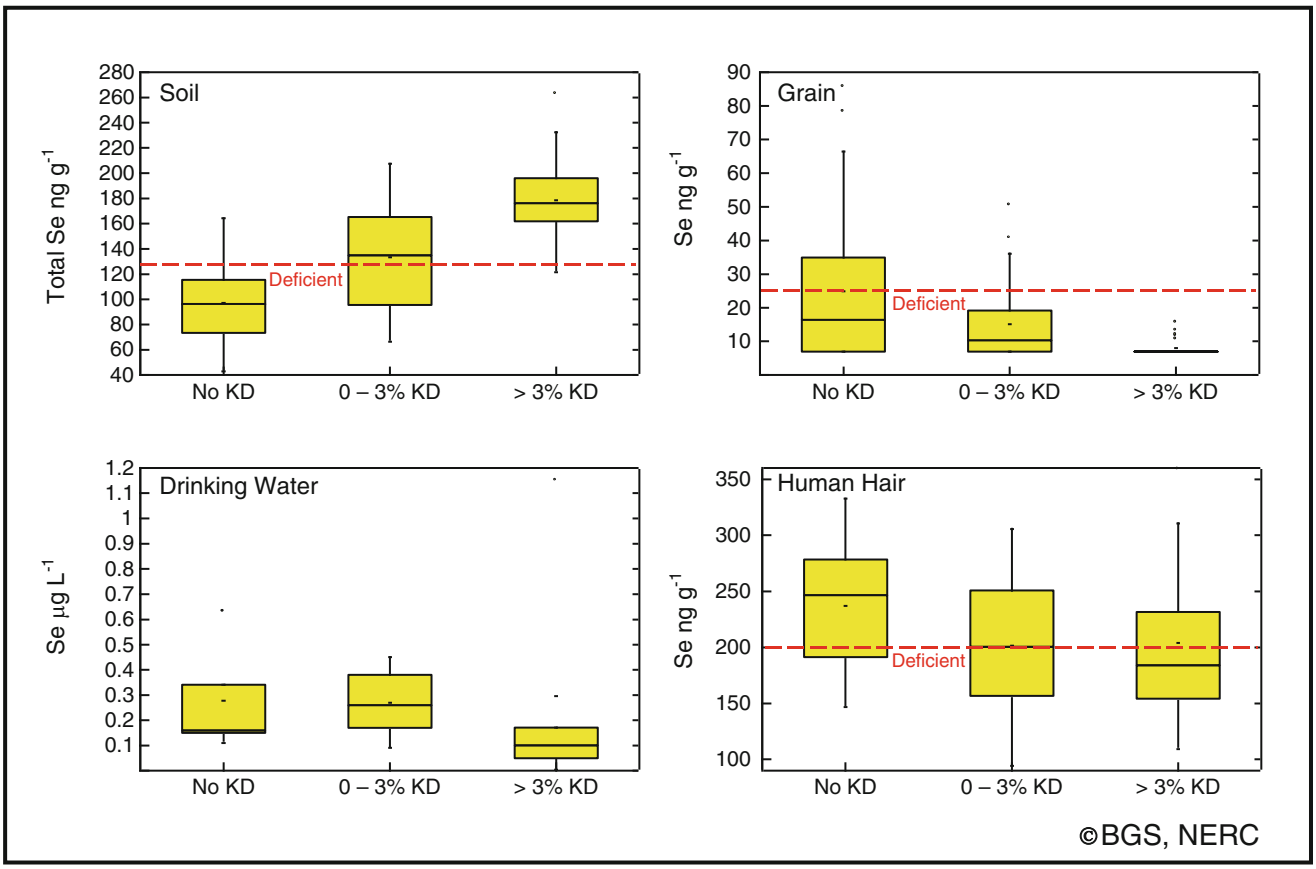


Fig. 16.12 Box and whisker plots of selenium concentrations in soil, grain, drinking water and human hair samples from Zhangjiakou District, Hebei Province, China where the selenium-deficiency related condition KD is prevalent in the local population. Samples were collected from three village groups classified into no KD (0% prevalence), moderate KD (0–3% prevalence), and high KD (>3% prevalence). Medians are represented by *horizontal bars* within the *box* and upper and lower quartiles ±1.5 times the quartile range are represented by the *whiskers*. Deficiency thresholds from Tan (1989) are shown as *dashed lines* on the plots for soil total selenium, grain selenium, and human hair selenium contents (Reproduced with permission from the British Geological Survey, Keyworth, Nottingham)

(Fig. 16.12). Indeed comparisons between the data collected from high prevalence villages for the study and selenium deficiency thresholds proposed by Tan (1989) indicated that the selenium contents of all sample types were very low, whereas hair (geometric mean 177 ng g^{-1}, threshold 200 ng g^{-1}) and grain (geometric mean 7.8 ng g^{-1}, threshold 25 ng g^{-1}) contents would be classed as deficient, soil total selenium contents would not (geometric mean 171 ng g^{-1}, threshold 125 ng g^{-1}). There was a strong correlation between the selenium content of grain and the selenium status of the local population determined by hair sampling, but relationships with local soils were less clear. Further examinations into the soil geochemistry demonstrated that soils in the high KD prevalence villages were black or dark brown with a high organic matter content and lower pH than other soils in the region. Although these soils contained high total selenium contents, it was not in a readily bioavailable form as it was held in the organic matter in the soil. Despite the higher total selenium contents, water-soluble selenium in the high prevalence villages was in fact lower than deficiency threshold values (geometric mean 0.06 ng g^{-1}, threshold 3 ng g^{-1}). This study concluded that when environmental concentrations of selenium are low, any factor that is responsible for reducing the mobility of selenium may have a critical effect and emphasizes the importance of determining the bioavailability of selenium rather than the total selenium content when assessing impacts on human health. On the basis of this study, conditioning treatments to raise the soil pH thus increasing the bioavailability of selenium in the organic-rich soils or foliar application of selenium fertilizer to crops to avoid selenium adsorption in the soils were recommended as remediation strategies to increase the levels of selenium in local diets. The study also demonstrates the importance of understanding the biogeochemical environment in the determination of selenium-deficient regions and appropriate remediation techniques. However, here as with elsewhere in China, no incidences of KD have been reported in the area since 1996 as economic and communication improvements diversify the diet and enhance the health of the population.

16.6.4 The Geological Impact of Selenium on Human Health—Deficiency and Toxicity, Enshi District, Hubei Province, China

If there is one place on Earth that demonstrates the importance of geological controls on selenium and human health it is Enshi District, Hubei Province, in China, which lies approximately 100 km south of the Yangtze River Gorges and 450 km west-southwest of the provincial capital Wuhan (Fig. 16.8). In Enshi District, selenium-deficiency related diseases (Keshan disease) and selenium toxicity (selenosis) occur within 20 km of each other; their incidence is controlled by geology. The area is very mountainous with little connectivity between villages; some of which can only be reached on foot, hence, populations are very dependent on the local environment for their food supply.

Jurassic sandstones, which contain low concentrations of trace elements including selenium, underlie the northwest part of Enshi District and KD is present in this area. Selenium toxicity, on the other hand, is associated with high environmental selenium derived from Permian age coal-bearing strata in the center and east of the Enshi District. Soils developed over Jurassic sandstones comprise red-purple sands whereas light-brown silts and clays containing many carbonaceous fragments are typical in areas underlain by the Permian strata.

Studies into the selenium balance of local populations were carried out during the 1960s and 1970s by the Enshi Public Health Department in response to outbreaks of selenium-related diseases in the area. Between 1923 and 1988, 477 cases of human selenosis were reported. Of these cases, 338 resulted in hair and nail loss and disorders of the nervous system. They occurred between 1959 and 1963 in Shadi, Xin Tang and Shuang He communities. In Yu Tang Ba village, Shuang He community, the population was evacuated after 82% (19 out of 23 people) suffered nail and hair loss and all livestock died from selenium poisoning. During the same period, 281 selenosis cases were reported in five villages in the Shadi area. Cases of selenosis in pigs reached peak prevalence between 1979 and 1987 when 280 out of 2,238 animals were affected in Shatou, which resulted in 122 deaths. No human cases of selenium toxicity have been reported in recent years but animals commonly suffer hoof and hair loss as a result of the high environmental selenium.

During the late 1960s and 1970s, an area of selenium deficiency in Lichuan County to the northwest of Enshi District was also identified and lies within the KD belt across China. In total, 312 people have suffered KD in the county, an average incidence rate of 103 per 100,000. Among the 312 cases, 136 recovered, 163 died, and 13 persons still suffer from the disease. The village of Chang Ping was the worst affected with a total of 259 cases out of a population of 20,368 and 117 of those affected died. Children between the ages of 3 and 8 accounted for 83.4% of the total cases and 80% of the children affected by the disease died. Following peak prevalence in 1969 (106 cases), the number of cases has fallen dramatically and current prevalence rates are unknown as no medical investigations have been carried out in recent years.

Table 16.20 Estimated daily dietary intake of selenium in three areas of Enshi district, China. Compared to recommended intakes elsewhere

Source	Daily dietary intake of selenium (μg)
Enshi low Se and KD	62–70
Enshi high Se and no selenosis	194–198
Enshi high Se and selenosis	1238–1438
U.S. National Research Council	50–200 RDA
UK Dept of Health	60–200 RDA
China	40–600 RDA

From Yang et al. (1983), Tan (1989), WHO (1996), and MAFF (1997)
RDA recommended daily allowance

Yang et al. (1983) were the first to compare levels of selenium in soil, crops, drinking water, human urine, blood, nail, and hair samples from the Enshi area with other regions of China and demonstrate that the endemic selenium intoxication of humans in Enshi was related to the occurrence of Permian selenium-enriched shaley coal, which contains up to 6,471 mg kg^{-1} selenium. There is some evidence to suggest that selenium in these rocks is in the form of micro particles of elemental selenium in association with organic carbon and that the carbon content of the rock controls the selenium content. However, some selenium is also found in the lattice of pyrite minerals. Selenium concentrations in soil, food, and human samples from areas underlain by carbonaceous strata were up to 1,000 times higher than in samples from selenium-deficient areas where KD was prevalent and dietary intakes of selenium greatly exceed the U. S. NRC and Chinese recommended standards (Table 16.20). It was estimated that locally grown crops constituted 90% of the diet in the Enshi area and cereal crops (rice and maize) accounted for 65–85% of the selenium intake, which indicated the importance of the local environment to selenium in the food chain. In addition to exposure via soils and foodstuffs, villagers in the selenosis region also mine the carbonaceous shale for fuel and use burnt coal residues as a soil conditioner. Although the epidemiological investigations revealed that selenosis occurred in areas of high environmental selenium associated with the carbonaceous strata, not all villages underlain by this strata were affected.

Table 16.21 Selenium in soil, grain, and hair from Enshi district, China. Compared to selenium deficiency and toxicity threshold values

Threshold	Soil total Se (mg kg^{-1})	Soil water soluble Se (ng g^{-1})	Grain Se (mg kg^{-1})	Hair Se (mg kg^{-1})
Deficient	0.125	3	0.025	0.2
Toxic	3	20	1	3
Enshi area village	Geometric mean Se		Concentration ranges (n = 15)	
Low Se	0.069–0.199	0.21–0.44	0.001–0.003	0.252–0.345
High Se	2.07–19.54	2–61	0.041–2.902	0.692–29.21

From Tan (1989) and Fordyce et al. (2000b)

Further studies carried out by Fordyce et al. (2000b) into three groups of villages: one suffering KD, one with high environmental selenium but no selenosis, and one with high environmental selenium and selenosis. In villages with high selenium, concentrations in soils and foodstuffs could markedly vary from low to toxic within the same village, with these variations dependent on the outcrop of the coal-bearing strata. The wide range in geochemical conditions could in part explain why some villages suffered selenosis and others did not as did practices such as using coal ash as a soil conditioner (Li et al. 2012). Therefore, villagers were advised to avoid cultivating fields underlain by the coal and were counseled against using coal-derived products, such as ash, to condition the soil. In the KD affected villages of Lichuan County, selenium concentrations in staple food crops (rice and maize), drinking water, and the human populations (measured in hair samples) were very low and soils in this area had lower pH contents than soils in the high-selenium villages, which would further inhibit the uptake of selenium into plants. Conditioning the soil with lime to increase the pH making selenium more mobile was suggested as a remediation strategy.

Although all the villages in the low-selenium area had a marginally deficient selenium status and the majority of villages in the high-selenium area had excessive amounts of selenium in the environment and human population using the thresholds defined by Tan (1989) (Table 16.21), no new incidents of either KD or overt selenium poisoning have occurred in recent years. This suggests that the local population may have adapted to the high and low selenium intakes present in the different environments and that the historical occurrences of clinical effects related to selenium imbalances were caused by other factors. The outbreaks of human selenosis in Enshi during the late 1950s and early 1960s coincided with a drought and the failure of the rice crop. The crop failure had serious implications for the dietary intake and health of the local population with less food available, reduced protein intake, and higher dependence on vegetables and maize and natural plants. These factors may have lead to the severe outbreaks of selenosis in the Enshi area and demonstrate that in geologically controlled high- or low-selenium environments additional stresses can lead to serious health outcomes in the local population.

16.6.5 The Geological Impact of Selenium on Animal Health—Deficiency and Toxicity, Queensland, Australia

Another example of the effects of geology on selenium and health has been reported in Queensland, Australia. Here seleniferous limestones and shales of the Tambo Formation cause selenium toxicity symptoms in livestock grazing plants in this area, whereas less than 100 km to the south selenium deficiency and WMD in grazing animals is a problem over Tertiary volcanic soils. Grain grown over the seleniferous limestone rocks contains >0.2 mg kg^{-1} selenium whereas over the southern selenium-deficient region concentrations rarely exceed 0–0.05 mg kg^{-1} in grain (Oldfield 1999). This is another example of how geologically controlled geochemical variation can influence selenium status and health over relatively short distances.

16.6.6 Selenium Status in Western Countries—Is Environment Still Important?

With the exception of Italy, the human case studies presented in this chapter refer to developing countries where populations are very dependent on the local environment to provide the correct mineral balance. Under these circumstances it is easy to see why considerations of selenium status may be important. But what of the western world where people generally move around more during their lifetimes, products are derived from all around the world, and people buy food in large supermarkets rather than growing it in their own back yard? Under these conditions, the links between environment and health are less direct. Nonetheless, the impact of the selenium status of the environment on animals and humans is still evident. It has already been pointed out in this chapter that New Zealand is a generally selenium-deficient country compared to other areas of the world; indeed cereal grains grown in New Zealand contain 10 times less selenium than grain from Canada and the United States. In years when the crops in New Zealand are poor, wheat is imported from Australia and a corresponding notable increase in blood selenium levels is seen in the population. Indeed, the selenium status of the New Zealand

population has been improved by greater importation of high-selenium Australian wheat (Thomson 2004). Studies have also shown that the average total body contents of selenium are only 3–6.1 mg kg^{-1} in New Zealand compared to 14.6 mg kg^{-1} in the United States, and studies into individual tissues of the body show that in New Zealand, concentrations are half that of the United States. A marked lowering of blood selenium levels has been noted in populations moving from selenium-adequate areas of the United States to New Zealand; however, the actual resultant values also depend on factors such as physiological status (WHO 1987). There is clear evidence therefore of the influence of the geochemical environment on food and human selenium status when either food or people move from one area to another. But what about variation within a country? Perhaps the most compelling evidence that environmental differences are important even within western countries comes from the United States where, despite one of the most diverse and mobile food supply chains in the world, selenium concentrations in animals and humans reflect the surrounding environment. Studies have shown that despite the widespread use of agricultural management practices including selenium supplementation, the selenium content of skeletal muscle in cattle shows marked geographic variation concordant with selenium contents in soils and grasses and perhaps even more surprisingly, human blood selenium levels are higher in the seleniferous western United States than in selenium-poor areas. For example, serum selenium contents average 0.161 μg L^{-1} in Ohio compared to 0.265 μg L^{-1} in South Dakota (WHO 1987). Hence, even in populations who now live one step removed from their natural environment, the cycling of selenium from nature into humans is still of fundamental importance to health.

16.7 Future Considerations

This chapter has demonstrated that human exposure to the biologically important element selenium is largely dependent on dietary intakes in food and water, which are significantly controlled by variations in the geology of the Earth's surface. Although much work has been done over the past 40 years to enhance our understanding of environmental selenium, over large areas of the globe information is still missing because until recently selenium was a difficult element to analyze. More work is required to understand not just the total amounts of selenium present but also the bioavailability of the element and cycling through the environment. For example, it is only relatively recently that the importance of the oceans in the cycling of selenium has been recognized. The selenium status of the human and animal populations around the globe closely reflect environmental levels and although overt clinical symptoms of selenium toxicity and deficiency are rarely reported, the possible subclinical effects and implications of selenium status are only beginning to be understood and should not be underestimated as medical science continues to uncover new essential functions for the element. In the future, closer collaboration between medical and environmental scientists will be required to evaluate the real environmental health impact of this remarkable element in diseases such as cancer, AIDS, and heart disease.

See Also the Following Chapters. Chapter 2 (Natural Distribution and Abundance of Elements) • Chapter 7 (Biological Functions of the Elements) • Chapter 15 (Bioavailability of Elements in Soil) • Chapter 17 (Soils and Iodine Deficiency) • Chapter 21 (Animals and Medical Geology) • Chapter 25 (Environmental Pathology) • Chapter 33 (Modeling Groundwater Flow and Quality)

Further Reading

Afzal S, Younas M, Ali K (2000) Selenium speciation studies from Soan-Sakesar Valley, salt range, Pakistan. Water Int 25:425–436

Akesson B, Steen B (1987) Plasma selenium and glutathione peroxidase in relation to cancer, angina pectoris and short-term mortality in 68-year-old men. Comprehens Gerontol A 1:61–64

Alfthan G, Xu GL, Tan WH, Aro A, Wu J, Yang YX, Liang WS, Xue WL, Kong LH (2000) Selenium supplementation of children in a selenium-deficient area in China—blood selenium levels and glutathione peroxidase activities. Biol Trace Elem Res 73(2):113–125

Amouroux D, Liss PS, Tessier E, Hamren-Larsson M, Donard OFX (2001) Role of the oceans as biogenic sources of selenium. Earth Planet Sci Lett 189(3–4):277–283

Ander EL, Fordyce FM, Johnson CC, Schofield DI (2010) Controls on the distribution of selenium in the soils of East Anglia, UK. Geological society of America abstract section meeting, Baltimore, Maryland, 13–16 March, p 143

Anonymous (2001) Observations on the effect of sodium selenite in the prevention of keshan disease (Reprinted from the Chinese Med J 92:471–476, 1979). J Trace Elem Exp Med 14(2):213–219

Aras NK, Nazli A, Zhang W, Chatt A (2001) Dietary intake of zinc and selenium in Turkey. J Radioanal Nucl Chem 249(1):33–37

Aro A, Alfthan G, Varo P (1995) Effects of supplementation of fertilizers on human selenium status in Finland. Analyst 120(3): 841–843

Arthur JR, Beckett GT (1994) New metabolic roles for selenium. Proc Nutr Soc 53:615–624

ATSDR (2003) Toxicology profile for selenium. Agency for Toxic Substances and Disease Registry, Atlanta

Beck MA (1999) Selenium and host defense towards viruses. Proc Nutr Soc 58(3):707–711

Bedwal RS, Nair N, Sharma MP, Mathur RS (1993) Selenium—its biological perspectives. Med Hypotheses 41(2):150–159

Birt DF, Pour PM, Pelling JC (1989) The influence of dietary selenium on colon, pancreas and skin tumorigenesis. In: Wendel A (ed) Selenium in biology and medicine. Springer, Berlin, pp 297–304

Boyle RW (1979) The geochemistry of gold and its deposits: together with a chapter on geochemical prospecting for the element. Bulletin 280. Geological Survey of Canada, Ottawa

British Geological Survey (2000) Regional geochemistry of Wales and part of West-Central England: stream sediment and soil. British Geological Survey, Keyworth

British Geological Survey (2011) London Earth. British Geological Survey, Keyworth. http://www.bgs.ac.uk/gbase/londonearth.html. Accessed Mar 2012

Broadley MR, White PJ, Bryson RJ, Meacham MC, Bowen HC, Johnson SE, Hawkesford MJ, McGrath SP, Zhao FJ, Breward N, Harriman M, Tucker M (2006) Biofortification of UK food crops with selenium. Proc Nutr Soc 65:1–14

Broadley MR, Alcock J, Alford J, Cartwright P, Fairweather-Tait SJ, Foot I, Fairweather-Tait SJ, Hart DJ, Hurst R, Knott P, McGrath SP, Meacham MC, Norman K, Mowat H, Norman K, Stroud JL, Tovey M, Tucker M, White PJ, Young SD, Zhao FJ (2010) Selenium biofortification of high-yielding winter wheat (*Triticum aestivum* L.) by liquid or granular Se fertilisation. Plant Soil 32(1–2):5–18

Burk RF (ed) (1994) Selenium in biology and human health. Springer, New York

Cao ZH, Wang XC, Yao DH, Zhang XL, Wong MH (2000) Selenium geochemistry of paddy soils in the Yangtze river delta. Environ Int 26(5–6):335–339

Chen X, Yang G, Chen J, Chen X, Wen Z, Ge K (1980) Studies on the relations of selenium and keshan disease. Biol Trace Elem Res 2:91–107

Chilimba ADC, Young SD, Black CR, Rogerson KB, Ander EL, Watts MJ, Lammel J, Broadley MR (2011) Maize grain and soil surveys reveal suboptimal dietary selenium intake is widespread in Malawi. Sci Rep 1(72):1–9

Clark LC, Combs GF, Turnbill BW, Slate EH, Chalker DK, Chow J, Davies LS, Glover RA, Graham GF, Gross EG, Krongrad A, Lesher JL, Park HK, Sander BG, Smith CL, Taylor JR (1996) Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin. J Am Med Assoc 276:1957–1963

Combs GF, Combs SB (1986) The role of selenium in nutrition. Academic, Orlando

Cutter GA, Bruland KW (1984) The marine biogeochemistry of selenium: a re-evaluation. Liminol Oceanogr 29:1179–1192

Darnley AG, Björklund A, Bølviken B, Gustavsson N, Koval PV, Plant JA, Steenfelt A, Tauchid M, Xuejing Xie X (1995) A global geochemical database for environmental and resource management. Recommendations for international geochemical mapping, earth science report 19. UNESCO Publishing, Paris

Davies EB, Watkinson JH (1966) Uptake of native and applied selenium by Pasture species. N Z J Agric Res 9:317–324

Dhillon KS, Dhillon SK, Thind HS (2008) Evaluation of different agroforestry tree species for their suitability in the phytoremediation of seleniferous soils. Soil Use Manage 24:208–216

DOE (1996) Code of practice for agricultural use of sewage sludge. Department of the Environment, HMSO, London

EA (2009) Soil guideline values for selenium in soil. Science report SC050021. UK Environment Agency, Bristol

Fairweather-Tait SJ, Collings R, Hurst R (2010) Selenium bioavailability: current knowledge and future research requirements. Am J Clin Nutr 91:1484S–1491S

Fergusson JE (1990) The heavy elements: chemistry, environmental impact and health effects. Pergamon, London

Finley JW, Ip C, Lisk DJ, Davis CD, Hintze KJ, Whanger PD (2001) Cancer-protective properties of high-selenium broccoli. J Agric Food Chem 49(5):2679–2683

Fleming GA (1980) Essential micronutrients II: iodine and selenium. In: Davis BE (ed) Applied soil trace elements. Wiley, New York, pp 199–234

Fordyce FM, Zhang G, Green K, Liu X (1998) Soil, grain and water chemistry and human selenium imbalances in Enshi District, Hubei Province, China, overseas geology series technical report WC/96/54. British Geological Survey, Keyworth

Fordyce FM, Johnson CC, Navaratne URB, Appleton JD, Dissanayake CB (2000a) Selenium and iodine in soil, rice and drinking water in relation to endemic goiter in Sri Lanka. Sci Tot Environ 263 (1–3):127–142

Fordyce FM, Zhang G, Green K, Liu X (2000b) Soil, grain and water chemistry and human selenium imbalances in Enshi District, Hubei Province, China. Appl Geochem 15(1):117–132

Fordyce FM, Brereton N, Hughes J, Luo W, Lewis J (2010) An initial study to assess the use of geological parent materials to predict the Se concentration in overlying soils and in five staple foodstuffs produced on them in Scotland. Sci Tot Environ 408(22):5295–5305

Fordyce FM, Nice SE, Lister TR, Ó Dochartaigh BÉ, Cooper R, Allen M, Ingham M, Gowing C, Vickers BP, Scheib A (2012) Urban soil geochemistry of Glasgow, open report, OR/08/002. British Geological Survey, Edinburgh

Franke KW (1934) A new toxicant occurring naturally in certain samples of plant foodstuffs. 1. Results obtained in preliminary feeding trials. J Nutr 8:597–608

FSA (2009) Survey on measurement of the concentrations of metals and other elements from the 2006 UK total diet study. Food survey information sheet 01/09. Food Standards Agency, London

Gawalko E, Garrett RG, Warkentin T, Wang N, Richter A (2009) Trace elements in Canadian field peas: a grain safety assurance perspective. Food Add Contamin 26(7):1002–1012

GSNI (2012) The Tellus project. Geological Survey of Northern Ireland, Belfast. http://www.bgs.ac.uk/gsni/tellus/geochemical_survey/index.html. Accessed Mar 2012

Hartikainen H, Xue TL, Piironen V (2000) Selenium as an anti-oxidant and pro-oxidant in ryegrass. Plant Soil 225(1–2):193–200

Hawkes WC, Turek PJ (2001) Effects of dietary selenium on sperm motility in healthy men. J Androl 22(5):764–772

Haygarth PM (1994) Global importance and cycling of selenium. In: Frankenberger WT, Benson S (eds) Selenium in the environment. Marcel-Dekker, New York, pp 1–28

Hem J (1992) Study and interpretation of the chemical characteristics of natural water, water supply paper 2254. U. S. Geological Survey, Reston

Hira CK, Partal K, Dhillon KS (2004) Dietary selenium intake by men and women in high and low selenium areas of Punjab. Public Health Nutr 7(1):39–43

Jacobs LW (ed) (1989) Selenium in agriculture and the environment, Soil Science Society of America special publication 23. SSSA, Madison

Johnson CC, Ge X, Green KA, Liu X (2000) Selenium distribution in the local environment of selected villages of the keshan disease belt, Zhangjiakou District, Hebei Province, People's Republic of China. Appl Geochem 15(3):385–401

Johnson CC, Fordyce FM, Rayman MP (2010) Factors controlling the distribution of selenium in the environment and their impact on health and nutrition. Proc Nutr Soc 69(1):119–132

Kohrle J (1999) The trace element selenium and the thyroid gland. Biochimie 81(5):527–533

Levander OA (1986) Selenium. In: Mertz W (ed) Trace elements in human and animal nutrition. Academic, London, pp 139–197

Li Y, Peng T, Yang Y, Niu C, Archard LC, Zhang H (2000) High prevalence of enteroviral genomic sequences in myocardium from cases of endemic cardiomyopathy (keshan disease) in China. Heart 83(6):696–701

Li S, Xiao T, Zheng B (2012) Medical geology of arsenic, selenium and thallium in China. Sci Tot Environ 421–422:31–40

Lippman SM, Klein EA, Goodman PJ, Lucia MS, Thompson IM, Ford LG et al (33 authors) (2009) Effect of selenium and vitamin E on risk of prostate cancer and other cancers: the selenium and vitamin E cancer prevention trial (SELECT). J Am Med Assoc 301(1):39–51

Longombe AO, Arnaud J, Mpio T, Favier AE (1994) Serum selenium in HIV-infected Zairian patients. Trace Elem Electrolytes 11(2):99–100

Lyons MP, Papazyan TT, Surai PF (2007) Selenium in food chain and animal nutrition: lessons from nature. Asian-Aust J Anim Sci 20(7):1135–1155
MAFF (1997) Food surveillance information sheet of the Joint Food Safety and Standards Group, 126. Ministry of Agriculture Fisheries and Food, London
Mahapatra S, Tripathi RM, Raghunath R, Sadasivan S (2001) Daily intake of selenium by the adult population of Mumbai, India. Sci Tot Environ 277(1–3):217–223
Mayland HF (1994) Selenium in plant and animal nutrition. In: Frankenberger WT, Benson S (eds) Selenium in the environment. Marcel-Dekker, New York, pp 29–47
Moreno-Reyes R, Egrise D, Neve J, Pasteels JL, Schoutens A (2001) Selenium deficiency-induced growth retardation is associated with an impaired bone metabolism and osteopenia. J Bone Miner Res 16(8):1556–1563
Moxon AL (1937) Alkali disease or selenium poisoning, Bulletin 311. South Dakota Agricultural Experimental Station, Vermillion
Moxon AL (1938) The effect of arsenic on the toxicity of seleniferous grains. Science 88:81
Murphy J, Cashman KD (2001) Selenium content of a range of Irish foods. Food Chem 74(4):493–498
Muth OH, Allaway WH (1963) The relationship of white muscle disease to the distribution of naturally occurring selenium. J Am Vet Med Assoc 142:1380
Neal RH (1995) Selenium. In: Alloway BJ (ed) Heavy metals in soils. Blackie Academic & Professional, London, pp 260–283
Nriagu JO (1989) Occurrence and distribution of selenium. CRC Press, Boca Raton
Nriagu JO (1991) Heavy metals in the environment, vol 1. CEP Consultants, Edinburgh
Oldfield JE (1999) Selenium world atlas. Selenium-Tellurium Development Association, Grimbergen
Papp LA, Lu J, Holmgren A, Khanna KK (2007) From selenium to selenoproteins: synthesis, identity, and their role in human health. Antioxid Redox Signal 9:775–806
Peng A, Wang WH, Wang CX, Wang ZJ, Rui HF, Wang WZ, Yang ZW (1999) The role of humic substances in drinking water in Kashin-Beck disease in China. Environ Health Perspect 107 (4):293–296
Pineda-Valdes G, Bullerman LB (2000) Thermal stability of moniliformin at varying temperature, pH and time in an aqueous environment. J Food Prot 63(11):1589–1601
Plant JA, Kinniburgh D, Smedley P, Fordyce FM, Klinck B (2004) Chapter 9.02: Arsenic and selenium. In: Lollar SB (ed) Environmental geochemistry, vol 9, Treatise on geochemistry series, Holland HD, Turekain KK (Exec Eds). Elsevier, Amsterdam, pp 17–66
Rapant S, Vrana K, Bodis D (1996) Geochemical atlas of Slovakia: Part 1 Groundwater. Geological Survey of the Slovak Republic, Bratislava
Rawlins BG, McGrath SP, Scheib AJ, Breward N, Cave M, Lister TR, Ingham M, Gowing C, Carter S (2012) The advanced soil geochemical atlas of England and Wales. British Geological Survey, Keyworth
Rayman M (2008) Food-chain selenium and human health: emphasis on intake. Br J Nutr 100:254–268
Rayman M (2012) Selenium and human health. Lancet 379:1256–1268
Rayman M, Goenaga Infante H, Sargent M (2008) Food-chain selenium and human health: spotlight on speciation. Br J Nutr 100:238–253
Reilly C (2006) Selenium in food and health. Springer, New York
Reimann C, Caritat P (1998) Chemical elements in the environment. Springer, Heidelberg
Romero CD, Blanco FL, Sanchez PH, Rodriguez E, Majem LS (2001) Serum selenium concentration in a representative sample of the Canarian population. Sci Total Environ 269(1–3):65–73
Rosenfield I, Beath OA (1964) Selenium, geobotany, biochemistry, toxicity and nutrition. Academic Press, New York
Rotruck JT, Pope AL, Ganther HE, Swanson AB, Hafeman DG, Hoekstra WG (1969) Selenium: biochemical role as a component of glutathione peroxidase. Science 179:588–590
Ruston D, Henderson L, Gregory J, Bates CJ, Prentice A, Birch M et al (2004) Nutritional status (anthropometry and blood analytes), blood pressure and physical activity. In: The national diet and nutrition survey: adults aged 19–64 years, vol 4. The Stationary Office, London
Sasaki S, Iwata H, Ishiguron N, Habuchi O, Miura T (1994) Low-selenium diet, bone and articular-cartilage in rats. Nutrition 10 (6):538–543
Schwarz K, Foltz CM (1957) Selenium as an integral part of factor 3 against dietary necrotic liver degeneration. J Am Chem Soc 79:3292–3293
Shamberger RJ, Frost DV (1969) Possible protective effect of selenium against human cancer. Can Med Assoc J 100:682
Shimada T, Murayama N, Tanaka K, Takenaka S, Guengerich FP, Yamazaki H, Komori M (2011) Spectral modification and catalytic inhibition of human cytochromes P450 1A1, 1A2, 1B1, 2A6, and 2A13 by four chemopreventive organoselenium compounds. Chem Res Toxicol 24(8):1327–1337
Smith MI, Westfall BB (1937) Public health report 52. U. S. Public Health Department, Washington, DC
Suetens C, Moreno-Reyes R, Chasseur C, Mathieu F, Begaux F, Haubruge E, Durand MC, Neve J, Vanderpas J (2001) Epidemiological support for a multifactorial etiology of Kashin-Beck disease in tibet. Int Orthop 25(3):180–187
Tan J (ed) (1989) The atlas of endemic diseases and their environments in the People's Republic of China. Science Press, Beijing
Tan J, Wang W, Yang L, Zhu W, Li R, Hou S (1999) The approaches to amelioration of selenium eco-cycle in selenium-deficient biogeochemical area and its effects on selenium dietary intake and health in China. In: Wenzel WW, Adriano DC, Alloway BE, Doner HE, Keller C, Lepp M, Mench R, Naidu R, Pierzynski GM (eds) Proceedings of the 5th international conference on the biogeochemistry of trace elements 2, Vienna, Austria, pp 620–621
Taylor SR, McLennan SM (1985) The continental crust: its composition and evolution. Blackwell, Oxford
Thomson CD (2004) Selenium and iodine intakes and status in New Zealand and Australia. Br J Nutr 91:661–672
Thornton I, Kinniburgh DG, Pullen G, Smith CA (1983) Geochemical aspects of selenium in British soils and implications to animal health. In: Hemphill DD (ed) Trace substances in environmental health XVII. University of Missouri, Columbia, pp 391–398
Tokunaga T, Pickering I, Brown G (1996) Selenium transformations in ponded sediments. Soil Sci Soc Am J 60:781–790
U.S.-EPA (2002) Toxic release inventory 2002. United States Environmental Protection Agency, Washington. Accessed via Scorecard. org. http://scorecard.goodguide.com/. Accessed Apr 2012
U.S.-EPA (2004) Draft aquatic life water quality criteria for selenium – 2004. United States Environmental Protection Agency, Washington, DC
U.S.-EPA (2012) Basic information about selenium in drinking water fact sheet. United States Environmental Protection Agency, Washington, DC
USGS (2007) 2006 minerals yearbook – selenium and tellurium. United States Geological Survey, Reston, Washington, DC
USGS (2011) 2010 minerals yearbook – selenium and tellurium. United States Geological Survey, Reston, Washington, DC
Varo P, Alfthan G, Ekholm P, Aro A (1998) Effects of selenium supplementation fertilizer on human nutrition and selenium status. In: Frankenberger WT, Engberg RA (eds) Environmental chemistry of selenium. Marcel-Dekker, New York
Vinceti M, Rothman KJ, Bergomi M, Borciani N, Serra L, Vivoli G (1998) Excess melanoma incidence in a cohort exposed to high levels of environmental selenium. Cancer Epidemiol Biomarkers Prev 7:853–856

Vinceti M, Rovesti S, Bergomi M, Vivoli G (2000) The epidemiology of selenium and human cancer. Tumori 86:105–118

World Health Organization (1987) Environmental health criterion 58—selenium. World Health Organization, Geneva

World Health Organization (1996) Trace elements in human nutrition and health. World Health Organization, Geneva

World Health Organization (2011) Guidelines for drinking water quality, 4th edn. World Health Organization, Geneva

Wu L (1994) Selenium accumulation and colonization of plants in soils with elevated selenium and salinity. In: Frankenberger WT, Benson S (eds) Selenium in the environment. Marcel-Dekker, New York, pp 279–326

Wu AS, Oldfield JE, Shull LR, Cheeke PR (1979) Specific effect of selenium deficiency on rat sperm. Biol Reprod 20:793–798

Wu L, Banuelos G, Guo X (2000) Changes of soil and plant tissue selenium status in an upland grassland contaminated by selenium-rich agricultural drainage sediment after ten years transformed from a wetland habitat. Exotoxicol Environ Saf 47(2):201–209

Xu G, Jiang Y (1986) Selenium and the prevalence of Keshan and Kashin-Beck diseases in China. In: Thornton I (ed) Proceedings of the first international symposium on geochemistry and health. Science Reviews Ltd., Northwood, pp 192–205

Yamashita Y, Yamashita M (2010) Identification of a novel selenium-containing compound, selenoneine, as the predominant chemical form of organic selenium in the blood of Bluefin Tuna. J Biol Chem 285(24):18134–18138

Yang G, Xia M (1995) Studies on human dietary requirements and safe range of dietary intakes of selenium in China and their application to the prevention of related endemic diseases. Biomed Environ Sci 8:187–201

Yang G, Wang S, Zhou R, Sun S (1983) Endemic selenium intoxication of humans in china. Am J Clin Nutr 37:872–881

Yang G, Yin S, Zhou R, Gu L, Yan B, Lui Y, Liu Y (1989) Studies of safe and maximal daily dietary selenium intake in a seleniferous area in China. Part 2. J Trace Elem Electrolytes Health Dis 3:123–129

Zhang WH, Neve J, Xu JP, Vanderpas J, Wang ZL (2001) Selenium, iodine and fungal contamination in Yulin district (People's Republic of China) endemic for Kashin-Beck disease. Int Orthop 25 (3):188–190

Soils and Iodine Deficiency

17

Ron Fuge

Contents

17.1 Introduction

Iodine has long been known as an essential element for humans, and mammals in general, where it is concentrated in the thyroid gland. It is a component of the thyroid hormone thyroxine. Deprivation of iodine results in a series of iodine deficiency disorders (IDD), the most obvious of which is endemic goiter, a condition where the thyroid gland becomes enlarged in an attempt to be more efficient. Iodine deficiency during fetal development and in the first year of life can result in endemic cretinism, a disease which causes stunted growth and general development along with brain damage. However, while these two diseases are easily recognizable, perhaps the more insidious problem is that iodine deficiency impairs brain development in children even when there is no obvious physical effect. Many researchers have suggested that a relatively low degree of iodine deficiency during fetal development can result in a significant reduction of IQ in children. Indeed it has been suggested that iodine deficiency is the most common preventable cause of mental retardation (see Chaps. 8 and 9, this volume). For a detailed discussion of the problems resulting from iodine deficiency in humans the reader is referred to the many publications on the topic by Basil Hetzel (e.g. Hetzel 1987, 2001; Hetzel et al. 1990).

Endemic goiter and cretinism along with related IDD have long been recognized as serious health problems and consequently much work has been carried out on the etiology and geographical distribution of these diseases. Many authors have suggested the involvement of several other elements in the etiology of these diseases, whereas a group of sulfur-containing compounds, thiocyanates and thiouracils, identified collectively as goitrogens, have been strongly implicated in some endemics. These compounds have been found to either inhibit iodine uptake by the thyroid gland or inhibit the formation of the thyroid hormones. However, it is generally agreed that the primary cause of IDD is a lack of iodine in the diet.

R. Fuge (✉)
Institute of Geography and Earth Sciences, Aberystwyth University, Aberystwyth, UK
e-mail: rrf@aber.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_17,

Table 17.1 Iodine in igneous and sedimentary rocks

Rock type	Mean iodine content (mg kg^{-1})
Igneous rocks	
Granite	0.25
All other intrusives	0.22
Basalts	0.22
All other volcanics	0.24
Volcanic glasses	0.52
Sedimentary rocks	
Shales	2.3
Sandstones	0.80
Limestones	2.3
Organic-rich shales	16.7

Iodine was the first element recognized as being essential to humans and the disease of endemic goiter was the first to be related to environmental geochemistry. Indeed, this disease appears to have been identified by the ancient Chinese. One Chinese medical writer from the fourth century AD noted the use of the seaweeds *Sargassum* and *Laminaria* (which are known to be very iodine rich) for treatment of goiter (Langer 1960). However, there are many earlier records in ancient literature of seaweeds and burnt sea sponges used in the treatment of endemic goiter possibly from as early as 2700 BC (Langer 1960).

Iodine was discovered by the French chemist Bernard Courtois in 1811 when he accidentally added concentrated sulfuric acid to the seaweed *Fucus vesiculosus*, one of the seaweeds used in goiter treatment. It was soon realized that iodine was the active ingredient in the treatment. Iodine was later identified as an essential element in human nutrition. Despite this early recognition of the role of iodine in endemic goiter and related disorders, it is apparent that IDD is still affecting large numbers of people worldwide, with Zimmermann (2010) suggesting that around two billion people are iodine deficient. While the majority of the iodine-deficient populations are in the developing world, with Africa and Asia particularly badly affected, it is apparent that iodine deficiency problems are re-emerging in the industrialized world, with reports that iodine deficiency is currently affecting populations in Australia, New Zealand, USA and Europe.

The areas where IDD are concentrated tend to be geographically defined. Thus many of the most severe occurrences of endemic goiter and cretinism have been found to occur in high mountain ranges, rain shadow areas, and central continental regions (Kelly and Sneddon 1960). This distribution of IDD results from the unique geochemistry of iodine. Geochemists generally agree that little iodine in the secondary environment is derived from weathering of the lithosphere. Most iodine is derived from volatilization from the oceans with subsequent transport onto land (Goldschmidt 1954). Therefore, to understand the role of soil iodine geochemistry in IDD it is essential to consider some aspects of the general geochemistry of iodine and its cyclicity in the environment.

17.2 Iodine Geochemistry

Iodine has been shown to be concentrated in seawater, in the biosphere, and in the atmosphere. Because of this, it is classified as a hydrophile, biophile, and atmophile element (Goldschmidt 1954). In addition iodine has been found to be concentrated in sulfur-containing minerals causing it to be classified as a chalcophile element.

17.2.1 The Lithosphere

In the lithosphere iodine is an ultra-trace element; its crustal abundance is estimated to be 0.25 mg kg^{-1} by Fuge (1988) and more recently 0.3 mg kg^{-1} by Muramatsu and Wedepohl (1998). Due to the large ionic radius of the iodide ion (220 pm), it is thought unlikely that iodine enters the crystal lattices of most rock-forming minerals. In fact the iodine content of the various rock-forming minerals has been shown to be fairly uniform and low, with some enrichment found in only the chlorine-containing minerals sodalite and eudialyte. Its distribution in the different igneous rocks is fairly uniform (Table 17.1) with only volcanic glasses showing comparative enrichment. Some data suggest that carbonatites are also somewhat richer in iodine than average igneous rocks.

Sedimentary rocks show a greater range of iodine content with clay-rich or argillaceous rocks more enriched than the sand-rich, arenaceous rocks (Table 17.1). The highest concentrations of iodine have been found in organic-rich shales, with concentrations as high as 44 mg kg^{-1} recorded in some bituminous shales. The iodine content of carbonate rocks (limestones) and shales is highly variable but generally correlates with the amount of organic matter.

Recent sediments of marine origin can be extremely enriched in iodine with concentrations of up to 2,000 mg kg^{-1} recorded in some surficial sediments (Wong 1991).

It seems likely that the iodine content of metamorphic rocks is similar to that of igneous rocks. Analytical data produced by Muramatsu and Wedepohl (1998) suggest that metamorphic rocks have uniformly low iodine contents of <0.025 mg kg^{-1}. The same authors suggest that felsic igneous rocks are also very low in iodine with <0.009 mg kg^{-1}. It is suggested that the low concentrations are due to volatilization of iodine during the formation of these rocks.

17.2.2 The Marine Environment

Seawater is by far the biggest reservoir of iodine; its average concentration is about 60 μgL^{-1} (Wong 1991). Iodine is thought to have fairly uniform concentrations with depth, but it is slightly depleted in surface ocean waters because it

is concentrated by organisms. Iodine can exist in several forms in seawater and the transformations between the various forms and their mechanisms are vital for understanding the cycling of iodine in the general environment.

Inorganic iodine is essentially present in two forms, the iodide anion, I^-, and the iodate anion, IO_3^-. Iodate is the thermodynamically stable form of inorganic iodine in oxygenated, alkaline seawater, whereas iodide, the reduced species, is in a metastable state. There is considerable variation of the I^-/IO_3^- ratio with depth with I^- enriched in surface waters and depleted in deeper waters. There is also considerable geographic variation of the I^-/IO_3^- ratio with waters from the shallower inner shelf areas containing more iodide than the deeper mid-shelf waters. The conversion of IO_3^- to I^- is thought to be due to biological activity, with the enzyme nitrate reductase implicated in the reaction. The increased iodide in surface and shallow shelf waters is thus thought to be due to the high biological activity in these zones. However, it is possible that some abiological mechanisms might also be involved in the conversion of iodate to iodide. Once formed, iodide is only slowly re-oxidized to iodate (Wong 1991).

It has long been suggested that some iodine in seawater is present as an organic phase, and it has been found that in some coastal waters dissolved organic iodine can constitute up to 40% of the total iodine content (Wong and Cheng 1998). Although Wong (1991) has indicated that few specific organo-iodine compounds have been identified in seawater, one of the substances identified, methyl iodide (CH_3I), is of interest due to its volatility and possible role in the transfer of iodine from the oceans to the atmosphere. Lovelock et al. (1973) were the first researchers to detect methyl iodide in seawater. This compound can be formed both biologically, by seaweeds and phytoplankton, and by photochemical reactions, with the likelihood that chemical formation of methyl iodide is more important in the open ocean. More recently it has been found that methyl iodide is oversaturated in surface ocean waters (Moore and Groszko 1999). In addition to methyl iodide, Amachi et al. (2005) have reported that CH_2I_2 and C_2H_5I are also produced by marine organisms and added to the atmosphere.

The strong enrichment of iodine in marine organisms has long been recognized with Courtois discovering the element in the brown alga *F. vesiculosus*. Since that time brown algae have been shown to strongly concentrate iodine with an enrichment factor of alga/seawater having been estimated as being over 32,000 (Fuge and Johnson 1986). Although concentration factors for red and green algae are lower, they are still enriched in iodine. Phytoplankton also concentrates iodine.

It is generally held that the iodide ion is preferentially incorporated into the organisms; however, it has been demonstrated that phytoplankton can take in iodate (Moison et al. 1994). Whatever the mechanism of uptake, these marine organisms play an important role in transformation of iodine species and ultimately in its transfer from the oceanic environment to the atmosphere. It has been demonstrated that organo-iodine compounds are released from brown seaweeds and this could represent a source of atmospheric iodine.

17.2.3 Transfer of Iodine from the Marine to Terrestrial Environment via the Atmosphere

Iodine has been shown to be concentrated in the atmosphere. This atmophile behavior is the most important part of its geochemical cycle, with the transfer of iodine from the oceans to the atmosphere ultimately governing its distribution in the terrestrial environment.

The mechanism of transfer of iodine from the oceans to the atmosphere has been a subject of some debate. Undoubtedly some iodine is transferred to the atmosphere as seawater spray. However, it has been shown that the amount of iodine relative to chlorine in the atmosphere is several hundred times that in seawater. Although some of the iodine is likely to be in aerosols, it has also been shown that a large percentage of atmospheric iodine is in a gaseous form (Duce et al. 1973). Therefore, the major mechanism of iodine transfer must reflect preferential volatilization of seawater iodine into the atmosphere.

The iodide ion can be converted to elemental iodine (I_2) by photochemical oxidation, and this has been proposed as a mechanism for volatilization of iodine from the oceans. More recently it has been proposed that I_2 (and possibly HOI) could be produced at the sea surface from oxidation of iodide by such species as ozone and nitrogen dioxide (Garland and Curtis 1981; Heumann et al. 1987). However, some workers have suggested that volatilization of elemental iodine is not likely to be a major source of transfer of iodine from the oceans to the atmosphere.

As methyl iodide has been found to occur in the surface waters of the sea and as this compound is volatile, it could represent a significant source of atmospheric iodine. In addition, CH_2I_2 and C_2H_5I are also volatilized from the marine environment. Yoshida and Muramatsu (1995) found that on the Japanese coast 90% of atmospheric iodine is gaseous with organically bound iodine as the dominant species. In addition, Gabler and Heumann (1993) found that organically bound iodine is the most abundant species of iodine in European air. In the atmosphere methyl iodide is broken down and has an estimated lifetime of about 5 days (Zafiriou 1974).

Whereas several authors have suggested that methyl iodide is the dominant form of iodine released from the oceans, it seems likely that there are several mechanisms of transfer of iodine from the oceans to the atmosphere and

ultimately to the landmasses. From studies in Antarctica, Heumann et al. (1990) suggested that during short distance transport into coastal areas iodine is carried as I_2, HI, and sea spray; for long distance transport CH_3I is responsible. Recently, Hou et al. (2009) have reported that the following iodine species occur in the atmosphere I_2, HI, HIO, CH_3I, CH_2I_2, and $CH_2CH_2CH_2I$. The residence time of inorganic iodine in the atmosphere has been estimated to be 10 days (Rahn et al. 1976; Chameides and Davis 1980) and the total residence time of iodine has been estimated to be 15 days (Kocher 1981).

Iodine is transferred from the atmosphere to the terrestrial environment by wet and dry deposition. On the basis of the marine origin of atmospheric iodine it might be expected that deposition should be highest at the coast and decrease inland. However, literature data are limited on the iodine content of rain from coastal and inland locations. Several workers have quoted data for the iodine content of rainfall in the UK, the whole of which has a strong maritime influence. Generally the samples were found to contain around 2 μgL^{-1} iodine with some higher values (up to 5 μgL^{-1}) recorded in coastal rainfall. It has been shown that rain collected from an upland area 12 km inland from the mid-Wales coast contained over three times as much iodine as rainfall from 84 km inland. Although some studies of iodine contents of rainfall have failed to demonstrate significantly higher values in coastal rain, analyses of rainfall from the continental United States (Missouri) showed iodine contents of <1.0 μgL^{-1}, and it does seem likely from the limited data that the iodine content of rain is influenced by its proximity to the oceans.

The form of iodine in rainfall has been the subject of much discussion. In general it has been shown that iodide is the most common form of iodine in rain making up over 50% of rainfall iodine, and the iodate ion is the second major component. It has also been demonstrated that the iodate content of rainfall decreases inland with iodide content showing a parallel increase. In view of the likely importance of organically bound forms in atmospheric transport of iodine, it would seem likely that some iodine in rain should be present as organo-iodine, with an early study suggesting as much as 40% is organically bound (Dean 1963). However, more recently, only small amounts of "non-ionic" dissolved iodine have been found to occur in Japanese rain; iodide and iodate are the dominant forms (Takagi et al. 1994). It seems likely that any atmospheric methyl iodide is converted to inorganic forms before deposition.

Dry deposition of iodine could be an important transfer mechanism for iodine from the atmosphere to the land surface. Few studies have attempted to quantify the amount of iodine deposited on land surfaces by dry deposition. Those studies suggest that dry deposition in marine influenced areas is a significant source of terrestrial iodine. However, there is no agreement on the relative importance of wet and dry deposition, and little is known with regard to the quantities of iodine in dry deposition in areas remote from the sea.

17.3 Iodine Geochemistry of Soils

There is a considerable body of data for iodine in soils and this shows a very broad range of concentrations from <0.1–150 mg kg^{-1}. The iodine content of soils is generally considerably higher than the rocks from which they derive. Most geochemists agree that the majority of the iodine in soils is derived from the atmosphere and ultimately the marine environment. The proximity of an area to the sea, therefore, is likely to exert a strong influence on the iodine content of soils in that area. This also results in considerable geographic variation of soil iodine content. The other important feature of soil iodine geochemistry is the fact that the element can become strongly adsorbed by various soil components, and thus its concentration and behavior in soils is going to depend on soil composition. In this respect the nature of the composition of the soil parent material indirectly exerts a strong influence on the iodine chemistry of the soil. The iodine geochemistry of soils then can be summarized as dependent on the quantity of iodine supplied coupled with the soil's ability to retain this iodine.

17.3.1 Factors Influencing the Supply of Iodine to Soils

It has generally been suggested that soils in close proximity to the coast are likely to be enriched in iodine with those far removed from the coast depleted of iodine (Goldschmidt 1954) (see also Table 17.2). However, in some cases such a relationship is not obvious and some workers have found no correlation between iodine and distance from the sea. This has led some workers to suggest that marine influence extends a considerable distance inland. Whitehead (1984), for instance, suggested that all soils from the UK are affected by the strong maritime influence over the whole country and that low iodine soils would only be found in the middle of continental areas. To some extent this is true, but it is perhaps pertinent to point out that in a soil traverse from the Welsh coast 120 km inland (Fig. 17.1), it was found that samples from beyond 100 km (8 samples) contained between 2 and 3 mg kg^{-1} of iodine. Similarly, a traverse from the Irish coast showed that soil samples collected beyond 80 km inland contained between 1.7 and 2.8 mg kg^{-1}. These values are not much higher than concentrations in soils from the continental United States and China (Table 17.2). Similar patterns of iodine in soils with increasing distance from the coast have been demonstrated for

Table 17.2 Iodine in coastal and inland soils

Sample origin	Iodine content (mg kg^{-1}) Range	Mean	References
Coastal areas			
Northwest Norway	5.4–16.6	9.0	Låg and Steinnes (1976)
Wales	1.5–149	14.7	Fuge (1996)
Ireland	4.2–54	14.7	Fuge (unpublished)
Inland areas			
East Norway	2.8–7.6	4.4	Låg and Steinnes (1976)
Wales/England	1.8–10.5	4.2	Fuge (1996)
Missouri, U. S.	0.12–5.6	1.3	Fuge (1987)
Whole of U. S.[a]	<0.5–9.6	1.2	Shacklette and Boerngen (1984)
Xinjiang Province, China	<0.3–3.9	1.05	Fordyce et al. (2003)

[a]Includes some from coastal localities

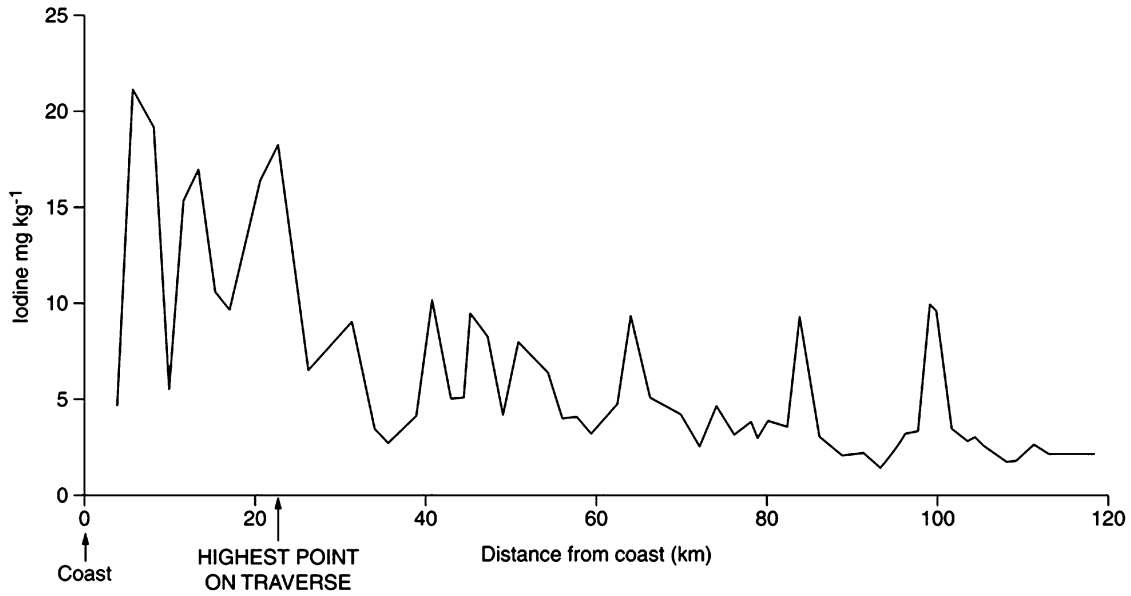


Fig. 17.1 Iodine in topsoils on a traverse from the west Wales coast to the English Midlands (After Fuge 1996)

Morocco with values dropping dramatically over the first 150 km (Johnson et al. 2002). On this basis it is quite likely that the strong influence of the marine environment on soil iodine contents does not extend very far inland, probably as little as 100–150 km.

The influence of topography on the iodine concentration of soils is well illustrated in Fig. 17.1. The highest values in the traverse occur in soils over the Welsh Mountains, which range up to 350 m (Ordinance Datum). In addition, several of the higher values in the latter half of the traverse occur on high ground. The greater precipitation that occurs in the upland areas causes a greater degree of washout of atmospheric iodine and hence a higher input of iodine. An additional factor is that in the area of the traverse, upland soils tend to be organic rich and thus are more able to retain the increased iodine input (see below).

The corollary of washout of iodine in upland areas is the low supply of iodine to rain shadow areas beyond. This is going to be particularly pronounced in the rain shadow areas of high mountain ranges such as the Himalayas and the Alps.

17.3.2 Factors Influencing the Retention of Iodine in Soils

As stated earlier soil iodine geochemistry reflects both the input of iodine and the ability of the soil to retain it. Many factors have been implicated in the retention of iodine in soil. Iodine is strongly enriched in organic-rich sediments and it seems likely that one of the most influential soil components with regard to retention of iodine is organic matter. Soils rich in organic matter are frequently enriched

Table 17.3 Iodine content of various soil types

Soil type	Iodine range (mg kg^{-1})	Mean	References
Peats (70% organic matter), UK	28–98	56	Johnson (1980), Fuge and Johnson (1986)
Peats, UK	18.7–98.2	46.8	Whitehead (1984)
Peaty tundra soils, Russian Plain	0.2–42	12.0	Vinogradov (1959)
Non-peaty soils, Russian Plain	0.3–9.8	2.8	Vinogradov (1959)
Iron-rich soils, UK	7.5–32.5	16.0	Fuge and Ander (1998) and unpublished information
Chalk and limestone parent material, UK	7.9–21.8	13.0	Whitehead (1984)
Over limestone, Derbyshire, UK	2.58–26.0	8.2	Fuge (1996)
Over shale, sandstone, and dolomite, Derbyshire, UK	1.88–8.53	3.44	Fuge (1996)
Carbonate-rich soils, Austria	1.64–5.63	3.75	Gerzabek et al. (1999)
Carbonate-free soils, Austria	1.08–4.80	2.58	Gerzabek et al. (1999)
Clay parent material, UK	2.1–8.9	5.2	Whitehead (1984)
Sand and sandstone parent material, UK	1.7–5.4	3.7	Whitehead (1984)
River and terrace alluvium parent material, UK	0.5–7.1	3.8	Whitehead (1984)
Marine estuarine alluvium parent material, UK	8.8–36.9	19.6	Whitehead (1984)

in iodine with the concentration of iodine correlated with the content of organic matter. Peaty soils are particularly enriched in iodine (see also Table 17.3).

Whereas organic matter has been shown to be the major contributor to the retention of soil iodine, it has also been suggested that iron and aluminum oxides play important roles in soil iodine retention (Whitehead 1974). With regard to iron oxide it is noteworthy that the weathered surfaces of iron meteorites have been found to be strongly enriched in iodine (Heumann et al. 1990). In addition, it has been found that iodine is concentrated in iron-rich soils (Table 17.3). In the case of aluminum oxide, it has been demonstrated in experimental work with both aluminum and iron oxides that sorption of iodide by aluminum oxide is similar to that by iron oxide (Whitehead 1974). The sorption of iodide by aluminum and iron oxides is strongly dependent on soil pH with sorption greatest in acid conditions, which is typical of anion adsorption. It has also been shown that iodate will be strongly sorbed by aluminum and iron oxides; however, this ion is not sorbed by organic matter. However, it has been demonstrated that both iodide and iodate added to organic-rich soil are converted to organic forms of iodine and subsequently strongly held in soil (Shimamoto et al. 2011).

Clay minerals have been thought to be involved in retention of soil iodine with the suggestion that the sorption of iodine to clays is also pH dependent (Prister et al. 1977). However, it is generally held that clay minerals are relatively unimportant in soil iodine retention, with organic matter and aluminum and iron oxides providing the bulk of the retentive capacity of soils.

The ability of soil to retain iodine has been called the iodine fixation potential (IFP) by Fuge and Johnson (1986). Thus soils that are rich in one or more of organic matter, iron oxides, and aluminum oxides are likely to have a high IFP, whereas those with low amounts of the major fixation components have a low IFP (Fig. 17.2).

Whereas little iodine is derived from the weathering of bedrock, the parent material governs the type of soil formed. In this context, the bedrock can exert a strong influence on the iodine retention capacity of soils (Table 17.3). Thus in the case of soils derived from sand-rich parent material, iodine contents tend to be low as the sandy soils derived will have little ability to trap iodine. A particularly interesting lithological control on the iodine content of soils occurs in areas underlain by carbonate rocks. In such soils, which

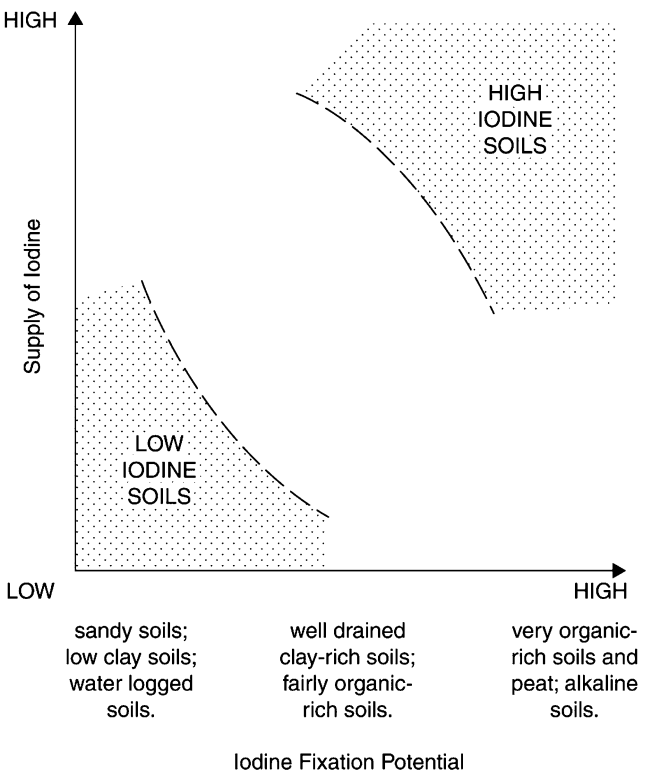


Fig. 17.2 A simplified model for iodine content of soils based on supply and fixation potential (Modified from Fuge and Johnson 1986)

generally have alkaline or circum-neutral pH values, iodine contents are generally elevated. This association of iodine with carbonate bedrocks is particularly well illustrated when comparing the iodine content of soils overlying limestone with neighboring soils overlying non-limestone lithologies as illustrated for the Derbyshire area of the U.K. (Table 17.3). A similar relationship has been demonstrated for carbonate-rich versus non-carbonate soils from Austria (Gerzabek et al. 1999) (see Table 17.3). It has been shown that the soils overlying limestone have distinctly higher iodine contents and hence have a high fixation potential.

Whereas soils do not normally derive much iodine from their parent materials, where the parent material is sediment of fairly recent marine origin and likely to be iodine-rich, the soils derived could inherit some of their iodine from the parental source (Table 17.3). In addition, it is likely that soils recently inundated by marine incursions or those that occur over reclaimed marine areas will be high in iodine.

As iodine is generally strongly sorbed within soils, very little is in a water-soluble form. Various researchers have found that up to 25% is in a soluble form with the water-soluble content of the majority of soils being less than 10% (Johnson 1980). In most high-iodine, organic-rich soils water-soluble iodine accounts for much less than 10% of the total. More recently, a study of German soils revealed that 2.5–9.7% of the iodine was water soluble; however, in the same soils ^{129}I, recently added from a nuclear reprocessing plant, was considerably more soluble (21.7–48.7%). This suggests that natural iodine had become strongly bound through time (Schmitz and Aumann 1994). For most soils it can be confidently predicted that water-soluble iodine will account for only a small percentage of the total, but in arid areas alkaline soils are likely to contain more elevated amounts of water-soluble iodine.

Whereas iodine sorbed in soils is generally strongly held and is not easily desorbed, it has been suggested that iodine is strongly desorbed in waterlogged soils. Thus Yuita et al. (1991) demonstrated that under flooding, and the resultant reducing conditions, 2–3 times as much iodine is solubilized from soils as in dry, oxidizing conditions. Such waterlogged, reducing conditions are typical of rice paddies and strong desorption of iodine has been found to occur in these soils (Muramatsu et al. 1996). Similarly lowland Japanese soils are low in iodine. It is suggested that this is due, in part, to flooding and desorption from the reducing conditions occasioned by microbial activity (Muramatsu and Yoshida 1999).

17.3.3 Chemical Form of Iodine in Soils

As outlined in the previous section of this chapter, in most soils iodine is strongly bound to organic matter and iron and aluminum oxides. However, it is of interest to establish the form of the soluble or easily leached iodine in soils as this is the fraction which should be plant available. Several authors have found that the iodide ion is the dominant form of soluble soil iodine in acidic soils, particularly in waterlogged soils. However, in dry, oxidizing conditions iodate was found to be the dominant form of soluble iodine. The dominance of iodide in acidic soils, with iodate dominating in alkaline soils, demonstrates the importance of pH in governing the form of soluble iodine. However, Eh will also exert an important control on the form of soil iodine. It has also been demonstrated that Fe^{3+} and SO_4^{2-} reducing bacteria found in soils are capable of reducing iodate to iodide (Councell et al. 1997). It has been shown that inorganic iodine added to organic-rich soil is converted to organic forms (Yamaguchi et al. 2010).

Despite some conflicting evidence in the literature, from the Eh-pH diagram for iodine (Fig. 17.3), it seems likely that in acid soils soluble iodine will predominate as iodide, whereas in alkaline soils iodate will be dominant. As will be discussed later, this is important when considering the bioavailability and possible volatility of soil iodine.

17.3.4 Volatilization of Soil Iodine

Volatilization of iodine from soils as iodine gas has long been suggested to be an important process in the iodine cycle. Not all authors agree on the degree of volatilization of iodine from soils, but it is generally suggested that its volatilization is significant. It seems likely from consideration of the Eh-pH diagram for iodine (Fig. 17.3), that in oxidizing conditions iodine gas is quite likely formed from iodide in acid soils. Perel'man (1977) has also suggested that the Fe^{3+} and Mn^{4+} ions could oxidize iodide under both acid and alkaline conditions.

In addition to the possible volatilization of gaseous elemental iodine from soils, several workers have demonstrated that soil iodine can be volatilized as methyl iodide, particularly under waterlogged and reducing conditions. Further, it has been suggested that volatilization of methyl iodide from the waterlogged soils may be a contributory cause of the low iodine contents of lowland soils in Japan (Muramatsu and Yoshida 1999).

Therefore, it seems that volatilization of iodine from soils is quite likely and as such could play a very important role in the iodine cycle and the transfer of iodine into the biosphere. Fuge (1996) has suggested that only a relatively small proportion of iodine derived from the marine environment is transported into central continental regions and regions generally remote from the sea. Some of the iodine that occurs in environments far removed from marine influence could have been volatilized from soils, with iodine deposited on land by

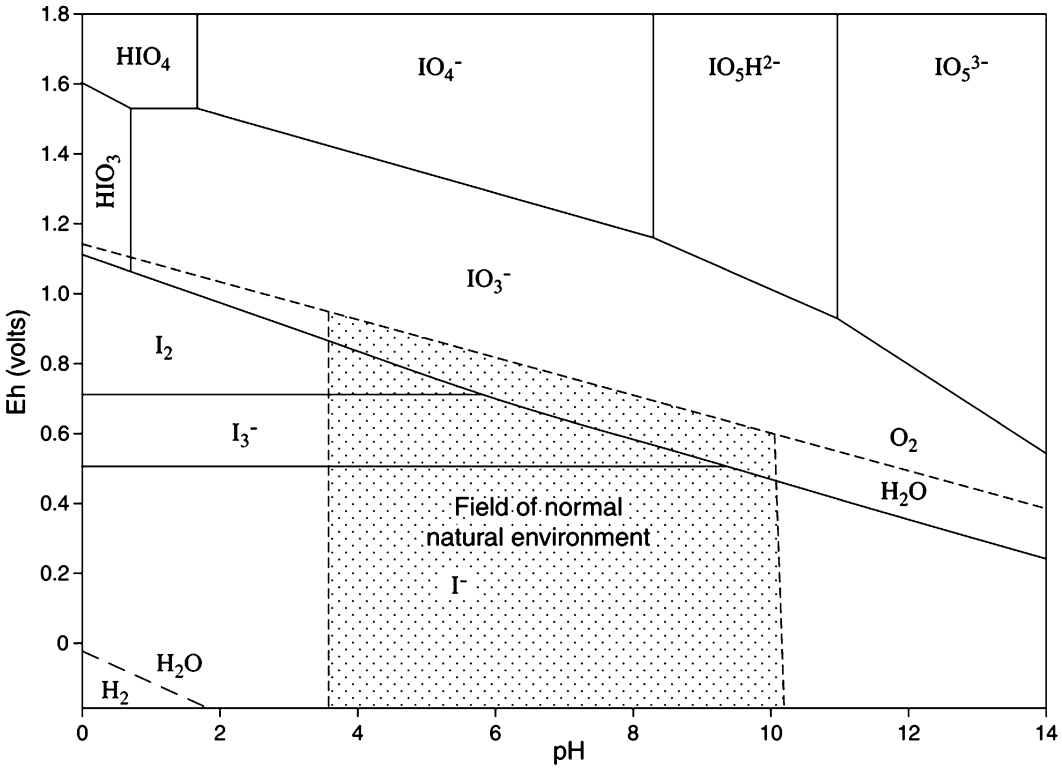


Fig. 17.3 Eh-pH diagram for iodine (Modified from Vinogradov and Lapp 1971 and Bowen 1979)

wet and dry precipitation being subsequently re-volatilized enriching the atmosphere in iodine. Such precipitation and re-volatilization could occur several times resulting in iodine migrating "stepwise" inland. In soils where iodine is strongly bound, such as organic-rich soils, it is likely that iodine is not available for re-volatilization. This could have a detrimental effect on the local iodine cycle by depriving plants of a potential source of iodine (see Sect. 17.4). It is also important to note that volatilization is going to be dependent on both Eh and pH, with volatilization under alkaline oxidizing conditions very unlikely, which is a possible explanation for the elevated iodine contents of the circum-neutral to alkaline soils occurring over limestone (Table 17.3).

17.4 Transfer of Iodine from Soil to Plants

When iodine is strongly sorbed in soils it will not be readily bioavailable. Therefore, the presence of high iodine concentrations in soil does not necessarily mean that plants growing in the soil will incorporate large concentrations of iodine; indeed, it has been shown that there is no correlation between the iodine content of soils and the plants growing on them (Al-Ajely 1985). This is particularly important when considering the distribution of IDD. An additional consideration is that iodine in high concentrations has been shown to be toxic to most plants (Sheppard and Evenden 1995), with high iodine uptake in rice plants thought to be responsible for Akagare disease (Yuita 1994b).

In most circumstances the major pathway of elements into plants is through the root system followed by translocation to the upper parts of the plant. For iodine it has been shown experimentally that it can be taken in through the root system of plants with the iodide ion more readily incorporated than iodate. However, it has also been demonstrated that there is little translocation from the roots to the aerial parts of the plant. Studies of rice plants grown on high-iodine soils in Japan have shown that concentrations of iodine in the roots were over three times greater than the concentrations in the above ground portion of the plants (Tsukada et al. 2008). In some circumstances high concentrations of iodine have been shown to occur in rice grown on flooded soil, which leads to Akagare disease. This has been suggested to be due to high soluble iodide in soils, which results in greater uptake by the roots; however, it has been demonstrated that there is only a relatively small increase in the iodine content of rice grown in flooded soil when compared with rice grown on drained soil, but submerged leaves when compared to other leaves showed dramatically increased iodine content.

From these considerations it seems likely that root uptake of iodine is relatively unimportant for the overall iodine content of plants. It is probable that the most important pathway into plants is from the atmosphere by direct absorption. Experiments utilizing radioactive isotopes of iodine

have demonstrated that plant leaves can absorb this iodine, and it has been found that the absorption of gaseous iodine by leaves increases with increasing humidity. This is probably due to increased opening of the leaf stomata. Iodine absorbed through the leaf can be translocated through the rest of the plant, albeit slowly.

While the uptake of gaseous iodine has been demonstrated to be significant, doubt has been expressed that significant amounts of methyl iodide could be taken in through the stomata. However, even if little is absorbed, atmospheric iodine, whatever the form, could be deposited on plant surfaces and as such could represent a significant source of iodine to grazing animals, etc. The source of the atmospheric iodine in near-coastal areas will be mainly marine. However, the source in inland areas could be derived in part from iodine volatilized from soil, in some circumstances this being very significant.

The uptake of iodine into plants is, therefore, accomplished in two ways: through the roots and through leaf stomata. The latter is probably the most important. However, it must be stressed that the bioavailability of iodine in soils is low, whatever the preferred uptake route. Thus, strongly bound iodine will not be bioavailable for either root uptake or for volatilization from the soils. Several workers have quoted soil-to-plant concentration factors for iodine (iodine in plant/iodine in soil); these are generally very low and are in the range of 0.01–1.5 with most falling between 0.01 and 0.1 (Ng 1982), while IAEA (1994) quoted a value of 0.0034 for grass. Variation of the plant concentration factor for different soil types has been demonstrated, and Muramatsu et al. (1993) quoted factors for brown rice grown on an andosol and gray lowland soil in Japan of 0.007 and 0.002, respectively.

Not surprisingly then, the iodine content of plants is generally low. Grass and herbage analyses from many different countries have shown iodine contents to be about 0.2 mg kg^{-1}; a typical example being for the UK with an estimated mean of 0.22 $\pm$ 0.16 mg kg^{-1} (Whitehead 1984). In a study of Japanese plants Yuita (1994a) found the mean iodine content of different plant parts to be green leaves 0.46 mg kg^{-1}, fruit 0.14, edible roots 0.055, and seeds 0.0039.

17.5 Iodine Sources for Humans

The optimum daily dietary intake of iodine is variously quoted but the WHO recommend 90 μg/day for children under the age of 5, 120 μg/day for children aged 6–12, and 150 μg/day for adults and adolescents, while greater amounts are required during pregnancy and lactation with the WHO recommending an intake of 250 μg/day at such times (see Chap. 8, this volume). The traditional view is that humans derive their iodine from consumption of crops and vegetables, etc. In near-coastal areas such a source may provide sufficient iodine. However, in inland areas this will provide only relatively low quantities of iodine. In this context it has been demonstrated that vegetarian diets result in low iodine intake, which could lead to iodine deficiency (Davidsson 1999; Remer et al. 1999) (see also Chap. 8, this volume). Krajcovicová-Kudlácková et al. (2003) in a study of vegetarians and vegans found that 25% of the vegetarians were iodine deficient compared to 80% of the vegans.

Seafood is a potentially rich source of dietary iodine and, where such food is a major part of the diet as in Japan and Iceland, some problems of excess iodine in the diet have been described. A high iodine intake causes a decrease of thyroid hormone production resulting in formation of "high iodine goiter." However, the major source of dietary iodine in many developed countries is dairy produce such as milk, yogurt, butter, and cheese, which are rich sources due to the addition of iodine to cattle feed and the use of iodine-containing disinfectants, iodophors, in the dairy industry. Recently, the use of iodophors in the dairy industry has been reduced, resulting in lower iodine contents of the produce. It has been suggested that the decreasing iodine content of milk is a possible reason for the re-emergence of iodine deficiency in Australia (Li et al. 2006). Therefore, it is likely that the recent re-emergence of IDD in some affluent countries is due partly to the consumption of lower quantities of dairy produce in the move to "healthy eating," and to the reduced amount of iodine in these produce.

It is also likely that even without iodine added to animal diets, animal products are probably enriched in iodine as grazing animals will take in iodine that has been deposited on the surfaces of grass and leaves; human preparation of food is likely to remove much of this surface-deposited iodine. A study of iodine in cow's milk in Peru demonstrated that cattle grazing on near coastal pasture produced milk which had six times as much iodine as that produced in mountain regions (Cárdenas Quintana et al. 2003), presumably due to greater deposition of marine-derived iodine in the coastal region. In addition animals are known to inadvertently consume soil and this could add iodine that is not bioavailable to plants. This source of iodine has been shown to prevent iodine deficiency in grazing animals in some areas where pasture is iodine deficient as in New Zealand and Tasmania (see Chap. 18, this volume), even though the soil contains only 1–2 mg kg^{-1}. In this context it is also of note that Lidiard (1995) found that farm animals in Exmoor, Somerset, UK, developed iodine deficiency symptoms when grazing areas of reclaimed land. These areas, which are about 40 km from the sea, were originally covered in peat and previous to reclamation no iodine-deficiency problems had occurred. Thus it seems likely that the animals had previously obtained iodine from inadvertent intake of peat.

It is possible that humans may obtain some iodine by inhalation, particularly in near-coastal environments. While

it is indeed possible that some atmospheric iodine might be inhaled even in areas remote from the sea, this is unlikely to be a major component of iodine intake. It is of note that the Nordic Project Group (1995) suggested that an individual acquires only 0.5 μg per day from inhalation.

Although it has been suggested that some dietary iodine derives from drinking water, generally this source is unlikely to provide more than 10% of the daily adult iodine requirement. However, very high iodine drinking waters derived from groundwater have been recorded in some areas of China. In these areas it has been found that goiter incidence is negatively correlated with the iodine content of the drinking water. In Denmark, humic-rich drinking water with 140 μg/L iodine was found to constitute a significant dietary source of iodine (Andersen, et al. 2008). In some areas of China, such as around Bohai Bay in Hebai and Shandong Provinces and in some areas of Xinjiang and Shanxi Provinces, drinking waters contain between 100 and 200 $\mu g L^{-1}$ and the incidence of goiter was found to be positively correlated with the iodine content of the water. The high iodine causes populations in the areas to suffer from high iodine goiter (Tan 1989).

It is perhaps pertinent to point out that in order to address the problem of iodine deficiency, many countries have developed a program of iodization of table salt. This has generally been successful and in many countries this source makes up a significant proportion of daily iodine intake. However, with current medical advice advocating major cuts in salt intake, it is likely that the amount of dietary iodine derived from this source will be significantly lowered.

17.6 The Global Distribution of Iodine Deficiency Disorders and the Iodine Cycle

The most important parts of the iodine cycle, in terms of the environmental distribution of iodine and its impact on human health, are summarized in Fig. 17.4.

Prior to about 1950 IDD had affected virtually every country in the world. Because the problem is known to be essentially due to the lack of iodine in the diet, it has been possible to introduce schemes for the mass treatment of affected populations such as the addition of iodine to the diet through the use of iodized salt and bread, etc., injections of iodized oil, or the addition of iodine to irrigation waters (see Chap. 9, this volume), which results in the alleviation of the symptoms in many countries. It was noted in the 1970s that IDD had been effectively eradicated in the developed world and endemic goiter was described as a disease of the poor that was largely confined to third world countries. However, it is important to point out that since the 1990s iodine deficiency has been reported to occur in several affluent countries in western Europe, in Australia, New Zealand and USA, probably as a result of dietary changes and the reduced use of iodophors in the dairy industry. This suggests that such changes could cause the reintroduction of IDD in many countries.

The global distribution of iodine-deficiency problems is shown in Fig. 17.5. This figure is based on the data in Dunn and van der Haar (1990) and is limited as data are lacking for some areas such as parts of Africa and the Middle East.

As outlined by Kelly and Sneddon (1960) many of the areas affected by iodine deficiency are remote from marine influence. Many of the areas highlighted are the mountainous regions and their rain shadow areas such as the Himalayan region, the European Alps region, and the Andean Chain. Tan (1989) has indicated that many of the extremely seriously affected IDD areas of China are the mountainous and hilly regions over the whole country. Central continental regions, such as those of Africa and China, are also well-documented areas of iodine deficiency. In all of these situations IDD can be explained according to the classic explanations of low iodine supply and hence low iodine availability. Thus in continental areas such as the central United States, where iodine-deficiency problems were described prior to the 1950s (Kelly and Sneddon 1960), soil iodine is relatively low with concentrations typically 1.3 mg kg^{-1} (Table 17.2).

However, many endemias are not explicable in these simplistic terms and several countries and regions that are close to the coast have been known to suffer from IDD problems. For example, large regions of the UK have histories of IDD, despite the strong maritime influence on the country (see Fig. 17.6). IDD in Sri Lanka occurs in coastal areas (Dissanayake and Chandrajith 1996) while Tan (1989) has documented coastal areas of China where endemic goiter have been recorded.

In some endemias the involvement of the sulfur-containing goitrogens has been invoked as a reason. Some of these may derive from geological sources and be incorporated into drinking water or food (Gaitan et al. 1993), but others occur naturally in vegetables such as those of the *Brassica* genus and in such staple items as cassava (see also Chap. 9, this volume).

Several other elements such as fluorine, arsenic, zinc, magnesium, manganese and cobalt have been suggested to be involved in the etiology of IDD, but there is no real evidence to support their involvement. However, more recently in some areas of IDD selenium has been shown to be strongly implicated (Vanderpas et al. 1990) (see also Chap. 16, this volume).

However, even allowing for the involvement of other elements and compounds in some endemias, it seems likely that the causes of IDD problems in several areas are

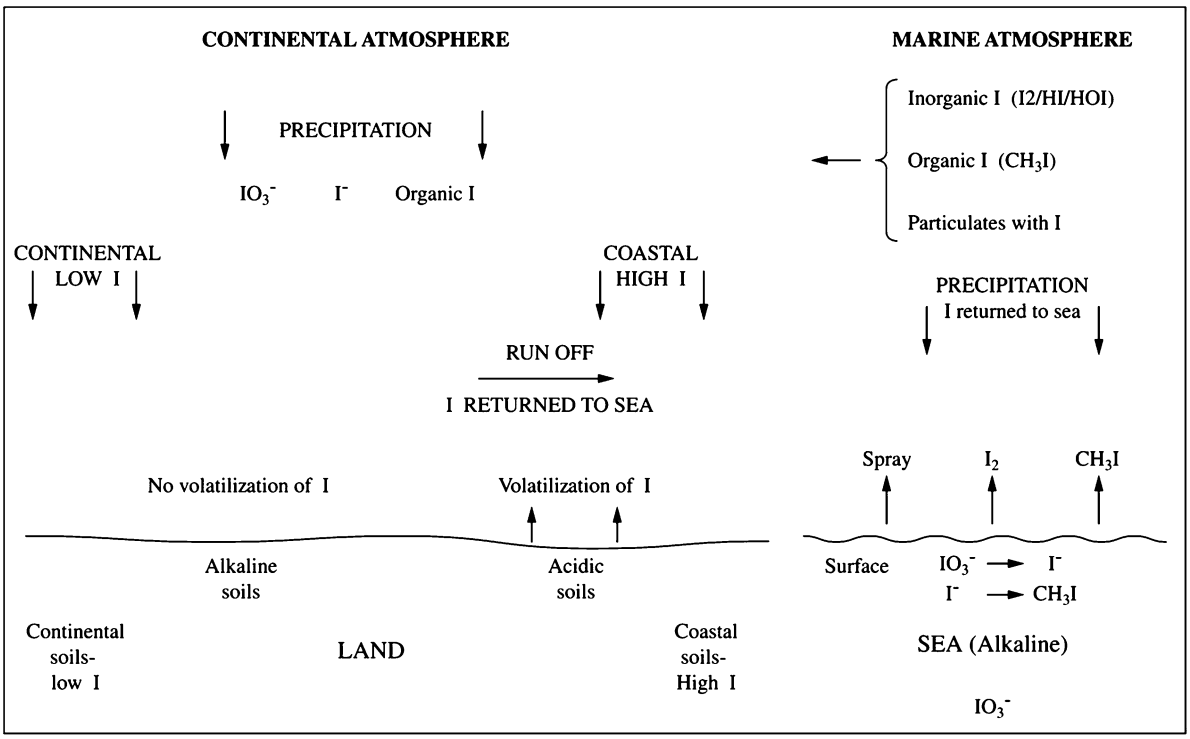


Fig. 17.4 A simplified model of part of the iodine cycle

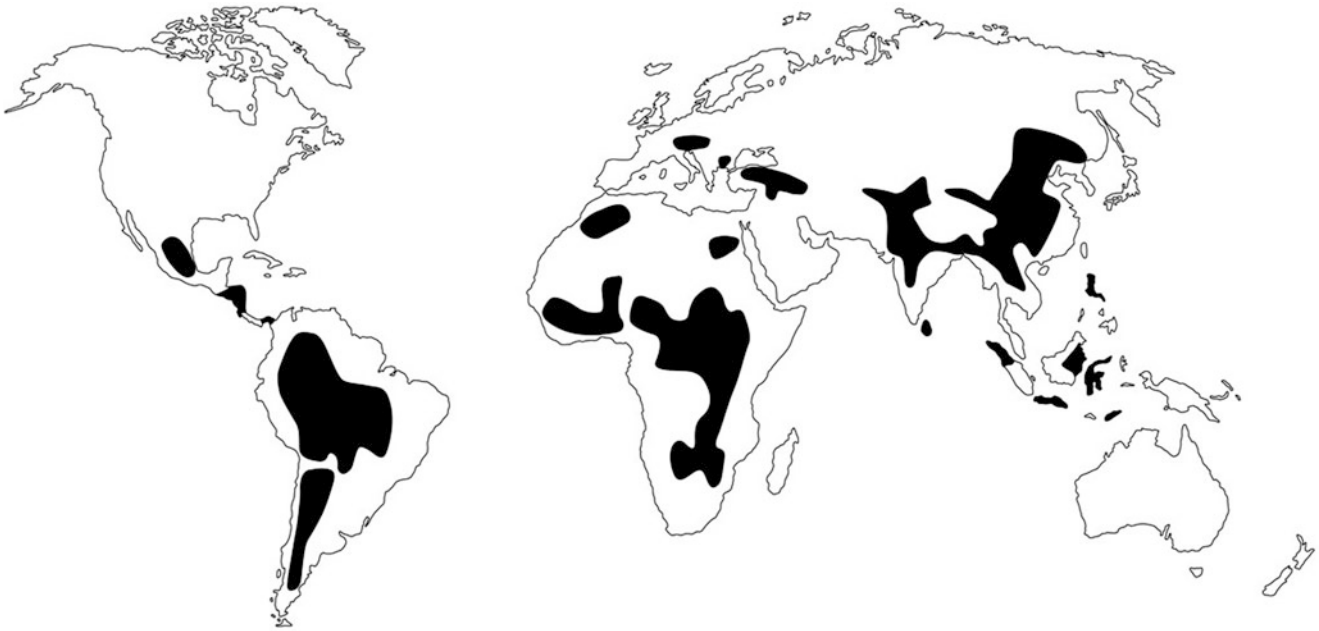


Fig. 17.5 Global distribution of IDD (Modified from Dunn and van der Haar 1990)

governed by the geochemistry of iodine and its bioavailability. In some cases sandy soils have been found to occur in goitrous areas such as the coastal regions of China. Here any iodine added to the soil from atmospheric sources is going to be leached from the soil rapidly, which results in low-iodine soils. This means little iodine is available for any plants or crops growing in the region, thus depriving humans and livestock of a source of iodine.

In sandy soils iodine is not retained, as outlined in Sect. 17.3 of this chapter. The iodine retention capacity of

Fig. 17.6 Distribution of IDD in the UK and Ireland (Modified from Kelly and Snedden 1960)

soils is related to composition with organic matter, and iron and aluminum oxides being the most important retentive components. So that in general, high-iodine soils are rich in one or more of these components. The iodine sorbed by these soil components is strongly held and little has been found to be easily leachable, so that in general terms this strongly held soil iodine is non-bioavailable. Thus Tan (1989) noted that some peaty areas of China were goitrous. Similarly, several peaty areas of the UK, even though some are in coastal regions, are listed as goitrous by Kelly and Sneddon (1960).

Another example of environmental control on IDD would seem to be the strong association of some goitrous regions with limestone bedrock as first mentioned by Boussingault (1831) for a region of Columbia, and subsequently by several other researchers such as Perel'man (1977). This association of limestone with goitrous areas led to the suggestion that calcium was a goitrogen, a claim that was subsequently disproved in clinical tests.

There are many examples of such limestone-associated endemias occurring in the UK and Ireland with the limestone regions of north Yorkshire historically renowned for its severe goiter and cretinism problems (see Fig. 17.6). Similarly County Tipperary in Ireland, underlain by Carboniferous Limestone, was historically one of the major areas of IDD in that country. One of the most well-documented areas of IDD in the UK is the Derbyshire region of northern England, an area that is between 150 and 180 km from the west coast in the direction of the prevailing wind. Here endemic goiter was rife, known as "Derbyshire neck," with the endemia confined to areas underlain by limestone bedrock (see Fig. 17.6). Analyses of soil in the former goitrous region have shown that iodine concentrations range up to 26 mg kg^{-1} with a mean value of 8.2 mg kg^{-1}. In neighboring areas underlain by non-limestone lithologies and with no history of IDD, soil iodine concentrations are lower with a mean of 3.44 mg kg^{-1}. Similarly, there is an area of north Oxfordshire, England, where endemic goiter was prevalent (Kelly and Sneddon 1960) and where iodine-deficiency problems in school children were recorded as recently as the 1950s. This area is underlain by limestone and soil iodine ranges between 5 and 10 mg kg^{-1}.

IDD occurs in these limestone areas despite relatively high iodine in soils. This would imply that iodine is not

bioavailable. Soils over the limestones would generally be well drained and circum-neutral to alkaline in nature. In these conditions any soluble soil iodine is likely to be present as the iodate anion (Fig. 17.3). It has been shown that iodate uptake through plant roots is more limited than the uptake of iodide, thus in neutral to alkaline soils root uptake of iodine may be low. It is also apparent that iodate is far more strongly adsorbed in soils than iodide (Hu et al. 2012).

In addition, as stated earlier, it is likely that plants derive much of their iodine from the atmosphere, through the upper parts of the plant. In the neutral to alkaline soils generally found overlying limestones where iodate is going to be the dominant soluble species of iodine, there is no possibility of conversion of this species to gaseous elemental iodine (see Fig. 17.3). An additional factor is that in the well-drained soils there is going to be no conversion of iodine to methyl iodide. Thus in the limestone areas, plants will be deprived of a local source of atmospheric iodine.

The distribution of IDD reflects the geochemistry of iodine and, as stated above, with large areas of iodine deficiency occurring in central continental regions and mountainous and rain shadow areas, this distribution fits in with the classical explanation of IDD governed by the external supply of iodine from the marine environment via the atmosphere. However, the geochemistry of iodine is more complex than this simplistic approach, and from a closer scrutiny of the distribution of IDD it is apparent that, in many cases, iodine deficiency problems are related to the bioavailability of iodine in soils and are not related directly to the external supply of iodine.

17.7 Radioactive Iodine in the Environment

A relatively recent problem regarding iodine and human health is that of anthropogenically produced radioactive iodine, a topic that has received considerable attention during the last three decades. Although the problem of radioactive iodine that is of anthropogenic origin is perhaps out of place in this chapter, it is included because the distribution of natural iodine is likely to have a marked influence on the health effects of the radioactive iodine.

Whereas natural iodine is essentially mono isotopic, the one stable isotope being ^{127}I, over 20 radioactive isotopes have been identified ranging from ^{117}I to ^{139}I. Of these radioactive isotopes only ^{129}I has a significantly long half-life of 1.6×10^7 years. Extremely small amounts of ^{129}I are produced naturally by spontaneous fission of uranium and also by spallation of xenon in the upper atmosphere. However, comparatively large quantities of ^{129}I are produced from nuclear fission fallout and reactors. The pre-nuclear age ^{129}I/^{127}I ratio has been estimated as 10^{-12} while present day "background values" have been estimated to be about 10^{-10}.

Relatively high concentrations of ^{129}I have been found in the biosphere around nuclear plants, and elevated concentrations of this isotope have been shown to occur in the thyroid glands of animals in the vicinity of nuclear installations. The release of fairly large quantities of ^{129}I from nuclear sources is of concern, but it is the allied release of the shorter lived isotopes ^{125}I (half-life 60 days), ^{131}I (half-life 8.04 days), and ^{133}I (half-life 20.9 h) that represents the greater threat to human health. Of these it is ^{131}I that has been suggested to pose the greatest risk because it is produced in fairly large quantities from fission of enriched uranium (the yield is about 3%). It has a particularly high specific activity, and hence is highly radioactive. It has been estimated that in the Chernobyl accident, in April 1986, 1.3 million terabecquerels (35 million curies) of ^{131}I was released. More recently, within 4 weeks of the accident at the Fukushima Daiichi nuclear plant, Japan, resulting from the earthquake and tsunami in March 2011, trace levels of ^{131}I were found in rainwater and milk in Washington State and elsewhere in western USA. This suggests that radioactive iodine from accidental releases can be transported great distances fairly rapidly.

The problem of the short-lived radioactive isotopes of iodine was first encountered in the early years of the nuclear industry after an accident at Windscale, UK, in 1957. Radioactive iodine was found to enter the biosphere rapidly and was found in plants and in cow's milk shortly after the accident. Presumably as the radioactive iodine would have been released as a gas, it could be absorbed through plant stomata and deposited on foliage. Thus the contaminant iodine passes fairly rapidly through the food chain and into humans where it can be taken into the thyroid gland. Once there, radioactive breakdown would result in an increased risk of thyroid cancer.

In areas where natural iodine is deficient, the problem of large releases of ^{131}I would be potentially very serious, as a sudden increase of iodine would result in a high percentage of bioavailable iodine that is radioactive. Thus much of the iodine entering the human body would be radioactive. The area around Chernobyl is situated in a central continental region and as such has a history of iodine deficiency (Kelly and Sneddon 1960) with goiter found to occur in schoolchildren in the area. Since the Chernobyl accident there has been an increased incidence of childhood thyroid cancers, and this has been shown to be causally linked to the release of radioactive iodine.

17.8 Summary

The lithosphere is generally depleted in iodine and although it contributes to iodine in soils through weathering of bedrock, this is not the most important part of the iodine

geochemical cycle. The oceans represent the largest reservoir of iodine on the Earth, and virtually all iodine in the terrestrial environment derives from the oceans by way of the atmosphere. Iodine is volatilized from the sea as methyl iodide (CH_3I), elemental iodine (I_2), and possibly as some other inorganic iodine compounds such as HI or HOI, with CH_3I probably the most important of these. This volatilized iodine is deposited on land by wet and dry precipitation and consequently soils from near-coastal environments are enriched in iodine and soils remote from the sea are depleted.

The iodine content of soils reflects not only the amount of iodine input from the atmosphere but is markedly dependent on soil composition. Thus organic matter, and iron and aluminum oxides in soils are able to strongly sorb iodine, and soils enriched in these components are frequently enriched in iodine.

It has generally been assumed that iodine in soils is transferred to plants and these, in turn, represent a major pathway of iodine into animals and humans. However, the soil-to-plant concentration factor for iodine has been shown to be low due to the strong sorption of iodine by soil components such as organic matter and iron and aluminum oxides. Little iodine in soils has been found to be easily leachable and no correlation of soil and plant iodine has been demonstrated. In addition, it has been shown that although iodine is taken into plant roots with the iodide ion more readily incorporated than iodate, little of this iodine is translocated from the roots of plants to the aerial parts. It is likely that most iodine in plants is taken in from the atmosphere through the stomata. Whereas in coastal areas such an atmospheric source of iodine is likely to be abundant, in inland areas such an atmospheric source would, to a large extent, be dependent on iodine volatilized from soils. Volatilization of soil iodine is likely to be of major importance in the iodine cycle, but in many areas such volatilization may be limited due to iodine being strongly bound in soil. This is likely to be particularly important in areas underlain by limestone where soils would be expected to be circumneutral to alkaline, which results in any labile iodine being present as the iodate ion and hence unable to be converted to gaseous elemental iodine.

Traditionally, crops and vegetables have been suggested to be important sources of dietary iodine for humans. However, this is unlikely to be true in all but coastal regions as little iodine in soils is generally bioavailable. Seafood is generally a rich source of dietary iodine, whereas in some areas drinking waters are important sources. In developed countries, dairy products are a major source of dietary iodine due to the addition of iodine to cattle feed and use of iodine-containing sterilants in the dairy industry. However, even without the addition of iodine to dairy products animal products are likely to be richer sources of dietary iodine as grazing animals will take in iodine that has been deposited on the surfaces of grass and leaves. In addition, many grazing animals are known to inadvertently take in soil, which has been shown to provide more iodine. The recent upsurge in iodine deficiency in industrialized countries is suggested to be due to dietary changes and the reduced use of iodophors in the dairy industry.

The global distribution of IDD reflects the geochemistry of iodine with large areas of iodine deficiency occurring in central continental regions and mountainous and rain shadow areas, which reflect the supply of iodine from the marine environment via the atmosphere. However, the geochemistry of iodine is more complex than this simplistic approach, and from a closer scrutiny of the distribution of IDD it is apparent that, in many cases, iodine-deficiency problems are related to the bioavailability of iodine in soils and are not related directly to the external supply of iodine.

A relatively modern problem concerning iodine is the release of radioactive iodine from anthropogenic sources. Of the radioactive isotopes of iodine it is ^{131}I, with a half-life of 8.04 days, which has been suggested to pose the greatest risk as it is produced in fairly large quantities from fission of enriched uranium and is highly radioactive. In areas where natural iodine is deficient the problem of large releases of ^{131}I would be potentially very serious, as a sudden increase of iodine would result in a high percentage of bioavailable iodine being radioactive. This problem was highlighted in the Chernobyl accident which occurred in a central continental area where iodine-deficiency problems have been described. A high incidence of childhood thyroid cancer since the accident has been shown to be causally linked to the release of the radioactive iodine.

See Also the Following Chapters. Chapter 7 (Biological Functions of the Elements) • Chapter 8 (Geological Impacts on Nutrition) • Chapter 9 (Biological Responses of Elements) • Chapter 15 (Bioavailability of Elements in Soil) • Chapter 16 (Selenium Deficiency and Toxicity in the Environment) • Chapter 18 (Geophagy and the Involuntary Ingestion of Soil)

Further Reading

Al-Ajely KO (1985) Biogeochemical prospecting as an effective tool in the search for mineral deposits. Ph.D. thesis, University of Wales, Aberystwyth

Amachi S, Fujii T, Shinoyama H, Muramatsu Y (2005) Microbial influences on the mobility and transformation of radioactive iodine in the environment. J Nucl Radiochem Sci 6:21–24

Andersen S, Pedersen KM, Iversen F, Terpling S, Gustenhoff P, Petersen SB, Laurberg P (2008) Naturally occurring iodine in humic substances in drinking water in Denmark is bioavailable and determines population iodine intake. Br J Nutr 99:319–325

Boussingault JB (1831) Recherches sur la cause qui produit le Goitre dans les Cordileres de la Novelle-Grenade. Ann Chim (Phys) 48:41–69

Bowen HJM (1979) Environmental chemistry of the elements. Academic, London

Cárdenas Quintana H, Gómez Bravo C, Pretell EA (2003) Iodine content of cattle milk from mountain and coast from Perú. Arch Latinoam Nutr 53:408–412 (in Spanish)

Chameides WC, Davis DD (1980) Iodine: its possible role in tropospheric photochemistry. J Geophys Res 85:7383–7398

Councell TB, Landa ER, Lovley DR (1997) Microbial reduction of iodate. Water Air Soil Pollut 100:99–106

Davidsson L (1999) Are vegetarians an "at risk group" to iodine deficiency? Br J Nutr 81:3–4

Dean GA (1963) The iodine content of some New Zealand drinking waters with a note on the contribution of sea spray to the iodine in rain. N Z J Sci 6:208–214

Dissanayake CB, Chandrajith RLR (1996) Iodine in the environment and endemic Goitre in Sri Lanka. In: Appleton JD, Fuge R, McCall GJH (eds) Environmental geochemistry and health. Geological Society Special Publication, London 113: 201–211

Duce RA, Zoller WH, Moyers JL (1973) Particulate and gaseous halogens in the Antarctic atmosphere. J Geophys Res 78:7802–7811

Dunn JT, van der Haar F (1990) A practical guide to the correction of iodine deficiency. International Council for the Control of Iodine Deficiency Disorders technical manual no. 3, The Netherlands

Fordyce FM, Stewart AG, Ge X, Jiang J-Y, Cave M (2003) Environmental controls in IDD: a case study in the Xinjiang Province of China, Technical report, CR/01/045N. British Geological Survey, Nottingham

Fuge R (1987) Iodine in the environment: its distribution and relationship to human health. Trace Subst Environ Health 21:74–87

Fuge R (1988) Sources of halogens in the environment, influences on human and animal health. Environ Geochem Health 10:51–61

Fuge R (1996) Geochemistry of iodine in relation to iodine deficiency diseases. In: Appleton JD, Fuge R, McCall GJH (eds) Environmental geochemistry and health. Geological Society Special Publication 113: 201–211

Fuge R, Ander L (1998) Geochemical barriers and the distribution of iodine in the secondary environment: implications for radio-iodine. In: Nicholson K (ed) Energy and the environment: geochemistry of fossil, nuclear and renewable resources. MacGregor Science, Insch, pp 163–170

Fuge R, Johnson CC (1986) The geochemistry of iodine-a review. Environ Geochem Health 8:31–54

Gabler HE, Heumann KG (1993) Determination of iodine species using a system of specifically prepared filters and IDMS. Fresnius J Anal Chem 345:53–59

Gaitan E, Cooksey RC, Legan J, Cruse JM, Lindsay RH, Hill J (1993) Antithyroid and goitrogenic effects of coal-water extracts from iodine-sufficient areas. Thyroid 3:49–53

Garland JA, Curtis H (1981) Emission of iodine from the sea surface in the presence of ozone. J Geophys Res 86:3183–3186

Gerzabek MH, Muramatsu Y, Strebl F, Yoshida S (1999) Iodine and bromine contents of some Austrian soils and relations to soil characteristics. J Plant Nutr Soil Sci 162:415–419

Goldschmidt VM (1954) Geochemistry. Oxford University Press, London

Hetzel BS (1987) The biological effects of iodine deficiency and their public health significance. Neuro Toxicol 8:363–368

Hetzel BS (2001) Iodine deficiency and the brain. Nutr Neurosci 2:375–384

Hetzel BS, Potter BJ, Dullberg EM (1990) The iodine deficiency disorders, nature, pathogenesis and epidemiology. In: Bourne G (ed) World review of nutrition and dietetics, vol 62. Karger, Basel, pp 59–119

Heumann KG, Gall M, Weiss H (1987) Geochemical investigations to explain iodine-overabundances in Antarctic meteorites. Geochim Cosmochim Acta 51:2541–2547

Heumann KG, Neubauer J, Reifenhauser H (1990) Iodine overabundances measured in the surface layers of an Antarctic stony and iron meteorite. Geochim Cosmochim Acta 54:2503–2506

Hou X, Hansen V, Aldahan A, Possnert G, Lind OC, Lujaniene G (2009) A review on speciation of iodine-129 in the environmental and biological samples. Anal Chim Acta 632:181–196

Hu QH, Moran JE, Gan JY (2012) Sorption, degradation, and transport of methyl iodide and other iodine species in geologic media. Appl Geochem 27:774–781

IAEA (1994) Handbook of parameter values for the prediction of radionuclide transfer in temperate environments, Technical report series, no. 364. International Atomic Energy Authority, Vienna

Johnson CC (1980) The geochemistry of iodine and a preliminary investigation into its potential use as a pathfinder element in geochemical exploration. Ph.D. thesis, University of Wales, Aberystwyth

Johnson CC, Strutt MH, Hmeurras M, Mounir (2002) Iodine in the environment of the high Atlas Mountains, Morocco. British Geological Survey Commissioned report CR/02/196N, Nottingham

Kelly FC, Sneddon FW (1960) Prevalence and geographical distribution of endemic goitre. In: Endemic goitre. World Health Organization, Geneva, pp 27–233

Kocher DC (1981) A dynamic model of the global iodine cycle and estimation of dose to the world population from release of I-129 to the environment. Environ Int 5:15–31

Krajcovicová-Kudlácková M, Bucková K, Klimes I, Seboková E (2003) Iodine deficiency in vegetarians and vegans. Ann Nutr Metab 47(5):183–185

Låg J, Steinnes E (1976) Regional distribution of halogens in Norwegian forest soils. Geoderma 16:317–325

Langer P (1960) History of goitre. In: Endemic goitre. World Health Organization, Geneva, pp 9–25

Li M, Waite KV, Ma G, Eastman CJ (2006) Declining iodine content in milk and re-emergence of iodine deficiency in Australia. Med J Aust 184(6):307

Lidiard HM (1995) Iodine in reclaimed upland soil of a farm in the Exmoor National Park, Devon, U.K., and its impact on livestock health. Appl Geochem 10:85–95

Lovelock JE, Maggs RJ, Wade RJ (1973) Halogenated hydrocarbons in and over the Atlantic. Nature 241:194–196

Moison TA, Dunstan WM, Udomkit A, Wong GTF (1994) The uptake of iodate by marine phytoplankton. J Phycol 30:580–587

Moore RM, Groszko W (1999) Methyl iodide distribution in the ocean and fluxes to the atmosphere. J Geophys Res 104:11163–11171

Muramatsu Y, Wedepohl KH (1998) The distribution of iodine in the earth's crust. Chem Geol 147:201–216

Muramatsu Y, Yoshida S (1999) Effects of microorganisms on the fate of iodine in the soil environment. Geomicrobiol J 16:85–93

Muramatsu Y, Uchida S, Ohmomo Y (1993) Root-uptake of radioiodine by rice plants. J Radiat Res 34:214–220

Muramatsu Y, Yoshida S, Uchida S, Hasebe A (1996) Iodine desorption from rice paddy soil. Water Air Soil Pollut 86:359–371

Ng YC (1982) A review of transfer factors for assessing the dose from radionuclides in agricultural products. Nucl Saf 23:57–71

Nordic Project Group (1995) Risk evaluation of essential trace elements–essential versus toxic levels of intake. Nordic Council of Ministers, Nord, Copenhagen, pp 18

Perel'man AJ (1977) Geochemistry of elements in the supergene zone. Keterpress Enterprises, Jerusalem

Prister BS, Grigor'eva TA, Perevezentsev VM, Tikhomirov FA, Sal'nikov VG, Ternovskaya IM, Karabin T (1977) Behaviour of iodine in soils. Pochvovedenie 6:32–40 (in Russian)

Rahn KA, Borys RD, Duce RA (1976) Tropospheric jalogen gases: inorganic and organic components. Science 192:549–550
Remer T, Neubert A, Manz F (1999) Increased risk of iodine deficiency with vegetarian nutrition. Br J Nutr 81:45–49
Schmitz K, Aumann DC (1994) Why are the soil-to-pasture transfer factors, as determined by field-measurements for I-127 lower than for I-129. J Environ Radioact 24:91–100
Shacklette HJ, Boerngen JG (1984) Element concentrations in soils and other surficial materials of the conterminous United States, United States Geological Survey professional paper 1270. Denver, CO
Sheppard SC, Evenden WG (1995) Toxicity of soil iodine to terrestrial biota with implications for I-129. J Environ Radioact 27:99–116
Shimamoto YS, Takahashi Y, Terada Y (2011) Formation of organic iodine supplied as iodide in a soil-water system in Chiba, Japan. Environ Sci Technol 45:2086–2092
Takagi H, Iijima I, Iwashima K (1994) Determination of iodine with chemical forms in rain water with fractional sampling NAA. Bunseki Kagaku 43:905–909
Tan J (ed) (1989) The atlas of endemic diseases and their environments in the People's Republic of China. Science Press, Beijing
Tsukada H, Takeda A, Tagami K, Uchida S (2008) Uptake and distribution of iodine in rice plants. J Environ Qual 37:2243–2247
Vanderpas JB, Contempre B, Duale NL, Gossans W, Bebe NGO (1990) Iodine and selenium deficiency associated with cretinism in Northern Zaire. Am J Clin Nutr 52:1087–1093
Vinogradov AP (1959) The geochemistry of rare and dispersed chemical elements in soils, 2nd edn. Consultants Bureau, New York
Vinogradov AP, Lapp MA (1971) Use of iodine haloes to search for concealed mineralisation. Vestn Leningr Univ Ser Geolog Geogr 24:70–76 (in Russian)
Whitehead DC (1974) The sorption of iodide by soil components. J Sci Food Agric 25:461–470
Whitehead DC (1984) The distribution and transformation of iodine in the environment. Environ Int 10:321–339
Wong GTF (1991) The marine geochemistry of iodine. Rev Aquat Sci 4:45–73
Wong GTF, Cheng XH (1998) Dissolved organic iodine in marine waters: determination, occurrence and analytical implications. Mar Chem 59:271–281
Yamaguchi N, Nakano M, Takamatsu R, Tanida H (2010) Inorganic iodine incorporation into soil organic matter: evidence from iodine K-edge structure. J Environ Radioact 101:451–457
Yoshida S, Muramatsu Y (1995) Determination of organic, inorganic, and particulate iodine in the coastal atmosphere of Japan. J Radioanal Nucl Chem Artic 196:295–302
Yuita K (1994a) Overview and dynamics of iodine and bromine in the environment. 1. Dynamics of iodine and bromine in soil-plant system. Jpn Agric Res Q 28:90–99
Yuita K (1994b) Overview and dynamics of iodine and bromine in the environment. 2. Iodine and bromine toxicity and environmental hazards. Jpn Agric Res Q 28:100–111
Yuita K, Tanaka T, Abe C, Aso S (1991) Dynamics of iodine, bromine and chlorine in soil. 1. Effects of moisture, temperature and pH on the dissolution of the triad from soil. Soil Sci Plant Nutr 37:61–73
Zafiriou OC (1974) Photochemistry of halogens in the marine atmosphere. J Geophys Res 79:2730–2732
Zimmermann MB (2010) Symposium on 'geographical and geological influences on nutrition' iodine deficiency in industrialised countries. Proc Nutr Soc 69:133–143

Geophagy and the Involuntary Ingestion of Soil 18

Peter W. Abrahams

Contents

Geophagy or geophagia can be defined as the deliberate ingestion of soil. This is a practice that is common among members of the animal kingdom, including people. Any person who studies geophagy undertaken by humans will invariably confront a problem during their research. Few people will believe them. This is perhaps understandable for members of a developed urban society that is educated, has ready access to modern pharmaceuticals, and which has increasingly, in both a physical and mental sense, become more remote from soils. Yet many of these people will readily accept that wild animals deliberately eat soil. For example, television programs which feature wildlife may show animals consuming soil, with the presenter commonly stating that the soils are being eaten for their mineral nutrient content (although, as will be seen in the following sections, there are a variety of reasons why animals consume soil). But many people find it more difficult to accept that humans can deliberately eat soil. This ignorance of geophagy is not restricted to the layperson, because academic writers have used adjectives such as curious, odd, perverted, and strange when commenting on human geophagy. The use of such words demonstrates a misunderstanding of geophagy. The practice is common in certain human societies and can be readily found in many countries provided one has appropriate knowledge of the subject. An understanding of geophagy also allows an appreciation of the practice. There are perfectly sensible reasons as to why certain people deliberately eat soil, and the consumer can benefit from indulging in geophagy in a number of ways. It has been suggested that the practice should be considered within the normal range of human behavior (Vermeer 1986), an enlightened viewpoint that I personally support. However, before the reader hurries away to indulge in geophagy, a word of warning is necessary. Aside from the benefits that eaten soils can impart to the consumer, very serious health problems may also result. These benefits and banes of soil consumption are considered in more detail later in Sect. 18.6 of this chapter, along with other aspects of geophagy undertaken by humans. This information follows a discussion about geophagy that is

P.W. Abrahams (✉)
Institute of Geography and Earth Sciences, Aberystwyth University, Aberystwyth, UK
e-mail: pwa@aber.ac.uk

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_18,

practiced by members of the animal kingdom other than humans. In addition to geophagy, many animals (including humans) also accidentally ingest soil. In order to appreciate geophagy in its proper context, this involuntary ingestion of soil is considered first (see also Chap. 15, this volume).

18.1 Involuntary Soil Ingestion: Domesticated Animals

Grazing and browsing animals are especially prone to what is variously called accidental, involuntary, or incidental soil ingestion. Although both wild and domesticated animals can ingest soil involuntarily, the majority of research on this topic has concentrated on the latter largely because of the economic and health implications to humans. On farmland, pasture plants growing close to the surface are subject to soil contamination resulting from the effects of trampling by grazing animals, rain splash, or wind. Grazing animals will ingest soil that has adhered to vegetation because of these processes. Soil can also be licked from snouts and, on closely cropped pastures, can be ingested directly from the ground surface. The soil adhering to roots can be another source. For example, in a study undertaken on a semi-arid range located in Idaho in the United States, soil was ingested by cattle primarily with the roots of cheatgrass (*Bromus tectorum*) that were pulled up and consumed along with the aboveground plant parts (Mayland et al. 1975).

The amount of soil ingested can be quantified in a variety of ways (Healy 1973). The ash content of feces from animals gives a measure of soil content, with a correction necessary for the ash contribution from undigested herbage. Treatment of ash with dilute acid gives an acid-insoluble residue (AIR) that allows a more accurate measure of soil content to be calculated. If the fecal output is known together with soil content of the feces, then the quantity of soil ingested can be determined. Alternatively, the titanium content of soil and feces can be used to estimate rates of soil ingestion (Miller et al. 1976). This method is based on the premise that titanium, which is abundant in soils (containing typically several thousand mg kg^{-1}), is present only in small quantities (usually <10 mg kg^{-1}) in plants not contaminated with soil. Any titanium recorded in fecal samples can thus be assumed to originate from a soil source. With animals absorbing a negligible amount of ingested soil titanium, soil ingestion can be calculated using the equation:

$$\% \text{ soil ingestion} = \frac{(1 - D_h)Ti_f \times 100}{Ti_s - D_h Ti_f} \qquad (18.1)$$

where D_h = digestibility of herbage, Ti_s = titanium in soil, and Ti_f = titanium in feces.

Research in New Zealand has indicated that sheep can ingest >75 kg of soil annually, and dairy cows can consume between 150 and 650 kg of soil $year^{-1}$ (Healy 1968). In New Zealand and elsewhere, seasonal variations of soil ingestion are marked, depending on soil type, weather, and management factors such as stocking rate and the use of supplementary feed. Within individual flocks and herds, significant differences in soil ingestion can be found between animals at any point in time, though there is some evidence to suggest that identical twins of cows have an inherited tendency to consume similar amounts of soil. In the UK, dairy cattle can ingest at certain times of the year >10% of their dry matter (DM) intake in the form of soil. For sheep, grazing typically closer to the ground, the figure may exceed 30% (Thornton 1974). In countries such as New Zealand and the UK, the rates of soil ingestion are low during the summer months when there is an adequate supply of herbage. Soil ingestion is greater in the autumn, winter, and early spring months attributable to factors such as the low rates of herbage production.

There are a number of economic and health implications that are associated with the involuntary ingestion of soil. For example, animal production will be adversely affected if the consumption of soil reduces the digestible DM intake (Pownall et al. 1980). Also, as soil is highly abrasive to dentine, excessive tooth wear attributable to a high ingestion of soil can lead to culling at a comparatively early age because the animal can no longer graze efficiently (Healy and Ludwig 1965). The abrasive effects on the alimentary tract could also prove irritating to animals, and may additionally increase their vulnerability to infections. The majority of research, however, has investigated the implications of ingested soils as a source of potentially beneficial mineral nutrients or of undesirable constituents such as pesticide residues, heavy metals, and radionuclides (Harrison et al. 1970; Beresford and Howard 1991; Green et al. 1996; Lee et al. 1996; Abrahams and Steigmajer 2003). As soil passes through the gastrointestinal tract of animals, en route it is exposed to digestive fluids that have the potential to extract soil elements thus contributing to the concentrations in these solutions. The release of elements such as cobalt, iodine, and selenium into the digestive fluids may be of benefit to the animal, because it is from the pool of elements in solution that essential mineral nutrients are absorbed into the bloodstreams of animals for distribution throughout their tissues. For example, in both New Zealand and Australia research has suggested that ingested soils can supplement iodine in potentially goitrous grazing sheep (Healy et al. 1972; Statham and Bray 1975). Soils are a significant source of cobalt to grazing animals because they contain 100–1,000 times more of this mineral nutrient than the pasture herbage they support. Cobalt is an essential constituent of vitamin B_{12}, the anti-pernicious anemia factor of liver that is

Fig. 18.1 Sheep grazing mineralized ground disturbed by historical mining activity, Derbyshire, UK. At this locality the ingestion of soils containing high concentrations of fluorine and lead can contribute to health problems suffered by young livestock (Photo: Peter W. Abrahams)

Table 18.1 Maximum and minimum soil and washed herbage arsenic (As) and copper (Cu) concentrations, rates of soil ingestion, and calculated intake by cattle of the two elements

	Soil ingested (%)	Soil concentration (mg kg^{-1})		Washed herbage concentration (mg kg^{-1} DM)		Daily intake as soil (mg day^{-1})		Daily intake as herbage (mg day^{-1})		Total daily intake (mg day^{-1})		% element ingested as soil	
		As	*Cu*	*As*	*Cu*	*As*	*Cu*	*As*	*Cu*	*As*	*Cu*	*As*	*Cu*
April	1.5–17.9			0.06–1.1	10–23	9–189	6–154	0.8–15	108–294	10–196	250–396	80–97	3.7–59
June	0.2–3.9	19–320	12–319	0.03–0.8	8–15	2–47	2–62	0.4–10	104–210	2.5–57	113–236	41–93	2.1–34
August	1.4–4.7			0.10–1.0	9–15	8–101	6–79	1.1–13	123–194	10–113	113–273	79–96	3.2–36

Data are taken from a study undertaken in the soil-contaminated province of southwest England
Reprinted from Agriculture, Ecosystems and Environment, 48, Peter W. Abrahams and Iain Thornton, The contamination of agricultural land in the metalliferous province of southwest England: implications to livestock, 125–137, Copyright 1994, with permission from Elsevier Science

produced by the synthesizing abilities of the symbiotic gastrointestinal bacteria of an animal. Research undertaken in the UK has shown that not only is cobalt extracted from soil in the rumen of sheep, but it is also synthesized into vitamin B_{12} as required by the animal (Brebner et al. 1987). Cobalt deficiency in sheep can thus be prevented by farmers dosing animals with soil (see also Chap. 21, this volume).

These examples demonstrate the importance of the direct soil–animal pathway of mineral nutrients that complements the soil-plant-animal route in agricultural systems. However, ingested soil can also be an important source of potentially harmful elements (PHEs) in geochemically anomalous areas such as mineralized and mined districts (Fig. 18.1). This was found in a study undertaken in southwest England, a province that is extensively contaminated because of mineralization and a long history of mining and smelting. Here concentrations of soil arsenic are high, typically several hundred mg kg^{-1}, yet the aboveground pasture herbage when free of soil contains only about 1 mg kg^{-1} of this element (Abrahams and Thornton 1994). Consequently, the soil–animal pathway is the dominant source of arsenic to grazing cattle in the province, with ingested soil contributing up to 97% of the total intake of this element (Table 18.1). In contrast, the ingestion of soil is not so important in supplying copper to grazing cattle in southwest England. The mineralization and mining activity within this province has contaminated large areas of soil with this metal, but relative to arsenic, more copper is absorbed and transferred to the aboveground parts of the pasture herbage species. This reduces the importance of soil ingestion in supplying copper to the animals, although sites of particularly high rates of soil ingestion (e.g., 17.9%) and some contaminated locations can be areas where soil ingestion of this metal is significant (e.g., 59% of the total intake; Table 18.1). To date, however, there is little information on the gut uptake of these elements and their transfer into animal products. In southwest England, much of the ingested arsenic is not available for uptake, and it is consequently found in animal feces. Still, some of the

arsenic is known to be absorbed, and it has been reported that owners of arsenic-contaminated land in southwest England rent their fields to farmers wishing to present their livestock at agricultural shows. The resulting elevated intake and absorption of arsenic leads to a nice "bloom" on the coats of the animals thus improving their appearance.

Ingested soils also have the potential to reduce the uptake of elements by animals. Adsorption on the organo-mineral cation exchange complex; sorption by hydrous oxides of aluminum, iron, and manganese; and the formation of stable complexes with soil organic matter are all mechanisms that can reduce the availability of elements to animals. In addition, antagonistic elements can be released from ingested soil. For example, the release of iron or molybdenum from soil in the stomach of sheep may interfere with copper metabolism, which leads to a disease known in the UK as swayback (Suttle et al. 1975, 1984). This disease affects the nervous system of newborn lambs. Its name is derived from the characteristic uncoordinated gait with the back legs swaying from side to side. In the UK this disease is more severe after mild winters, and forecasts as to the severity of the disease in lambs can be made on the basis of the number of days with snow cover. During mild, snow-free winters, soil ingestion rates are elevated and a high risk of swayback in spring is likely since intake of soil-contaminated herbage reduces the availability of copper to the ewe. Conversely, severe winters encourage supplementary feeding, limiting the intake of soil, and reducing the risk of swayback (although an additional factor here is that supplementary feeding of in-lamb ewes is likely to provide more copper than from pasture herbage alone) (see also Chap. 21, this volume).

18.2 Geophagy in the Animal Kingdom

The deliberate ingestion of soil has been observed in both domesticated and wild animals, although research investigations have concentrated almost exclusively on the latter. Among terrestrial vertebrates, geophagy has been reported in many species of birds, reptiles, and mammals. Whereas humans may be geophagists and are a member of the latter class, this section of the chapter considers only other animal species. Table 18.2 illustrates that only carnivores have not been observed deliberately eating soil. The listing shown in this table, however, is certainly not complete with ongoing research adding knowledge to the number of species that indulge in geophagy. For example, avian geophagy was reported from the tropical island of New Guinea for the first time in 1999 when 11 bird species (all predominantly herbivores, and especially frugivores) were observed consuming soil (Diamond et al. 1999). It is also worth reporting that geophagy is not restricted to vertebrates; for example, isopods and butterflies are also known to deliberately consume soil.

Geophagy is a widespread practice that is reported from many parts of the globe. However, most observations on geophagical behavior come from North America and the savanna of Africa (Kreulen and Jager 1984), perhaps reflecting a bias in the study of the ecosystems and animal species of these areas. Usually the soil intake is selective, with specific sites and sometimes even particular soil horizons being exploited (Fig. 18.2). In the literature, these locations are variously referred to as mining sites, salt licks, natural salt licks, salines, mineral licks, natural mineral licks, or natural licks (Klaus and Schmid 1998). The use of such terminology at times can be misleading, because the words "mineral" and "salt" suggest a chemical enrichment of soil, with animals indulging in geophagy to satisfy a mineral nutrient imbalance. This may be the case, but there are a variety of reasons as to why soils are deliberately consumed by animals, and not all are enriched in mineral nutrients. Consequently, the simple term "lick" is perhaps best, because its use does not imply a specific benefit that is gained from the soil.

The size of lick sites varies from small, unspectacular scrapes to large, treeless sites like those found within tropical rainforests. For example, in a study undertaken in Dzanga National Park (Central African Republic), the licks varied in size from 2,000 to 55,000 m^2, with holes and caves excavated by the trunks, tusks, and front legs of the forest elephant, *Loxodonta africana cyclotis* (Klaus et al. 1998). There is considerable variability in the use of licks. Sites are not necessarily used by all species in an area, and while a particular species may utilize a lick at one location, the same species may ignore licks in other areas. Some observations on ungulates have recorded no differences in lick use among different sex and age groups, whereas other reports observe geophagy mainly or exclusively in pregnant or lactating females and/or juveniles. In the Yankari Game Reserve (Nigeria), all ages of warthog (*Phacochoerus aethiopicus*) exploit licks, in contrast to (mainly) adolescent hartebeest (*Alcelaphus buselaphus*; Henshaw and Ayeni 1971). A seasonal use of licks is also evident (Kreulen and Jager 1984). In both North America and Europe, a peak in lick use by ungulates is linked to forage changes during leaf flush in the spring. Similarly, in the arid areas of southern Africa, seasonality of lick use is associated with leaf flush at the beginning of the wet season. For the humid tropics, information on the seasonality of lick use is limited, although red leaf monkeys (*Presbytis rubicunda*) have been observed consuming mineral nutrient enriched soil from termite mounds in the rainforests of northern Borneo from April to August (Davies and Baillie 1988). In contrast, the forest elephants in Dzanga National Park visit licks throughout the year though lick use decreases during the main fruiting season (Klaus et al. 1998).

Table 18.2 Taxonomic categories of reptiles, birds, and mammals that engage in geophagy

Class	Families	Representatives
Reptilia	Iguanidae	Iguana
	Emydidae	Box turtle
	Testudinidae	Tortoise
Aves	Struthionidae	Ostrich
	Anatidae	Goose
	Aegypiidae	Palm-nut vulture
	Phasianidae	Pheasant
	Numididae	Guinea fowl
	Columbidae	Dove, pigeon
	Psittacidae	Parrot
	Musophagidae	Turaco
	Coliidae	Mousebird
	Sturnidae	Starling
	Ploceidae	Sparrow, weaver
	Fringillidae	Canary, bunting
Mammalia	Leporidae	Rabbit
	Sciuridae	Squirrel, woodchuck
	Erethizontidae	American porcupine
	Elephantidae	Elephant
	Equidae	Horse, ass, zebra
	Tapiridae	Tapir
	Rhinocerotidae	Black rhino
	Suidae	Bushpig, warthog, wild boar
	Tayassuidae	Peccary
	Camelidae	Camel
	Cervidae	Caribou, moose, mule deer, roe deer, sambar, white-tailed deer
	Giraffidae	Giraffe
	Bovidae	Antelope (e.g., duiker, gazelle), bighorn (e.g., Dall sheep), mountain goat, African buffalo, banteng, gaur, domestic ox, goat, sheep
	Indridae	Lemur
	Cercopithecidae	Baboon
	Colobidae	Colobus, langur
	Hylobatidae	Gibbon
	Pongidae	Chimpanzee, gorilla
	Hominidae	Man

After Kreulen and Jager (1984)

The selective use of lick sites indicates that soils of these locations have certain qualities that animals find desirable. In northeastern Peru, mustached tamarins (*Saguinus mystax*) have been observed eating soil from a broken mound of leaf-cutting ants (Heymann and Hartmann 1991). Geochemical analysis of the fine-textured soil revealed elevated concentrations of a number of elements (e.g., iron and potassium), which are attributable to the ants constructing the mound with deeper soil materials that are less leached than the surrounding surface soil. Red leaf monkeys in northern Borneo (see above) and chimpanzees (*Pan troglodytes schweinfurthii*) in Tanzania (Mahaney et al. 1996) are examples of other primates that have been reported to feed on soils (in both cases from termite mounds) that are similarly enriched in potentially beneficial mineral nutrients (such as calcium, magnesium, potassium, and phosphorus) and clay minerals. In the Kalahari sandveld of Botswana, geomorphological processes are important in the formation of licks, with fine-textured and nutrient-rich material accumulating in depressions (called pans) by sheet flow from adjacent areas following periods of heavy rainfall (Kreulen and Jager 1984). The properties of lick soils can vary a great deal. For example, in the 1,500 km^2 Mkomazi Game Reserve located in Tanzania, the three known lick soils show considerable chemical and mineralogical variability (Abrahams 1999) (Table 18.3). Despite such differences, some common properties of lick soils are

- A high content of clay-sized particles

Fig. 18.2 A soil lick located in the Mkomazi Game Reserve, Tanzania. The reserve ranger is standing in an excavation mined by animals, and other holes made by elephant tusks can be seen. This extremely alkaline, highly calcareous and saline-sodic soil, referred to as site 2 in Table 18.3, may provide a range of benefits (e.g., sodium supplementation, an antacid function) to animals if consumed in appropriate amounts (From Abrahams 1999; Plenum Publishing Corporation)

- A high salinity
- Among saline (halomorphic) licks, sodium chloride and/or sodium sulfate may predominate in neutral or slightly acidic soils, whereas sodium carbonate and sodium bicarbonate are associated with alkaline lick sites
- High quantities of calcium and/or magnesium carbonate
- Licks may be chemically enriched (e.g., in nitrogen, sulfur, potassium, and phosphorus) because of fecal and urinary contamination; the excreta from diseased animals will heavily laden sites with pests, cysts, and nematodes

A common practice by research investigators is to compare the properties of lick samples with non-lick soils found in the same region. Differences, if any, between the two soil types can then lead to suggestions as to why animals are indulging in geophagy. Sodium often appears to be the major attracting substance of many licks (e.g., site 2, Table 18.3), and it is known that herbivores can seek extra sources of this macronutrient because they have a sodium-specific perception and hunger mechanism that is activated, among other things, by depletion of this element (Denton 1969). It is also appreciated that terrestrial plants may not accumulate sufficient sodium to satisfy the nutrient demands of an animal. There are a number of factors that can account for a seasonal demand of sodium that matches the periods of use of licks as noted above. For example, a temporary large increase in urinary and fecal output of sodium (attributable to a dietary change caused by the sudden transition from winter or dry season roughage to lush grass/browsing plants at the onset of spring or the wet season) will create a seasonal demand for this macronutrient that may be satisfied by the ingestion of appropriate soils (Kreulen and Jager 1984).

Even though there is strong evidence that sodium-rich soils are a cause of geophagy, there are licks that are not enriched in this element (e.g., site 1, Table 18.3) and clearly other benefits are sought by the geophagists. Licks may be exploited for mineral nutrients other than sodium, and calcium, iron, phosphorus, and sulfur have all been suggested as target elements. However, most licks contain relatively low concentrations of phosphorus, and wild animals indulge in osteophagy (the consumption of bone) as a source of this element, rather than geophagy. At high altitudes in the tropics, it has been suggested that ingested soils supplement African buffaloes (*Syncerus caffer caffer* (Sparrman)) and mountain gorillas (*Gorilla gorilla beringei*) with iron (Mahaney and Hancock 1990; Mahaney et al. 1990). These animals may require relatively large amounts of this mineral nutrient for erythrocyte formation, in the same way that it is known that humans living at high altitudes need iron-rich food to increase erythrocytes in the blood.

The consumption of soil to obtain calcium has been substantiated for reptiles (e.g., the desert tortoise, *Gopherus agassizii*; Marlow and Tollestrup 1982) and birds, though for the latter the most common explanation for geophagy is to provide grit. Because birds lack teeth, many ingest pebbles or coarse soil particles for the grinding of food in their gizzards (reptiles and ruminants are also known to ingest soil for the breakdown of food). But licks used by parrots in Peru (Figs. 18.3 and 18.4) are fine-textured, which strongly suggests that these soils are not consumed to aid digestion (Gilardi et al. 1999). Instead the lick soils have a higher cation exchange capacity (CEC) than those from non-preferred sites, and bioassays have shown the ability of the soils to adsorb toxins (such as quinine) associated with the birds' plant diet. Geophagy by seed-eating birds has also been observed elsewhere, and it would appear that this soil consumption represents one weapon in the escalating "biological warfare" between plants and animals (Diamond 1999). From a plant's evolutionary perspective, a seed needs to be enriched in nutrients both to support germination and subsequent growth, while nutrient-rich fruits attract animals like birds that disperse seeds following plucking and consumption. However, chemical toxins in seeds and fruit will be repulsive to animals thus inducing regurgitation or defecation of the former and deterring the harvest of the latter

Table 18.3 Selected geochemistry[a] and mineralogy of three lick soils from the Mkomazi Game Reserve, Tanzania

Geochemistry												
		Calcium		Iron	Potassium		Magnesium		Sodium		Phosphorus	
Site	pH	Total	Extr.	Total	Total	Extr.	Total	Extr.	Total	Extr.	Total	Extr.
1	4.3	280	230	42,000	2,000	225	940	88	120	9	1,250	8
2	11.0	91,000	360	8,400	160	7.5	8,000	13	12,600	11,875	<100	4
3	8.2	7,450	2,400	29,400	3,300	188	5,400	1,500	5,040	3,188	1,600	<8

Mineralogy												
	Smectite	Illite	Kaolinite	Amphibole	Quartz	K Feldspar	Plagioclase	Calcite	Dolomite	Halite	Pyrite	Total
Site												
1	0.0	0.0	41.0	0.0	52.6	2.4	1.8	0.0	0.0	0.0	2.2	100
2	11.6	0.0	7.3	0.0	17.8	5.7	26.3	26.1	3.8	1.4	0.0	100
3	2.3	8.8	12.0	8.7	32.2	10.0	16.9	0.0	1.5	3.7	3.9	100

[a]pH measured in 1:2.5 w/v water suspension. Total and extractable (Extr.) concentrations in mg kg^{-1}. The extractable concentrations are the water-soluble and exchangeable (i.e., adsorbed on soil particle surfaces) fraction of the element in the soil
From Abrahams (1999), Plenum Publishing Corporation

Fig. 18.3 Red and Green Macaw eating a chunk of soil, Manu National Park, Peru (Frans Lanting/Minden/FLPA)

until the seed is viable. From an animal's evolutionary perspective, by overcoming the plant's toxin defenses a creature will obtain nutrients from seeds and fruit, and will outcompete other animals that find the diet repulsive and unpalatable. The ability of parrots (and other birds) to overcome plant toxins by indulging in geophagy would appear to suggest that they excel at the evolutionary "arms race" that exists between plants and animals. Many other animals may also benefit from geophagy and the ability of ingested soils to effectively detoxify plants. For example, this hypothesis has been proposed to explain the deliberate ingestion of soil undertaken by at least 14 species of non-human primates (Krishnamani and Mahaney 2000), including apes (e.g., Sumatran orangutan, *Pongo pygmaeus abelii*), prosimians (mongoose lemur, *Eulemur mongoz*), New World monkeys (masked titi monkey, *Callicebus personatus melanochir*), and Old World monkeys (guereza monkey, *Colobus guereza*).

The research undertaken on the geophagous parrots of Peru has indicated that ingested soils may also serve a function other than the adsorption of toxins. Gastrointestinal cytoprotection results from the interaction of high surface area clays such as smectite and attapulgite with the gut lining (Gilardi et al. 1999). By increasing mucus secretion and preventing mucolysis, clays in the gastrointestinal tract enhance the ability of the mucus barrier in protecting the gut lining from either chemical or biological insults, thereby alleviating the symptoms of diarrhea. The ingestion of soils containing clay minerals with a moderate to high surface area, together with the long time of passage of soil through the gastrointestinal tract, suggests the possibility of cytoprotection as an important function of geophagy undertaken by the Peruvian parrots.

Diarrhea (and other gastrointestinal upsets) can also be cured by geophagical practices because clay minerals are able to adsorb bacteria and their toxins. For example, chimpanzees in the Mahale Mountains National Park (Tanzania) have been observed consuming soil from termite mounds containing the clay minerals metahalloysite and smectite (Mahaney et al. 1996). This mineralogy makes the soil similar to the pharmaceutical Kaopectate that is used to treat minor gastric ailments in humans. The soils consumed by the chimpanzees could also function as antacids, with the commonly alkaline termite mound soils acting as a buffering agent to counteract the effects of acidic foods. Acidosis can

Fig. 18.4 Red and Green Macaw group feeding at a lick site, Manu National Park, Peru (Frans Lanting/Minden/FLPA)

also afflict wild ruminants such as the giraffe (*Giraffa camelopardalis*) and wild ungulates such as the mountain goat (*Oreamnos americanus*). The problem of acidosis associated with these animals arises due to the sudden dietary changes and the lush growth of vegetation that is coincident with the onset of spring or the early wet season (Kreulen 1985). A sudden lack of fiber and the increase in readily fermentable carbohydrates (e.g., sugars) and soluble proteins lead to a drop in stomach pH that causes several ailments such as anorexia, diarrhea, and gastrointestinal irritation. As a source of calcium carbonate, potassium carbonate, sodium bicarbonate, sodium chloride, and montmorillonite clays, ingested soils can avert acidosis by preventing a decline in stomach pH and by improving digestion efficiency through altering the sites of digestion and absorption of carbohydrates and proteins.

Other motives that may lead to the deliberate ingestion of soil have been suggested (Kreulen 1985; Klaus and Schmid 1998; Krishnamani and Mahaney 2000). These include:

- The use of soil as a famine food to ease the pangs of hunger during periods of starvation
- Microbial inoculation, where the ingestion of feces-contaminated soil facilitates the transfer of bacteria between animals, thus accelerating digestive adaptation within a population during periods of dietary change
- A behavioral tradition, where animals ingest soil because others are doing likewise

With licks providing a number of potential benefits to consumers, it is not surprising that geophagy has such a wide distribution in the animal kingdom. However, there are a number of costs that are also associated with geophagy, which include:

- The adverse physical effects of excessive tooth wear, erosion of the mucosal surfaces of the stomach and intestines, and obstruction of the digestive tract. Soils enriched in silica or sodium bicarbonate may be responsible for the development of kidney stones.
- The adsorption of nutrients by (for example) clay minerals may cause deficiency symptoms, while an excessive intake of an element can lead to mineral nutrient imbalances or problems of toxicity.
- Feces and urine accumulation at lick sites may cause problems of parasitism and disease. Soil fungi produce antibiotics that may have a bacteriostatic effect in the stomachs of animals such as ungulates.
- The attraction of animals to licks is associated with energetic costs and time lost for foraging. Lick sites may also be focal points of disease transmission and predation (including poaching).

Clearly, licks must provide benefits that enhance both animal performance and resource utilization, which compensate for the costs and risks associated with their use. "Aversion learning" may also lessen some of the adverse effects that are linked with geophagy (Kreulen 1985). This practice is important not only to individual animals but, by influencing population densities and structures, it also has broader ecological consequences. Yet the extent of research dedicated to this practice is, to date, relatively limited. Consequently, much of

what we know about geophagy practiced by wild animals is speculative, and many questions relating to (for example) how animals find appropriate soils for consumption, why they ingest them, the quantities that are consumed, and the implications of the soil ingestion still remain to be fully answered.

18.3 Involuntary Soil Ingestion: Humans

All members of an exposed human population will ingest at least small quantities of soil. Foods, for example, may be contaminated with soil particles that are then inadvertently ingested. This contamination is especially likely in the tropics because of the tradition of drying foodstuffs like cassava and millet outdoors. Soil can also be ingested via inhalation. Particles entrained in the aboveground air can be inhaled, but while some will reach and be retained in the lungs, the bulk is trapped and ultimately taken over the epiglottis into the esophagus before passing through the gastrointestinal tract. Soil particles adhering to the skin of fingers can also be involuntarily ingested by so-called hand-to-mouth activity. Young children in particular can ingest significant amounts of soil through this behavior; their hands are typically contaminated with soil through normal play activities (Fig. 18.5).

Most research undertaken on the involuntary ingestion of soil by humans has concentrated on young children. This group of the population can be expected to ingest the greatest quantities of soil involuntarily, and it will be the most vulnerable to any health effects. Research is difficult to undertake on such people, because observations that do not disturb children are difficult to conduct. Attempts have been made to estimate soil ingestion rates through recording the amount of soil on a child's hand and estimating the frequency of finger or thumb sucking (Ferguson and Marsh 1993). However, ingestion cannot be estimated reliably without some knowledge of how much soil is removed during each mouthing action. This information has not been well recorded, and there is further inadequate knowledge about the frequency of mouthing, and how much soil is retained on the hands following skin contact. Simple "soil-on-finger" estimates have therefore proved inadequate, and they have been superceded by more elaborate experiments using tracer elements. The ideal tracer element for estimates of soil ingestion by humans is one that is not present in food (or water or air or medications), is uniformly present in high concentrations in soil, and is poorly absorbed via the gastrointestinal tract (Binder et al. 1986). Another criterion for a gold standard tracer element is that the soil concentration should not vary significantly by particle size. No element exactly meets these ideal criteria with, for example, all tracer elements found to some extent in food. The mass-balance equation for a tracer element can be written:

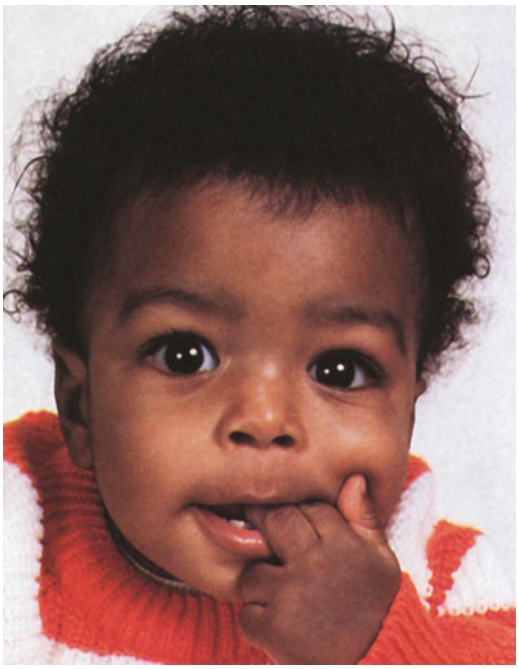


Fig. 18.5 Hand-to-mouth activity displayed by a young child (Erika Stone; Photo Researchers, Inc.)

$$I_a + I_{fo} + I_s + I_w = O_f + O_u \tag{18.2}$$

where the subscripts refer to intakes (I) of the element in air, food, soil, and water, respectively, and outputs (O) in feces and urine. Because some of these inputs and outputs are negligible for an ideal tracer element, Eq. 18.2 can be simplified to:

$$I_{fo} + I_s = O_f \tag{18.3}$$

that results in the soil ingestion estimate:

$$S_a = \left(O_f - I_{fo}\right)/S_c \tag{18.4}$$

where S_a is the mass of soil ingested and S_c is the concentration of the tracer element in the soil (Stanek and Calabrese 1991).

Table 18.4 records data from three tracer studies estimating soil ingestion by children, and the varying values illustrate the difficulties in interpretation. For example, Study 1 made a number of assumptions including that the daily stool output averaged 15 g (dry weight) per child. A later adjusted recalculation using measured stool weights instead of the 15 g assumed in the original work gives lower estimates of soil ingestion. Table 18.4 also illustrates the problem of poor intertracer consistency caused by errors in the mass-balance

Table 18.4 Estimates of soil ingestion (mg day^{-1}) by children

	Study 1		Study 1 (adjusted)		Study 2		Study 3		Study 3 (adjusted)
Tracer element	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean
Aluminum	181	121	97	48	39	25	153	29	136
Silicon	184	136	85	60	82	59	154	40	133
Titanium	1,834	618	1,004	293	246	81	218	55	208
Vanadium							459	96	148
Yttrium							85	9	97
Zirconium							21	16	113

From Calabrese and Stanek (1995)
Note: Study 1: 59 children 1–3 years of age; Study 2: 101 children 2–7 years of age; Study 3: 64 children 1–4 years of age

studies. For example, because titanium is widely used in inks, soil ingestion can be overestimated if a child eats printed paper or ingests ink residues sticking to the fingers. Such errors can be quantified, leading to adjustments in the soil ingestion estimates. This was done for the third study illustrated in Table 18.4. Using the original uncorrected data for six tracer elements, the mean soil ingestion estimates ranged from 21 to 459 mg day^{-1}. A marked improvement following adjustment led to a narrower range of 97–208 mg day^{-1}. Aluminum, silicon, and yttrium were considered to be the most reliable tracer elements in this particular study.

Despite efforts to improve the design of these soil ingestion studies, there still remains the problem of intertracer variability and the determination of which tracer element provides the best estimate of soil ingestion. There is a lack of information regarding the true variability of ingestion and the uncertainty of any average intake values. The limited number of investigations also provides little knowledge regarding factors such as the seasonal, regional, or ethnic variations in the rates of soil ingestion. Research undertaken in The Netherlands suggests that soil ingestion by children occurs mainly when the weather is dry and more time is spent outdoors (van Wijnen et al. 1990). The studies to date, however, do not address those children of tribal societies who live in a subsistence economy and who are most likely to inadvertently ingest the highest amounts of soil (Simon 1998).

Young children with high hand-to-mouth activity will ingest more soil than older children, who in turn will involuntarily consume more soil than adults. Relative to studies investigating children, data quantifying the rates of soil ingestion by adults are more limited. The U. S. Environmental Protection Agency (EPA) concluded that adults could ingest 100 mg of soil day^{-1}, but this guidance figure would appear to be an overestimate. A pilot study undertaken on just six individuals, and using the tracers aluminum and yttrium, suggested a soil ingestion rate of approximately 50 mg day^{-1} (Calabrese et al. 1990). Further studies on ten adults used a tracer-based mass-balance study over 28 days of observation. The findings, representing the largest amount of data available on soil ingestion by adults, indicated an average estimate of 10 mg of soil day^{-1} (Stanek et al. 1997).

Regarding health, the implications of this involuntary soil ingestion to humans may prove to be beneficial. With foods in developing countries often contaminated with iron from soil residues on vegetables and cereals, it has been suggested that the contamination could be a good dietary source of this important mineral nutrient (Hallberg and Björn-Rasmussen 1981). Considerably more work has emphasized the potentially deleterious effects of soil ingestion on health. In particular, ingested soils are likely to be a significant source of contaminants such as dioxins, and PHEs such as lead and the radionuclide isotopes, because there is only a limited uptake into the aerial parts of plants of these constituents from soils. In parts of Derbyshire (UK), soils are enriched in lead due to the weathering of mineral deposits and contamination associated with metalliferous mining. In a study undertaken in Derbyshire during the early 1970s, soil ingestion was found to be prevalent to an unexpected degree with 43% of children aged 2–3 years showing a pica for soil (Barltrop et al. 1974). An increased absorption of lead was found among children residing in villages near the extensive old mine workings, but the values found in blood and hair were still within the accepted normal range (Table 18.5). A subsequent investigation in the same area showed that handwipe samples from children have relatively high lead concentrations, which suggested the importance of hand-to-mouth activity in transferring significant quantities of the metal to the child (Cotter-Howells and Thornton 1991). The same investigation showed that many of the lead-rich soil grains were composed of pyromorphite, a stable soil–lead mineral that is only very slightly soluble and (presumably) has a low bioavailability to humans following ingestion.

Lead poisoning is a very important issue in the United States with medical, learning, and social costs having broad and long-term implications. There has been a substantial decline in blood lead levels during the last decade, yet there are still 900,000 American children under 6 years old that have blood lead concentrations high enough to suggest impairment of intelligence, behavior, and development. Urban soils in large American cities form a reservoir of lead and other PHEs such as cadmium and zinc because of pollutants that include leaded gas and paint (Mielke et al. 1999). The ingestion of soil from

Table 18.5 Lead in blood and hair samples collected from children residing in high and low soil-lead areas of Derbyshire

	Geometric mean	
	Blood (μg 100 ml^{-1})	Hair (mg kg^{-1})
High soil-lead area:[a]		
No current pica ($n = 27$)	23.6	10.8
Present pica for soil ($n = 16$)	26.4	21.1
Low soil-lead area[b]		
No current pica ($n = 17$)	19.9	5.7
Present pica for soil ($n = 16$)	22.1	9.0

From Barltrop et al. (1974)
[a]Mean soil-lead concentration about 10,000 mg kg^{-1}
[b]Mean soil-lead concentration about 500 mg kg^{-1}

gardens, school playgrounds, and other open spaces may therefore constitute a significant risk, especially because these urban soils will contain more soluble forms of lead (e.g., chloride and bromide) than the previously mentioned pyromorphite. Soil ingestion estimates are now routinely incorporated into all risk assessment procedures for contaminated sites in the United States (see also Chaps. 9 and 25, this volume).

Other health issues may also be associated with involuntary soil ingestion. For example, doctors in the United States have recorded eosinophilia (a high count of a type of white blood cell that is usually found when a toxin or parasitic infection is present) in children admitted for treatment of lead poisoning following hand-to-mouth activity (Berger and Hornstein 1980). The cause was attributable to infection with the larval form of the dog or cat parasite *Toxocara canis* or *T. cati*, respectively, which led to toxocariasis. Physicians managing children with lead intoxication following soil ingestion should therefore be aware of the possibility of concurrent parasitic infection. Toxoplasmosis is another disease associated with soil ingestion, attributable to the protozoan parasite *Toxoplasma gondii*, that sexually reproduces in cats who then release eggs in their feces to the soil. The ingestion of feces-contaminated soils by pigs, cattle, and sheep leads to their infection, and most people get toxoplasmosis from undercooked meat. However, the direct ingestion of soil by humans is a secondary source of infection. Medical opinion has insisted that *Toxoplasma* is nearly always harmless to people (Randerson 2002), but recent research has suggested that by adversely affecting human behavior latent toxoplasmosis, the mildest form of *T. gondii* infection, might represent a serious and highly underestimated economic and public health problem (Flegr et al. 2002).

18.4 Human Geophagy: Historical and Contemporary Perspectives

The recognition that geophagy is widespread among non-human primates suggests that the deliberate ingestion of soil predates our evolution as a species. The oldest evidence of geophagy practiced by humans comes from the prehistoric site at Kalambo Falls on the border between Zambia and Tanzania (Root-Bernstein and Root-Bernstein 2000). Here a calcium-rich white clay, believed to have been used for geophagical purposes, has been found alongside the bones of *Homo habilis* (the immediate predecessor of *Homo sapiens*). Migration transferred the practice from Africa to every other continent that has been permanently settled by humans, though there are some areas – Japan, Korea, much of Polynesia, Madagascar, and the south of South America – where geophagy is limited or unknown (Laufer 1930). However, a lack of reporting on the practice is common, and because many geophagists are also reluctant to admit to soil consumption, undoubtedly the prevalence of geophagy is greater than suggested from literature sources.

Throughout history a large number of writers, including anthropologists, explorers, scientists and physicians have commented on geophagy. In the first century AD, both Dioscorides and Pliny mention a famous medicament known as *terra sigillata* (earth that has been stamped with a seal) otherwise known as Lemnian Earth (Thompson 1913). This soil, derived from the Greek island of Lemnos, was mixed with the blood of a sacrificed goat, shaped into tablets somewhat larger than a thumbnail, stamped with an impression of a goat, and dried. Lemnian Earth was used for many maladies (but most notably for poisoning), and so great was the demand from the thirteenth to the fourteenth centuries, that almost every country in Europe strove to find within its boundaries a source of supply. Thus varieties of *terra sigillata* emanated from numerous localities including Bohemia, England, Italy, Malta, Portugal, Sicily, and the Mediterranean island of Samos. Some of these rival medicaments, notably *terra sigillata strigoniensis* or Strigian Earth derived from Silesia, acquired a considerable reputation. So valuable and respected were these that false Earths were also sold, leading Thevet to comment in 1554: "...adulterate it considerably when they sell it to people who have no knowledge of it" (Thompson 1913, p. 438). The fame of *terra sigillata* reached a peak at the end of the sixteenth century, and throughout the following 200 years the medicament was mentioned in most of the official medical books published in Europe. Its last appearance in any important pharmacopoeia was in 1848.

In the New World, geophagy was widespread previous to its discovery by Columbus. The oldest written history of Native Americans is provided by the explorer Alvar Nuñez Cabeza de Vaca who for 8 years (1528–1536) traveled through what is now known as the southeast United States (Loveland et al. 1989). Cabeza de Vaca writes of a tribe that was often exposed to starvation and they ate as much as they could, including soil. In another passage, the same explorer states that the fruit of the mesquite tree (*Prosopis juliflora*) was eaten with soil, making the food sweet and palatable. The Portuguese colonist and chronicler Gabriel Soares de Sousa provides the earliest (i.e., 1587) account of geophagy in South America. Commenting on the Tupinamba of Brazil, Soares describes how members of the tribe would commit suicide by eating soil "when they are seized by disgust or when they are grieved to such a degree that they are determined to die." The association of geophagy and suicide became a tragic part of American history. Slaves shipped across the Atlantic were responsible for the large-scale transfer of the practice from Africa. These slaves ingested soils perhaps to fill their stomachs as well as for medical (including nutritional) and cultural reasons (Hunter 1973). Additionally large numbers indulged in excessive soil consumption not only to become ill and to avoid work, but also to commit suicide in the belief that their spirit would return to the African homeland (Haller 1972). There are records of mass suicides caused by geophagy among plantation workers, and some estates were abandoned because of this practice (this problem was not confined to the Americas, since in 1687 approximately 50% of slaves in Jamaica died because of geophagy). Methods to deter the practice were harsh and included the use of facemasks, iron gags, chaining to plank floors, whipping, and confinement. The dismemberment of bodies of those who perished as a result of geophagy also proved effective, possibly because slaves believed that the spirit of a mutilated body could not return to the African homeland.

It was generally believed that African-Americans would cease the practice of geophagy following the termination of the slave trade and through the influence of Christianity. However, although the amelioration of life in the New World during the nineteenth century led to a decline of their medical problems that were associated with geophagy, the practice persisted for reasons other than as a means of committing suicide. For three-quarters of a century following the Civil War, geophagy in the United States was mentioned for the most part only incidentally in a few articles on the poor Whites of the American South as one of their numerous eccentricities. An anonymous writer in 1897 commented on the "clay-eaters" of Winston County, Alabama, who consumed a "dirty white" or "pale yellow" colored soil found along the banks of a small mountain stream (Anonymous 1897). The quantity eaten at any time varied from a pea-sized lump consumed by a child or beginner, to a piece the size of a man's fist for those who had eaten it for many years. Although the life expectancy of the clay-eaters was apparently not affected, they developed anemic, pale complexions. Several years later, Dr. Charles Wardell Stiles demonstrated that the anemia was attributable to hookworm disease that also caused the geophagy (ingested soils are reported to alleviate the gastric pain associated with hookworm disease).

A pioneering study reporting geophagy among African-Americans in 1941 proved to be the beginning of a number of important investigations on the subject (e.g., Dickins and Ford 1942). Geophagy was found to be extensive among black children, (especially pregnant) women of the American South, and the U. S. postal system used to deliver soil to friends and relatives who had migrated to the North. By the early 1970s, the practice could still readily be found, and was recorded as a structured custom embedded in a well-defined system of beliefs and rituals. Within a decade it was reported that the forces of urbanization and modernization had caused a decline of geophagy among the African-Americans (additionally the same report noted the increasing consumption of baking soda and laundry starch substitutes used instead of soil; Frate 1984). Nevertheless the practice of geophagy can still be found relatively easily. A medical report in the early 1990s indicated that the prevalence of pica (of which geophagy was a predominant form) had stabilized among pregnant women, affecting about one-fifth of high-risk patients (defined as rural blacks, with a positive family history of pica; Horner et al. 1991).

Relative to the number of studies undertaken on African-Americans, investigations on North American Indians are very limited. Nevertheless, geophagy was described in the early 1980s as widespread among certain desert-dwelling Indian tribes of the American southwest (Fisher et al. 1981). Here the consumption of soil fluorine has been reported to lead to skeletal fluorosis. The problem was exacerbated because people of such tribes have a high prevalence of renal impairment that results in a decreased excretion of fluoride.

Elsewhere in the world, geophagy can be commonly found in particular areas and among certain societies. For example, the practice remains widespread throughout Africa (Abrahams and Parsons 1996), and some recent research undertaken in Kenya indicates the contemporary prevalence of geophagy. In a cross-sectional study of 285 school children aged 5–18 years, 73% were reported to consume soil (Geissler et al. 1997). The prevalence decreased with age for both sexes up to the age of 15, then remained stable for girls between 15 and 18 years but continued to decrease for boys in that age range. The median amount consumed daily was 28 g, but it varied between individuals from 8 to 108 g. A cross-sectional survey of 275 pregnant women undertaken in the same country revealed that 56% consumed soil regularly

(Geissler et al. 1998). The median estimated daily intake was 41.5 g (range 2.5–219.0 g). Namibia is an example of an African country where the practice had not been reported until relatively recently. Then, in 1997, a study undertaken on 171 pregnant women in eastern Caprivi found that some 44% admitted to eating earth and utilized soil taken primarily from termite mounds (Thomson 1997). This investigation suggests that a lack of reporting on the practice may be contributing to a significant underestimation of geophagy. This is likely not only in Africa but also elsewhere such as the Middle East and Southeast Asia where few contemporary reports on the practice have been published. Hawass et al. (1987) recorded the first cases of geophagy undertaken by adults in Saudi Arabia, while the widespread and reasonably frequent occurrence of geophagy in Indonesia, where the practice has been ongoing for generations, has been reported by Mahaney et al. (2000).

Although geophagy can be found relatively easily in many developing countries and among the more tribally oriented people (e.g., the Aborigines of Australia; Abrahams and Parsons 1996), in many of the developed nations the forces of modernization and urbanization could be expected to lead to only infrequent reported cases. In the UK, pica undertaken by humans in the 1920s was described as very uncommon and confined almost entirely to the "abnormal cravings" of pregnant women and the "dirt-eating propensities" of children (Foster 1927, p. 72). In the early 1970s, a study of 100 pregnant women revealed cravings for normal food and drink substances, but pica was not recorded (Dickens and Trethowan 1971). Nevertheless, geophagy can be easily found in the UK by those with a desire to investigate the practice. Soil, variously known as sikor, mithi, patri, khuri, kattha, poorcha, or slatti, is traditionally taken by Asian ethnic groups as a remedy for indigestion and as a tonic during pregnancy (Fig. 18.6). The soil is imported from the Bengal region of south Asia, and is sold by weight in shops throughout the country (e.g., Birmingham, Bristol, London, and Swansea). Geophagy appears to be so well established as a custom of Bengali society, that immigration into the UK has resulted in the cultural transfer of the practice. The likelihood is that immigration into other developed nations has also transferred the practice in recent years or decades.

It is clear that geophagy is not limited to any particular age group, race, sex, geographic area, or time period. Nevertheless, the practice is especially associated with certain regions and people (e.g., contemporary developing nations, people of low socioeconomic status, children, pregnant women). Typically only specific soils are consumed, and are selected for desired qualities of (for example) color, odor, flavor, texture, and plasticity. Often the material is a ferruginous clay, but other soils are certainly sought. In Africa, sand may be exploited both within the Sahara Desert of Mauritania and Ghana's Volta River delta (Vermeer 1987).

Fig. 18.6 A 1-kg bag of sikor purchased from a shop in Birmingham, UK. Local pregnant Bengali women mainly consume the tablets that may be a significant source of iron (Photo: Peter W. Abrahams)

However, alluvial clay-enriched soil, from depths of 30–90 cm, are a common source of geophagical material. Scrapings from mudwalls of structures can also provide soil for occasional needs. In West Africa shales are mined and processed for geophagical use (Vermeer and Ferrell 1985), and a field study in eastern Sierra Leone indicated that 50% of pregnant women regularly consume clay found in the interior of termite mounds (Hunter 1984). In the same country another (less commonly used) geophagical source of clays are the nests of the mud-daubing wasp (genus *Synagris*).

Following collection, some rudimentary preparation of the soil may then occur. For example, tablets of sikor are made by compressing the soil prior to baking. The latter operation has the desired effect of destroying the eggs of potential intestinal parasites and produces the distinct smell of smoke to which the Bengali women are attracted. In eastern Guatemala, holy clay tablets known as pan del Señor (bread of the Lord) or *tierra santa* (sacred Earth) are produced following excavation (Hunter et al. 1989). Lumps of clay are pounded into small pieces before being crushed and passed through a 1-mm sieve. Water is added to produce a smooth dough-like material, and ceramic or wooden molds are then firmly pressed into the clay to produce tablets with a holy image. The tablets are sliced and trimmed with a carving knife and sun-dried for 24–48 h before redistribution to retailers throughout Guatemala and the neighboring countries of Belize, Honduras, and El Salvador (Fig. 18.7).

The qualitative and quantitative estimates of soil ingestion attributable to geophagy indicate the large quantities that can be consumed. For example, 3–4 tablets of sikor (about 64 g) may be ingested daily by pregnant Bengali women. The average daily consumption of soil by pregnant women in Africa has been reported as 30–50 g (Vermeer 1987). Sometimes geomania is encountered, whereby people develop a craving and uncontrollable urge for eating soil

Fig. 18.7 A vendor selling tablets of *tierra santa* in Central America. The soil is used as a pharmaceutical, and with its religious associations it provides psychological comfort. Here, the woman is daubing candy makers' red dye onto the tablets to simulate the blood of Jesus (Photo: John M. Hunter)

(Halsted 1968). This is evident from the quotations given to the interviewers of black women of the American South during the middle decades of the twentieth century: "I feel awful, just about crazy when I can't get clay"; "I craves it"; "I crave something sour like the taste of clay" (Edwards et al. 1959, p. 811; Ferguson and Keaton 1950, p. 463). The cravings may be difficult to control. Avicenna, the Arabian philosopher and physician who lived about 1,000 years ago, talks of the necessity of the whip in controlling boys from geophagy, while restraints or prison was used for older people. He also records how "incorrigible ones are abandoned to the grave," demonstrating the hazards of persistent and excessive soil consumption. Avicenna was the first to mention the benefit of iron preparations in treating geophagy, and the association between iron and the ingestion of soil is still of considerable debate today as will be discussed further in the next section.

18.5 The Causes of Human Geophagy

It may be thought that the causes of soil consumption can be easily established by interviewing geophagists. However, many people who undertake the practice are reluctant to admit to soil eating. Perhaps they have a sense of shame or guilt, and fear that the interviewers may view the trait negatively. In some societies, as commonly found in Africa, the practice is more overt and open discussion is possible (Vermeer 1987), but even in such situations many geophagists are at a loss to explain their desire for soil. One evident feature is that they generally like the practice, since many speak positively about the qualities (e.g., taste, feel, or odor) of the soil that is consumed. Despite the difficulties of obtaining information from geophagists, the practice is known to have multiple causes including those listed below.

18.5.1 Soil as a Food and Food Detoxifier

The use of soil as a food supplement during periods of famine has been frequently recorded, as the ingested soil gives a sensation of fullness to the stomach. For example, Alexander von Humboldt on his travels in South America at the beginning of the nineteenth century recorded the eating of clay by the Ottomac tribe in the Orinoco Valley (Ross 1895). Local supplies of fish and turtles were curtailed during the time of annual flooding, and Humboldt records how 12.5- to 15-cm diameter balls of mainly alluvial material were prepared and eaten by the Ottomacs in "prodigious quantities." Such consumption of soil is not restricted to distant times. In China, a country where traditional knowledge of famine foods has been transmitted between generations, soils have been utilized during famine as recently as about 60 years ago (Aufreiter et al. 1997), while in 2002 food shortages were reported to be causing geophagy in Malawi.

Closely associated to the practice of geophagy undertaken during periods of famine is the use of soils by humans in plant detoxification. Many plants containing toxins are consumed during periods of food shortage, but the mixing of soil with such plants adsorbs the potentially harmful chemicals and renders the food palatable. In this way the most important African famine food, the wild yam *Dioscorea dumentorum* Pax., is detoxified through the use of clay. Elsewhere, Native American populations have been reported to mix soil with acrid acorns, tubers, and berries as a corrective of taste. Indeed, the Aymara and Quechua people of the Andes Mountains of Bolivia and Peru still continue to eat wild potatoes by dipping them in a thick slurry of clay that effectively adsorbs potentially toxic glycoalkaloids (Johns 1986). The importance of these "potato clays" can be judged from the suggestion that because all wild potatoes are poisonous to humans, the domestication of the modern potato (the world's premier vegetatively propagated cultigen) may have required geophagy at first.

Perhaps as an extension of its use during times of famine, soil can also be consumed as a regular food item. A study of

African-American women in the late 1950s reported that many ate soil as part of the menu, and the meal seemed incomplete without it (Edwards et al. 1959). In Turkey, clay has been reported as used for snacks in place of candy or chewing gum. The literature also reports the use of soil as a sort of relish, condiment, or delicacy.

18.5.2 Psychiatric and Psychological Causes of Geophagy

Pica is common in institutionalized mentally retarded people, and soil is one of the nonfood items commonly sought. It may be considered that the mentally retarded cannot discriminate between food and nonfood items, but a study in the United States showed that individuals are often aggressive in seeking the nonfood item of their choice, and they are quite deliberate about what they ingest (Danford et al. 1982).

Other research undertaken in the United States has demonstrated that pica in early childhood is related to elevated, extreme, and diagnosable problems of bulimia nervosa in adolescents. This suggests that pica may be a symptom of a more general tendency to indiscriminant or uncontrolled eating behaviors.

A psychological cause for geophagy is also evident. For example, pregnant African-Americans have commented on feelings of anxiety and agitation that are overcome by experiencing a sense of satisfaction following the consumption of soil. Similarly, the holy clay tablets that are consumed in Central America (Fig. 18.7) provide psychological comfort by helping to allay anxieties associated with ill health or pregnancy (Hunter et al. 1989).

18.5.3 The Consumption of Soil as a Pharmaceutical

Soil may be the world's oldest medicine recorded as a pharmaceutical throughout history. The pre-modern Chinese extensively utilized soil in medicines with, for example, the pharmacologist Li Shi-Chen in 1590 listing 61 uses for soil materials in treating a variety of conditions (Root-Bernstein and Root-Bernstein 2000). Such multiple applications of soil in treating ailments is still evident (e.g., Fig. 18.8), although their effectiveness as a medicine for treating so many maladies must be questioned. The varieties of *terra sigillata* utilized throughout Europe for some 2,000 years are recorded as used for a number of ailments including plague, the bites and stings of venomous animals, malignant ulcers, nose bleeds, gout, dysentery, and poisoning (Thompson 1913). Certainly, *terra sigillata* would have proved effective in treating the latter due to the ion exchange capability of the soil constituents. Indeed, so effective are soils in treating cases of poisoning that fuller's earth and bentonite, both enriched in the clay mineral montmorillonite, are used in contemporary developed nations as an antidote for poisons such as the herbicides paraquat and diquat.

Contemporary modern societies also utilize kaolin and smectite clays for treating gastrointestinal disorders. It should be of no surprise, therefore, that the literature describing human geophagy mentions the effectiveness of ingested soils in treating gastrodynia (stomachache), dyspepsia (acid indigestion), nausea, and diarrhea. For example, the chief geophagical clay entering the well-developed West African market system comes from the village of Uzalla, Nigeria (Vermeer and Ferrell 1985). The clay, called *eko* by the villagers who prepare the material, is obtained from the working of Paleocene shales, and it is used as a traditional medicine including its use for, among other treatments, stomach and dysenteric ailments. Some 400–500 tonnes of *eko* is reportedly produced each year and widely redistributed to markets greater than 1,600 km from the source. Mineralogical analysis of *eko* indicates a kaolinitic composition similar to the clay in the modern pharmaceutical Kaopectate.

18.5.4 A Cultural Explanation for Geophagy

For some people there are symbolic links between themselves, fertility, blood, ancestors, and graves that are strengthened through the ingestion of soil. For example, among the Luo people of western Kenya, soil is eaten openly by women of reproductive age (Geissler 2000). There is a particular preference for soils from termite mounds that have a symbolic significance because of their red color (i.e., the color of blood), intense taste, fertility, the use of their material for building dwellings, and their location that may be coincident with sites of burials or former habitation. Luo children learn earth-eating by observation and imitation within the family, but girls and boys have different views on the practice. Boys have to stop indulging in geophagy in order to "become men."

18.5.5 The Consumption of Soil for Physiological Reasons

It is commonly assumed that humans ingest soils to satisfy a nutritional deficiency. Such a physiological explanation for geophagy is an attractive hypothesis, and certainly soils have the potential to supply mineral nutrients to the geophagist. Hydrochloric acid is secreted in the stomach and is a major component of gastric juice that is consequently strongly acidic (pH is dependent upon physiological parameters, varying from 1–2 when fasting to 2–5 when fed; Oomen et al. 2002). But the pancreatic secretion of bicarbonate ions

Fig. 18.8 A vendor selling traditional soil/herbal remedies (the cylindrical-shaped objects) in the central market of Kampala, Uganda. The distinctive medicines, typically broken up and mixed with water before drinking, are used for treating a large variety of ailments such as asthma, nausea and vomiting, syphilis, poisoning, and anemia (Photo: Peter W. Abrahams)

that neutralizes acid from the stomach modifies conditions in the small intestine where a higher pH (i.e., less acidic) environment exists (the small intestine consists of three sections: duodenum, pH 4–5.5; jejunum, 5.5–7; and ileum, 7–7.5). The human digestive system can thus be considered to be a two-part, acid-alkaline extraction system that operates on any soil constituent passing through the gastrointestinal tract. For example, ingested clays encountering the acidity of the stomach will release elements by cation exchange reactions, while iron oxide and other minerals can be expected to be partially solubilized. Mineral nutrients released in the stomach may then be adsorbed by soil constituents as they enter the small intestine, because the adsorption of nutrients tends to increase with pH. Soil extractants such as 0.1M hydrochloric acid can be used in simple laboratory experiments to simulate human digestion and its effects on the availability of mineral nutrients to the geophagists. Table 18.6 provides a summary of such experimentation that has been undertaken on samples collected within Africa. This table illustrates the varying ability of the different soils in supplying mineral nutrients to the geophagist following their consumption.

As previously explained there is strong evidence that sodium-rich (saline) soils are a cause of soil ingestion by members of the animal kingdom. But salt has never been shown to be a stimulus in primate geophagy, and most soils utilized by humans are reported to be essentially salt free. This means that its craving is an improbable cause of the practice. Indeed geophagists can deliberately add salt to the soil that they then consume: a practice that is, for example, commonly undertaken by pregnant women in Nigeria. Calcium and iron are the two mineral nutrients that are frequently implicated in the physiological explanation for geophagy. Daily calcium needs increase for pregnant women from 800 to 1,200 mg day^{-1}, primarily to provide the nutrient for fetal skeletal growth and development. In Africa, research has demonstrated that geophagy is especially common among women belonging to non-dairying tribes that have consequently a depleted intake of calcium (Wiley and Katz 1998). The ingestion of calcium-rich soil therefore provides a plausible explanation for geophagy as was made, for example, in a study reported in 1966 on the non-milk-drinking Tiv tribe of Nigeria (see Table 18.6). However, such explanations remain speculative because there are no detailed, consistent, and well-controlled data to support any observation that human geophagy represents a craving generated by a nutritional deficiency (Feldman 1986; Reid 1992). Nevertheless, some soils do have the potential to supply various mineral nutrients to the geophagist in significant quantities, even if the soils are not consumed to satisfy a physiological requirement. Table 18.6 shows the varying quantities of iron that can be extracted from geophagical soils following laboratory experimentation. Some soils appear to be poor providers of iron, but others are capable of contributing toward a significant proportion of the Reference Nutrient Intake (RNI) for this element. For example, research undertaken on Kenyan girls and boys in 1998 (see Table 18.6) indicated that on average ingested soils were providing 32 and 42%, respectively, of the RNI for iron. However, such findings have been criticized more recently, because the laboratory extractions ignore the effect of changes in the Eh/pH regime and kinetics during passage of soil through the gastrointestinal tract. Consequently, new experimental methods for the estimation

Table 18.6 Extractable concentrations (mg kg^{-1}) of selected macro- and micronutrients determined from geophagical materials collected within Africa

Date of study and origin of sample	Calcium	Copper	Iron	Potassium	Magnesium	Manganese	Sodium	Zinc
1966, Nigeria[a]	3,910	–	–	53	2,005	–	44	–
1971, Ghana	120	–	–	165	31	–	–	–
1973, Ghana[b]	1,133	10	95	130	331	<1	–	15
1984, Nigeria[c]	265	0.6	134	41	179	29	–	30
1991, Cameroon	77	–	9	45	–	–	–	–
Gabon	68	–	4	87	–	–	–	–
Kenya 1	791	2	7	432	135	63	–	3
Kenya 2	220	1	12	793	112	349	–	5
Nigeria	19	2	10	102	9	*nd*	–	3
Togo	120	–	5	177	–	–	–	–
Zambia[d]	142	11	74	93	60	19	–	2
Zaire	16	–	497	84	–	–	–	–
1997, Uganda[e]	1,341	2.1	528	763	458	59	186	6.7
1997, Uganda	1,800	–	326	460	1,180	50	143	4
Zaire	440	–	380	1,730	4,100	12	3,140	2
1998, Kenya[f]	–	–	169	–	–	–	–	2.7
1998, Kenya[g]	–	–	103	–	–	–	–	1.7

Adapted from Abrahams and Parsons (1996)
nd not detected
[a]Mean of two samples collected from soil pits that are utilized by the Tiv tribe
[b]Median concentrations determined from 12 samples
[c]*Eko* clay
[d]Sample from the Kalambo Falls archaeological site
[e]Median concentrations determined from 12 samples used as traditional medicines
[f]Mean concentrations determined from 48 samples of soil that are typically consumed by boys and girls
[g]Mean concentrations determined from 27 samples of soil that are typically consumed by pregnant women

of the bioaccessible fraction of elements (defined as the fraction of a substance that is soluble in the gastrointestinal environment and is available for absorption) are being developed. One such method, the physiologically based extraction system (PBET), is an *in vitro* procedure that incorporates gastrointestinal tract parameters representative of a human (such as stomach and small intestine pH and chemistry, soil-to-solution ratio, stomach mixing, and stomach emptying rates). Experimental work simulating this extraction system has confirmed the bioaccessibility of iron from Ugandan soils including those from termite mounds. This indicates that their consumption will satisfy a major proportion of the geophagist's RNI for this nutrient (Smith et al. 2000). However, bioaccessibility estimates vary according to the type of *in vitro* procedure employed, and more research is required to establish which method most accurately reflects the human *in vivo* situation (Oomen et al. 2002).

18.6 Human Geophagy: Benefits and Banes

The preceding section reveals how ingested soils, as a medicament, food detoxifier, psychological comforter, and a supplier of mineral nutrients, can have a positive role in human society. Although people may seek a particular outcome by indulging in geophagy, at times the ingestion of soil can confer multiple benefits to the consumer. For example, there is a long association of the practice with pregnancy, and it has been suggested that during the first trimester ingested soils will adsorb dietary toxins that are potentially teratogenic to the embryo, while simultaneously quelling the common symptoms of pregnancy sickness. In the second trimester, when pregnancy sickness usually ends, soils may serve as a source of mineral nutrients, and calcium supplementation may aid in the formation of the fetal skeleton, and reduce the risk of pregnancy-induced hypertension (Wiley and Katz 1998). Despite such benefits, problems may arise if inappropriate quantities or types of soil are ingested. Paradoxically, even though soils can be a source of mineral nutrients to the geophagist, the cation exchange and adsorption properties of soil constituents have been reported to result in deficiency symptoms of certain elements. In Turkey, iron-deficiency anemia has been linked to the consumption of clays (mainly sepiolite and montmorillonite) of high CEC (Minnich et al. 1968). Clinical trials confirmed the effectiveness of the clays in adsorbing iron, and the conclusion of the study indicated the prominent role of geophagy in contributing to the problem, although other nutritional and parasitic factors were also probably involved in the anemia.

The adsorption of potassium by soil constituents can induce hypokalemia in an individual that is reflected by abnormally low concentrations of this element in the blood. Literature from the mid-1800s indicates that the condition was, along with iron-deficiency anemia, common among black slaves. The concurrent iron and potassium deficiency was associated with a disease known as cachexia Africana, the symptoms of which could be relieved through the use of iron- and potassium-containing tonics (Cragin 1836). Today, hypokalemia attributable to geophagy is only occasionally reported. For example, an isolated case was recorded in the late 1980s of an African-American woman who had a 25-year history of geophagy. An increase in her soil consumption produced a condition similar to that associated with cachexia Africana. However, symptoms abated following potassium replacement and cessation of soil consumption (Severance et al. 1988).

The association between geophagy and zinc deficiency has been noted in a number of countries. In Turkey, soil consumption can be commonly found among village women and children, and the practice is linked to a combined deficiency of iron and zinc, and the latter causes symptoms of growth retardation and delayed puberty (Çavdar et al. 1980). Physical growth and improved sexual maturation were observed in patients following zinc supplementation. There may be a number of causes of the pathogenesis of the zinc deficiency in these patients. The high cereal diet of the Turkish villagers provides little zinc because the phytate-rich food depresses the bioavailability of the metal (the phytate in cereals binds zinc to produce a highly insoluble complex that prevents its absorption). Thus in people who already have a low intake of zinc, geophagy can be considered as an accelerating factor leading to a deficiency of this metal, and ingested soil adsorbs significant quantities of zinc through cation exchange reactions. Additional factors may also be important; for example, of 300 Aboriginal people examined in a study located in the northwest of Australia, half of the individuals had low plasma zinc concentrations (hypozincemia). Geophagy and the high cereal diet of the Aborigines causes a decreased absorption of zinc, and an excessive loss of the metal from these individuals also occurs attributable to intestinal parasites and excessive perspiration. All the requirements are therefore present in the north of Australia for zinc deficiency to be widespread among Aboriginal people (Cheek et al. 1981).

Potentially life-threatening hyperkalemia, an abnormally high potassium concentration in the blood, has been associated with geophagy (Gelfand et al. 1975). This condition is attributable to the absorption of potassium released from ingested soils that are enriched in this element. However, hyperkalemia and its links with geophagy have only been occasionally reported, but the widespread contamination of soils with lead provides another example of toxicity that can be associated with the ingestion of soil, deliberate or otherwise. With lead being especially harmful to the developing brains and nervous systems of young people, there must be a concern if children are consuming soils enriched in this metal, especially with recent research suggesting that there may be no safe level of lead for children (Canfield et al. 2003). Furthermore, lead toxicity will not be restricted to children. As an example, Wedeen et al. (1978) reported on a 46-year-old American black woman who was found to have lead poisoning. The consumption of garden soil with a lead content of 700 mg kg^{-1} resulted in damage to the patient's red blood cells, brain, and kidneys. Yet studies investigating the lead intoxication of geophagists remain limited, and bearing in mind that potentially deleterious quantities of this metal may be bioavailable even from soils that contain normal amounts of lead, further investigations are urgently required.

The biotic component of soils can pose hazards to geophagists because the eggs or larvae of parasitic worms (geohelminths) can be consumed, although infection is likely to be significantly reduced if subsoil or baked soil is utilized. Ascariasis and trichuriasis are caused by the ingestion of *Ascaris lumbricoides* and *Trichuris trichiura* eggs, respectively, while toxocariasis occurs through infection with the larvae of *Toxocara canis* or *T. cati*. Hookworm infection can occur via the oral ingestion of *Ancylostoma duodenale* and *A. ceylanicum* (though skin contact with soil is the main cause of infection). It has been suggested that chronic liver disorders and cirrhotic changes may be associated with ingested soil bacteria and fungi.

Excessive tooth wear is another consequence of human geophagy, though the problem has been seldom reported in the literature (Abbey and Lombard 1973). Rather, more attention has focused on the internal accumulation of soil that can lead to constipation, the reduction of the power of absorption of food materials by the body, severe abdominal pain, and obstruction and perforation of the colon. In pregnant women, this can lead to dysfunctional labor and maternal death (Key et al. 1982). Deleterious outcomes in fetuses and infants of mothers who practice geophagy are also likely, although the lack of research means that the quality of any evidence is poor. Clearly, the strong association of geophagy with pregnancy warrants further investigations on this important topic.

18.7 Conclusions

This chapter considers both the involuntary and deliberate ingestion of soil by humans and other members of the animal kingdom. To many people, the word soil is commonly understood to be the material directly underfoot, as with organic-enriched topsoil. Although this material may be ingested involuntarily by humans and grazing animals, it is often strenuously avoided by geophagists. For example,

human geophagists in Africa commonly exploit material from excavations that extend into clay-enriched subsoils or even further to underlying soft shales *in situ*. This mining zone is free of most organic matter and parasitic infestation. In such cases, clay eating may be a more accurate term for the geophagical practice, rather than soil consumption. With the close association between geophagy and pregnancy, sometimes the expression pregnancy clay may be appropriate (Hunter 1993).

To date soil ingestion, whether involuntary or deliberate, has received relatively sporadic attention in the medical, sociologic, veterinary, or soil science literature. Yet such ingestion is demonstrably widespread and has important consequences for members of the animal kingdom. For example, it has been controversially suggested that modern urban human societies may experience health problems since their contact with (and ingestion of) soil is diminishing (Hamilton 1998). Consequent decreasing human exposure to soil mycobacteria may contribute to the increasing prevalence of allergic and autoimmune diseases (e.g., asthma, diabetes, rheumatoid arthritis) that have been observed in affluent societies over the past 30 years (Fig. 18.9). Similarly, a relatively recent decline of intestinal worm infection (e.g., *Ascaris lumbricoides*) in people of developed societies may be the cause of the increasingly common inflammatory bowel diseases that are now being recorded.

Relative to involuntary ingestion, geophagy is associated with a more substantial intake of soil. The functionality of non-human geophagy is not argued, whereas human geophagy has been typically viewed as a low-status, deviant, or highly suspect behavior, which is limited to marginal or deprived societies. A more enlightened appraisal is to realize that humans have frequently turned to geophagy as a useful way of overcoming problems that they experience. So important was the practice that it became embedded in the culture and customs of societies, which were perpetuated by learning rather than instinct. Although advances in education and medicine have caused a significant decline in the practice, geophagists still continue their indulgence in spite of the ill effects that can occur (perhaps some people may not connect any health problems with geophagy, or their beliefs are strong enough to overcome the fear of any ill effects).

There is evidence to suggest that geophagy is now attaining renewed and serious interest within the academic

Fig. 18.9 Dietary soil supplements for sale. Behind the humor of the cartoon is the serious message that contemporary urban societies may be at risk for ill health because they are not ingesting soil that can afford them appropriate protection (From Kate Charlesworth)

fraternity. Hopefully this will lead to future multidisciplinary research that will investigate the issues that have been hinted at in this chapter. For example, the role of ingested soil in either supplying mineral nutrients such as iron, or PHEs such as lead or radionuclides to humans needs to be quantified by undertaking properly controlled *in vitro* and/or *in vivo* studies. Such research will create a better understanding of the implications of soil ingestion that would benefit epidemiological and risk assessment studies. This should be considered as urgent bearing in mind the widespread nature of soil ingestion and human nutritional imbalances.

See Also the Following Chapters. Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 9 (Biological Responses of Elements) • Chapter 15 (Bioavailability of Elements in Soil) • Chapter 21 (Animals and Medical Geology) • Chapter 25 (Environmental Pathology)

Further Reading

Abbey LM, Lombard JA (1973) The etiological factors and clinical implications of pica: report of case. J Am Dent Assoc 87:885–887

Abrahams PW (1999) The chemistry and mineralogy of three savanna lick soils. J Chem Ecol 25:2215–2228

Abrahams PW, Parsons JA (1996) Geophagy in the tropics: a literature review. Geogr J 162:63–72

Abrahams PW, Steigmajer J (2003) Soil ingestion by sheep grazing the metal enriched floodplain soils of mid-Wales. Environ Geochem Health 25:17–24

Abrahams PW, Thornton I (1994) The contamination of agricultural land in the metalliferous province of southwest England: implications to livestock. Agric Ecosys Environ 48:125–137

Anonymous (1897) The clay eaters. Sci Am 76:150

Aufreiter S, Hancock RGV, Mahaney WC, Stambolic-Robb A, Sanmugadas K (1997) Geochemistry and mineralogy of soils eaten by humans. Int J Food Sci Nutr 48:293–305

Barltrop D, Strehlow CD, Thornton I, Webb JS (1974) Significance of high soil lead concentrations for childhood lead burdens. Environ Health Perspect 7:75–82

Beresford NA, Howard BJ (1991) The importance of soil adhered to vegetation as a source of radionuclides ingested by grazing animals. Sci Total Environ 107:237–254

Berger OG, Hornstein MD (1980) Eosinophilia and pica: lead or parasites? Lancet 1:553

Binder S, Sokal D, Maughan D (1986) Estimating soil ingestion: the use of tracer elements in estimating the amount of soil ingested by young children. Arch Environ Health 41:341–345

Brebner J, Suttle NF, Thornton I (1987) Assessing the availability of ingested soil cobalt for the synthesis of vitamin B_{12} in the ovine rumen. Proc Nutr Soc 46:66A

Calabrese EJ, Stanek EJ (1995) Resolving intertracer inconsistencies in soil ingestion estimation. Environ Health Perspect 103:454–457

Calabrese EJ, Stanek EJ, Gilbert CE, Barnes R (1990) Preliminary adult soil ingestion estimates: results of a pilot study. Regul Toxicol Pharmacol 12:88–95

Canfield RL, Henderson CR, Cory-Slechta DA, Cox C, Jusco TA, Lanphear BP (2003) Intellectual impairment in children with blood lead concentrations below 10 μg per deciliter. N Engl J Med 348:1517–1526

Çavdar AO, Arcasoy A, Cin S, Gümüs H (1980) Zinc deficiency in geophagia in Turkish children and response to treatment with zinc sulphate. Haematologica 65:403–408

Cheek DB, Smith RM, Spargo RM, Francis N (1981) Zinc, copper and environmental factors in the aboriginal peoples of the North West. Aust N Z J Med 11:508–512

Cotter-Howells J, Thornton I (1991) Sources and pathways of environmental lead to children in a Derbyshire mining village. Environ Geochem Health 13:127–135

Cragin FW (1836) Observations on cachexia Africana or dirt-eating. Am J Med Sci 17:356–364

Danford DE, Smith JC, Huber AM (1982) Pica and mineral status in the mentally retarded. Am J Clin Nutr 35:958–967

Davies AG, Baillie IC (1988) Soil eating by red leaf monkeys (*Presbytis rubicunda*) in Sabah, northern Borneo. Biotropica 20:252–258

Denton DA (1969) Salt appetite. Nutr Abstr Rev 39:1043–1049

Diamond JM (1999) Dirty eating for healthy living. Nature 400:120–121

Diamond J, Bishop KD, Gilardi JD (1999) Geophagy in New Guinea birds. Ibis 141:181–193

Dickens G, Trethowan WH (1971) Cravings and aversions during pregnancy. J Psychosom Res 15:259–268

Dickins D, Ford RN (1942) Geophagy (dirt eating) among Mississippi Negro school children. Am Sociol Rev 7:59–65

Edwards CH, McDonald S, Mitchell JR, Jones L, Mason L, Kemp AM, Laing D, Trigg L (1959) Clay- and cornstarch-eating women. J Am Diet Assoc 35:810–815

Feldman MD (1986) Pica: current perspectives. Psychosomatics 27:519–523

Ferguson JH, Keaton AG (1950) Studies of the diets of pregnant women in Mississippi: I. The ingestion of clay and laundry starch. New Orleans Med Surg J 102:460–463

Ferguson C, Marsh J (1993) Assessing human health risks from ingestion of contaminated soil. Land Contam Reclam 1:177–185

Fisher JR, Sievers ML, Takeshita RT, Caldwell H (1981) Skeletal fluorosis from eating soil. Ariz Med 38:833–835

Flegr J, Havlícek J, Kodym P, Maly M, Smahel Z (2002) Increased risk of traffic accidents in subjects with latent toxoplasmosis. BMC Infect Dis 2:11, www. biomedcentral. com/1471–2334/2/11

Foster JW (1927) Pica. Kenya East Afr Med J 4:68–76

Frate DA (1984) Last of the earth eaters. Sciences 24:34–38

Geissler PW (2000) The significance of earth-eating: social and cultural aspects of geophagy among Luo children. Africa 70:653–682

Geissler PW, Mwaniki DL, Thiong'o F, Friis H (1997) Geophagy among school children in western Kenya. Trop Med Int Health 2:624–630

Geissler PW, Shulman CE, Prince RJ, Mutemi W, Mnazi C, Friis H, Lowe B (1998) Geophagy, iron status and anaemia among pregnant women on the coast of Kenya. Trans R Soc Trop Med Hyg 92:549–553

Gelfand MC, Zarate A, Knepshield JH (1975) Geophagia: a cause of life-threatening hyperkalemia in patients with chronic renal failure. JAMA 234:738–740

Gilardi JD, Duffey SS, Munn CA, Tell LA (1999) Biochemical functions of geophagy in parrots: detoxification of dietary toxins and cytoprotective effects. J Chem Ecol 25:897–922

Green N, Johnson D, Wilkins BT (1996) Factors affecting the transfer of radionuclides to sheep grazing on pastures reclaimed from the sea. J Environ Radioact 30:173–183

Hallberg L, Björn-Rasmussen E (1981) Measurement of iron absorption from meals contaminated with iron. Am J Clin Nutr 34:2808–2815

Haller JS (1972) The Negro and the southern physician: a study of medical and racial attitudes 1800–1860. Med Hist 16: 238–253

Halsted JA (1968) Geophagia in man: its nature and nutritional effects. Am J Clin Nutr 21:1384–1393
Hamilton G (1998) Let them eat dirt. New Sci 159(2143):26–31
Harrison DL, Mol JCM, Healy WB (1970) DDT residues in sheep from the ingestion of soil. N Z J Agric Res 13:664–672
Hawass ND, Alnozha MM, Kolawole T (1987) Adult geophagia–report of three cases with review of the literature. Trop Geogr Med 39:191–195
Healy WB (1968) Ingestion of soil by dairy cows. N Z J Agric Res 11:487–499
Healy WB (1973) Nutritional aspects of soil ingestion by grazing animals. In: Butler GW, Bailey RW (eds) Chemistry and biochemistry of herbage. Academic, London, pp 567–588
Healy WB, Ludwig TG (1965) Wear of sheep's teeth. I. The role of ingested soil. N Z J Agric Res 8:737–752
Healy WB, Crouchley G, Gillett RL, Rankin PC, Watts HM (1972) Ingested soil and iodine deficiency in lambs. N Z J Agric Res 15:778–782
Henshaw J, Ayeni JSO (1971) Some aspects of big-game utilization of mineral licks in Yankari game reserve, Nigeria. East Afr Wildl J 9:73–82
Heymann EW, Hartmann G (1991) Geophagy in moustached tamarins, *Saguinus mystax* (Platyrrhini: Callitrichidae), at the Río Blanco, Peruvian Amazonia. Primates 32:533–537
Horner RD, Lackey CJ, Kolasa K, Warren K (1991) Pica practices of pregnant women. J Am Diet Assoc 91:34–38
Hunter JM (1973) Geophagy in Africa and the United States: a culture-nutrition hypothesis. Geogr Rev 63:170–195
Hunter JM (1984) Insect clay geophagy in Sierra Leone. J Cultur Geogr 4:2–13
Hunter JM (1993) Macroterme geophagy and pregnancy clays in southern Africa. J Cultur Geogr 14:69–92
Hunter JM, Horst OH, Thomas RN (1989) Religious geophagy as a cottage industry: the holy clay tablet of Esquipulas, Guatemala. Natl Geogr Res 5:281–295
Johns T (1986) Detoxification function of geophagy and domestication of the potato. J Chem Ecol 12:635–646
Key TC, Horger EO, Miller JM (1982) Geophagia as a cause of maternal death. Obstet Gynecol 60:525–526
Klaus G, Schmid B (1998) Geophagy at natural licks and mammal ecology: a review. Mammalia 62:481–497
Klaus G, Klaus-Hügi C, Schmid B (1998) Geophagy by large mammals at natural licks in the rain forest of the Dzanga National Park, Central African Republic. J Trop Ecol 14:829–839
Kreulen DA (1985) Lick use by large herbivores: a review of benefits and banes of soil consumption. Mamm Rev 15:107–123
Kreulen DA, Jager T (1984) The significance of soil ingestion in the utilization of arid rangelands by large herbivores with special reference to natural licks on the Kalahari pans. In: Gilchrist FMC, Mackie RI (eds) Herbivore nutrition in the subtropics. The Science Press, Johannesburg, pp 204–221
Krishnamani R, Mahaney WC (2000) Geophagy among primates: adaptive significance and ecological consequences. Anim Behav 59:899–915
Laufer B (1930) Geophagy. Anthropology series. Field Museum of Natural History, Chicago, Publication 280, vol 18, pp 97–198
Lee J, Rounce JR, Mackay AD, Grace ND (1996) Accumulation of cadmium with time in Romney sheep grazing ryegrass-white clover pasture: effect of cadmium from pasture and soil intake. Aust J Agric Res 47:877–894
Loveland CJ, Furst TH, Lauritzen GC (1989) Geophagia in human populations. Food Foodways 3:333–356
Mahaney WC, Hancock RGV (1990) Geochemical analysis of African buffalo geophagic sites and dung on Mount Kenya, East Africa. Mammalia 54:25–32
Mahaney WC, Watts DP, Hancock RGV (1990) Geophagia by mountain gorillas (*Gorilla gorilla beringei*) in the Virunga Mountains, Rwanda. Primates 31:113–120
Mahaney WC, Hancock RGV, Aufreiter S, Huffman MA (1996) Geochemistry and clay mineralogy of termite mound soil and the role of geophagy in chimpanzees of the Mahale Mountains, Tanzania. Primates 37:121–134
Mahaney WC, Milner MW, Mulyono H, Hancock RGV, Aufreiter S, Reich M, Wink M (2000) Mineral and chemical analyses of soils eaten by humans in Indonesia. Int J Environ Health Res 10:93–109
Marlow RW, Tollestrup K (1982) Mining and exploitation of natural mineral deposits by the desert tortoise, *Gopherus agassizii*. Anim Behav 30:475–478
Mayland HF, Florence AR, Rosenau RC, Lazar VA, Turner HA (1975) Soil ingestion by cattle on semiarid range as reflected by titanium analysis of feces. J Range Manag 28:448–452
Mielke HW, Gonzales CR, Smith MK, Mielke PW (1999) The urban environment and children's health: soils as an integrator of lead, zinc, and cadmium in New Orleans, Louisiana, U. S. A. Environ Res 81:117–129
Miller JK, Madsen FC, Hansard SL (1976) Absorption, excretion, and tissue deposition of titanium in sheep. J Dairy Sci 59:2008–2010
Minnich V, Okçuoğlu A, Tarcon Y, Arcasoy A, Cin S, Yörükoğlu O, Renda F, Demirağ B (1968) Pica in Turkey. II. Effect of clay upon iron absorption. Am J Clin Nutr 21:78–86
Oomen AG, Hack A, Minekus M, Zeijdner E, Cornelis C, Verstraete W, van de Wiele T, Wragg J, Rompelberg CJM, Sips AJAM, van Wijnen JH (2002) Comparison of five *in vitro* digestion models to study the bioaccessibility of soil contaminants. Environ Sci Technol 36:3326–3334
Pownall DB, Lucas RJ, Ross AD (1980) Effects of soil-contaminated feed on dry matter and water intake in sheep. Proc N Z Soc Anim Produc 40:106–110
Randerson J (2002) All in the mind? New Sci 176(2366):40–43
Reid RM (1992) Cultural and medical perspectives on geophagia. Med Anthropol 13:337–351
Root-Bernstein RS, Root-Bernstein MR (2000) Honey, mud, maggots, and other medical marvels. Pan, London
Ross T (trans) (1895) Personal narrative of travels to the equinoctial regions of America during the years 1799–1804 by Alexander von Humboldt and Aime Bonpland (eds), vol 2. Routledge, London
Severance HW, Holt T, Patrone NA, Chapman L (1988) Profound muscle weakness and hypokalemia due to clay ingestion. South Med J 81:272–274
Simon SL (1998) Soil ingestion by humans: a review of history, data, and etiology with application to risk assessment of radioactively contaminated soil. Health Phys 74:647–672
Smith B, Rawlins BG, Cordeiro MJAR, Hutchins MG, Tiberindwa JV, Sserunjogi L, Tomkins AM (2000) The bioaccessibility of essential and potentially toxic trace elements in tropical soils from Mukono district, Uganda. J Geol Soc 157:885–891
Stanek EJ, Calabrese EJ (1991) A guide to interpreting soil ingestion studies. 1. Development of a model to estimate the soil ingestion detection level of soil ingestion studies. Regul Toxicol Pharmacol 13:263–277
Stanek EJ, Calabrese EJ, Barnes R, Pekow P (1997) Soil ingestion in adults – results of a second pilot study. Ecotoxicol Environ Saf 36:249–257
Statham M, Bray AC (1975) Congenital goitre in sheep in southern Tasmania. Aust J Agric Res 26:751–768
Suttle NF, Alloway BJ, Thornton I (1975) An effect of soil ingestion on the utilization of dietary copper by sheep. J Agric Sci, Cambridge 84:249–254
Suttle NF, Abrahams P, Thornton I (1984) The role of a soil x dietary sulphur interaction in the impairment of copper absorption by ingested soil in sheep. J Agric Sci, Cambridge 103:81–86
Thompson CJS (1913) *Terra sigillata*, a famous medicament of ancient times. In Proceedings of the 17th international congress of medical sciences (Section 23, History of medicine), London, pp 433–444
Thomson J (1997) Anaemia in pregnant women in eastern Caprivi, Namibia. S Afr Med J 87:1544–1547

Thornton I (1974) Biochemical and soil ingestion studies in relation to the trace element nutrition of livestock. In: Hoekstra WG, Suttie JW, Ganther HE, Mertz W (eds) Trace element metabolism in animals 2. University Press, Baltimore, pp 451–454

van Wijnen JH, Clausing P, Brunekreef B (1990) Estimated soil ingestion by children. Environ Res 51:147–162

Vermeer DE (1986) Geophagy in the American South. Bull Shreveport Med Soc 37:38

Vermeer DE (1987) Geophagy in Africa. Bull Shreveport Med Soc 38:13–14

Vermeer DE, Ferrell RE (1985) Nigerian geophagical clay: a traditional antidiarrheal pharmaceutical. Science 227:634–636

Wedeen RP, Mallik DK, Batuman V, Bogden JD (1978) Geophagic lead nephropathy: case report. Environ Res 17:409–415

Wiley AS, Katz SH (1998) Geophagy in pregnancy: a test of a hypothesis. Curr Anthropol 39:532–545

Natural Aerosolic Mineral Dusts and Human Health

19

Edward Derbyshire

Contents

19.1 Introduction

Fine atmospheric dust (including fine mineral aggregates, fibrous minerals, and fibrous organic materials) reaches concentrations in many parts of the world sufficient to constitute a major influence upon both human and animal health.

The visible effect of dust in the atmosphere has been noted in written records since at least 1150 BC in China, since ancient times in the Mediterranean, and over the North Atlantic to leeward of the Sahara since at least the eighteenth century. Written records of Saharan dust falls in western Europe became increasingly common from the mid-nineteenth century. Following the "dust bowl years" of the 1930s, awareness of soil-derived atmospheric dusts increased considerably in the United States, particularly after 1945. Understanding of the complex role of atmospheric dust as a factor influencing climate and climatic change has made notable progress in the past 20 years, although there is still much to learn (Houghton et al. 2001). In contrast, the impact of high concentrations of natural dust on human and animal health has received relatively little attention when compared to work on artificially generated particulates, smoke and gases.

Aerosols include gases, liquids, and solid particles suspended in the atmosphere for varying lengths of time. Solid aerosols include particles injected into the atmosphere, such as mineral dust and sea salt, and those that form within the atmosphere, notably sulfates. Natural and man-made fires, including extensive burning of vegetation, generate smoke plumes that are often carried several thousands of kilometers from their sources, a sequence that contributes to regional air pollution and adds to atmospheric health hazards. Biomass burning yields black carbon which, together with mineral dust, is monitored by ultraviolet and other sensors on Earth-orbiting satellites. This provides increasingly detailed information on the incidence and seasonality of aerosol plumes over both land and water surfaces. Emphasis here is given to the release, transportation, and

E. Derbyshire (✉)
Royal Holloway, University of London, Egham, Surrey, UK
e-mail: ed4gsl@sky.com

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_19,

deposition of mineral particulate aerosols derived from soils, sediments, and weathered rock surfaces and their impact on human health when in suspension in the atmosphere. The finer components (<10 μm) of respirable natural atmospheric dusts include single particles, aggregates of very fine mineral grains (notably silica), fibrous minerals (e.g., the asbestos group), and fibrous organic materials.

Wind-borne dusts may affect human health by way of direct and indirect pathways. The elemental composition of dust, both when airborne and when accumulated on a land surface to form loess, can enhance the toxicity of the air breathed as well as that of the soil and the waters that drain through it. Inhaled dusts derived from fine-grained sediment sources such as seasonally dry rivers and dry lakebeds make up the direct pathway (Fig. 19.1).

A significant indirect pathway arises from generation of respirable mineral dusts by both natural and human-induced re-working (erosion) of loess and loessic soils. A second indirect pathway (Fig. 19.1), involving transfer and concentration of some toxic minerals by groundwater movement through thick loess accumulations, is not considered here. Some accumulations of mineral aerosols, including varying amounts of fine volcanic ash (tephra), contribute to the minerogenic dust in the atmosphere, which often endows surface sediments and soils with a distinctive mineralogy and chemistry (Sect. 19.2.2). Large volumes of ash and dust have been emitted since 1995 during eruptions of the Soufrière Hills volcano on the island of Montserrat in the eastern Caribbean. This volcano is rich in andesite (ca. 57–63 wt% silica), its explosive eruptions yielding abundant crystalline silica including metastable cristobalite (Horwell et al. 2010). The finer dusts may contain up to 24% of cristobalite, which poses a potential health threat to local populations during periods of prolonged eruption.

The detachment of mineral dust from the ground surface and its entrainment and subsequent transport by the wind are functions of several variables, which include the wind speed (both mean regional wind speed and the critical wind speed or threshold velocity required to dislodge particles), the degree of instability of the atmosphere, the size of the particles, the roughness and moisture content of the land surface, and the degree of particle exposure.

Source environments of mineral aerosol dusts are diverse, and some dust take-up by the atmosphere (the process of deflation) is a natural phenomenon that occurs at some time in most terrestrial environments. However, certain types of landscape, notably the sparsely vegetated terrains characteristic of the world's drylands, are particularly susceptible to the massive deflation that accounts for most of the atmospheric dust plumes thought to have a bearing on human and animal health. Seasonally deposited fine water-lain sediments (notably the finer grades in the silt and clay range carried in typically turbid glacial melt waters: Fig. 19.2), actively aggrading alluvial fan deposits, and fine lake sediments exposed in extensive basins by climatic desiccation are important examples of terrain types serving as atmospheric dust sources. Mineral particles are released by a variety of surface processes grouped together under the general heading of "weathering." These processes include breaking up of rock surfaces by the action of frost, salt, and chemical reactions and the biochemical complex of processes involved in soil formation; the latter accounts for the presence in some aerosol dust of plant fibers, phytoliths (biogenic opal), pollens, and spores. The silt-rich wind-lain deposits known as loess, which accumulated to great thickness after about 2.5 million years ago in Eurasia and the Americas, and particularly in central and eastern Asia (Derbyshire 2001; Derbyshire et al. 2000), are readily eroded in certain circumstances, thus constituting a secondary source of minerogenic atmospheric dust (Fig. 19.3).

Silt-sized particles, especially those in the ~10–50 μm range, are readily entrained by the wind from dry, unvegetated surfaces, but the clay-size (<2 μm) component of soils and sediments is not easily detached by the wind as individual particles because of the high interparticle cohesive forces typical of such colloidal materials. Entrainment of material finer than 2 μm usually occurs in association with the coarser (silt-sized) grains, and also in the form of coarse or medium silt-sized aggregates made up of variable mixtures of fine silt and clay-grade particles (Fig. 19.4). Critical wind speeds for dust entrainment (threshold velocities) vary notably; those for the semi-arid/subhumid, silt-covered terrains of northern China are approximately twice those required to initiate dust storms in the Sahara (Wang et al. 2000).

Silicon, making up more than one-quarter of the elements in the Earth's crust, is highly reactive and readily combines with oxygen to form crystalline silica (SiO_2), the most common form of which is quartz. SiO_2 dominates the composition of dust from North Africa (60.95%) and China (60.26%). These values closely match the world mean (59.9%) and its average content in the rocks of the Earth's crust (58.98%). Although less common, some other silica polymorphs, including cristobalite, are more pathogenic than quartz. Silicon also combines with other elements in addition to oxygen to form the dominant mineral group known as the silicate family, which includes the group of fibrous amphibole minerals grouped together under the general term asbestos. In the finest (<2 μm) fractions of many dryland surface sediments and soils, quartz is an important, and sometimes a dominant mineral, ranging in type from lithic fragments to biogenic opal. Varying amounts of clay minerals (hydrous aluminous phyllosilicates) are also common (notably kaolinite, illite, chlorite, vermiculite, smectite, and several mixed layer clays), with varying amounts of calcite, gypsum, and iron compounds. Amorphous (non-crystalline) silica, found in

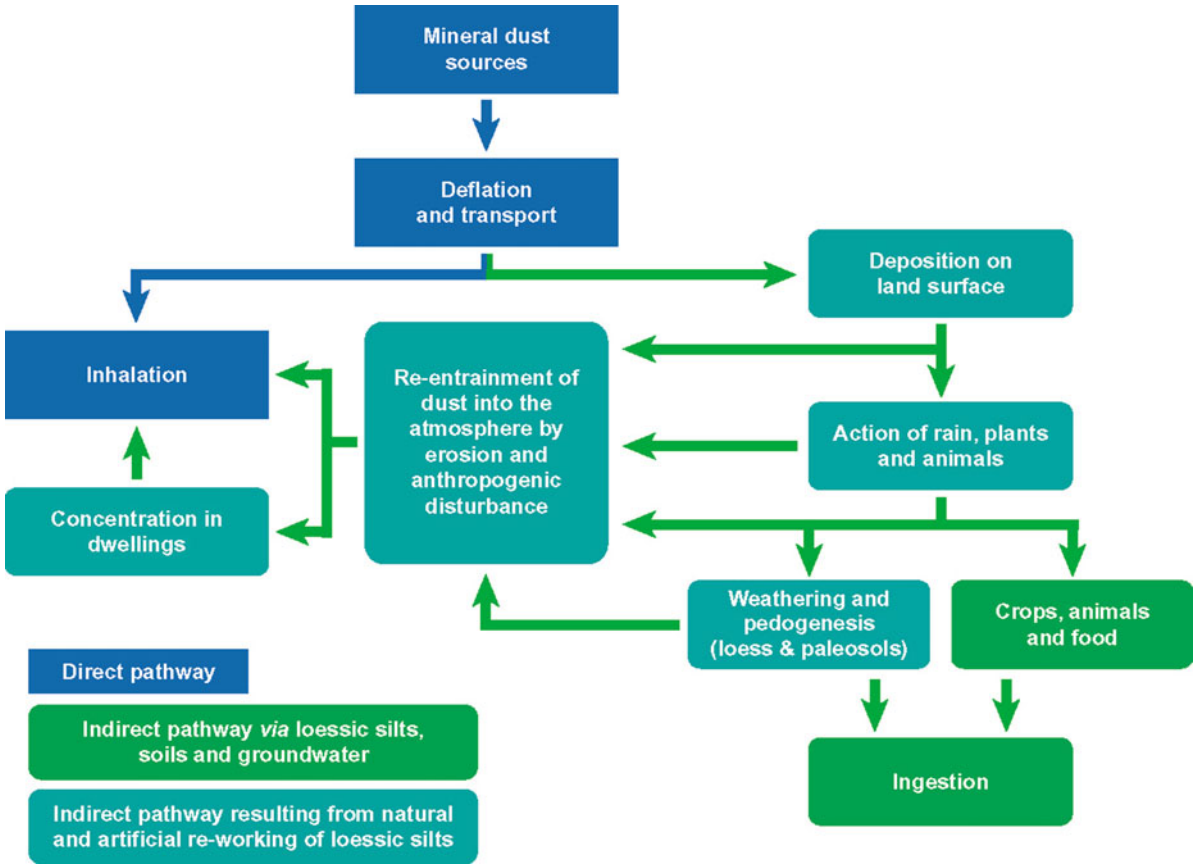


Fig. 19.1 Some essential links in the direct and indirect pathways from dust sources to human inhalation and ingestion (Note: The wind-lain and variously weathered sediment known as loess commonly contains buried soils (paleosols) marking phases of relatively stable former land surfaces during the accumulation of the loess)

Fig. 19.2 Dust storm in the upper Hunza valley, Karakoram mountains, northern Pakistan, summer 1980. The thick pall is mixed fine sand and silts carried by a cold, dense, gravity-enhanced airflow (katabatic wind) from the 59 km long Batura Glacier (not visible from this viewpoint). Such glacially induced winds are frequent in summer in the dry mountains of High Asia, deflating the finer components of extensive, dried-out meltwater deposits that accumulate around glacier margins

phytoliths, diatoms and testate amoebas, is generally believed to be benign on inhalation. However, given that the sponge spicules found in soils are only 5–3 μm in diameter and so conform to the definition of 'hazardous mineral fiber' (Skinner et al. 1988), some types of amorphous silica may prove to be a respirable hazard (Clarke 2003).

Fig. 19.3 Bare, eroded slopes in thick (>200 m) loess of the subhumid Xining Basin, northwestern China

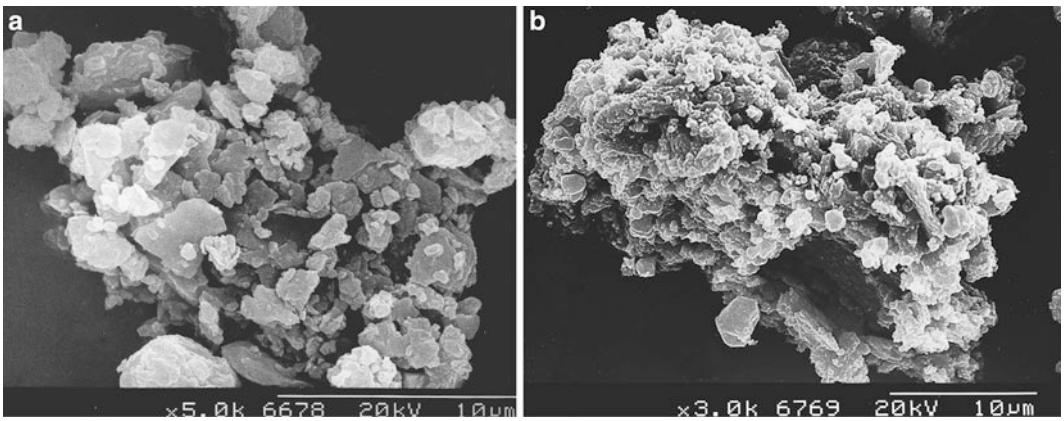


Fig. 19.4 Scanning electron micrographs of windblown dust aggregates. (**a**) Silt size aggregate made up of clay-grade mineral particles, as commonly found in the young (*Last Glacial*) loess deposits of central and eastern Asia. (**b**) Silt size aggregate taken from the dust on a house beam in Ladakh. Elemental composition of such beam dust is dominated by silica, with lesser amounts of oxygen, aluminum, sulfur, potassium, calcium, and iron. Both scale bars = 10 µm

The natural process by which substantial volumes of mineral dust are injected into the atmosphere is usually periodic, sometimes strongly seasonal, and mainly located in the subtropical arid and semi-arid regions. These potential dust sources cover about 30% of the total land area of the Earth. The dominant dust source regions lie in the northern hemisphere continents and include the subtropical and temperate deserts stretching from the Sahara of North Africa, through the Middle East and the northwest of the Indian subcontinent, and into central and eastern Asia. More modest sources of atmospheric dust have been identified in the Great Basin (United States) and in the Southern Hemisphere (the Lake Eyre Basin, Australia; central and northern Argentina; and a small part of southern Africa).

The relative contribution to global dust palls from these dominant source regions may have changed over recent geological time. For example, some sedimentary records from both the continents and the oceans indicate that rates of minerogenic dust accumulation during the last glacial maximum (about 20,000 years ago) were up to ten times greater than at present (Kohfeld and Harrison 2001). Such variable rates reflect changes in the location and size of dust source regions as glaciers waxed and waned, changing wind regimes (especially those associated with the monsoons) and

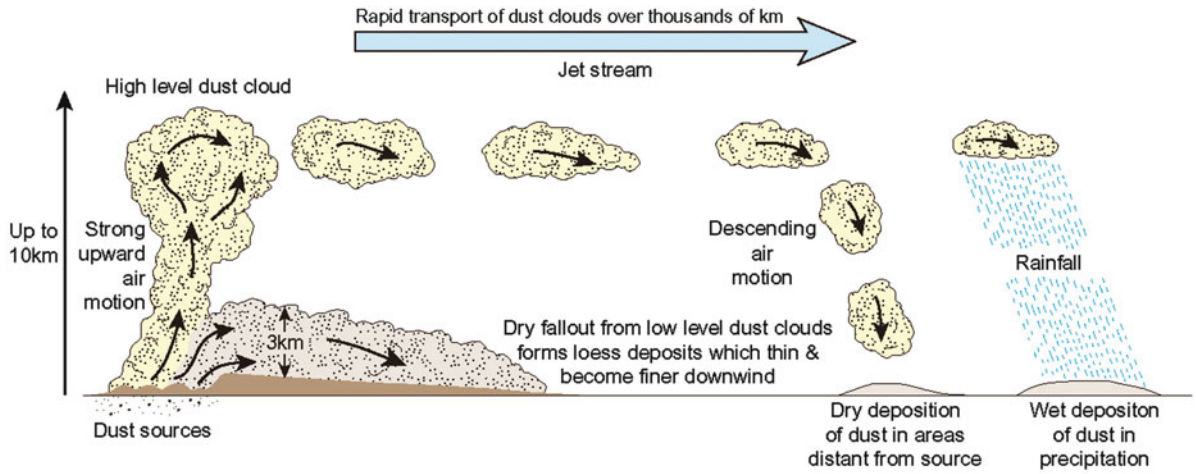


Fig. 19.5 Sketch showing the two principal modes of aeolian dust transport and deposition, based on a transect from the Chinese drylands to the Loess Plateau and the North Pacific Ocean (Re-drafted from Pye and Zhou 1989)

climatically driven fluctuations in the hydrological cycle that affected surface conditions including vegetation cover. Considerable contrasts in dust accumulation, expressed as calculated dust fluxes (mass accumulation rates in g/m^2/year, e.g., Derbyshire 2003), are beginning to emerge from studies of the Earth's loess deposits.

The process of entrainment and transport of mineral dusts varies from a local to a global scale. It is important to discriminate between source-proximal and source-distal dust plumes (Fig. 19.5). In general, the size of particles entrained by the wind declines with transport distance. As a result, the proportion consisting of the respirable fractions (commonly regarded as <10 μm) makes up an increasing proportion of the dust plume with distance from the source, although the absolute mass of the respirable fraction is greatest close to the source, as suggested by the colloquial term "desert lung" to describe pneumoconiosis in North Africa and the Middle East. High atmospheric dust concentrations show considerable variety in terms of their periodicity (from days to decades) and extent and proximity to sources, as well as in the percentage of the dust in the respirable range. In some regions of the world, extensive dust plumes have become an integral part of regional culture, and the washing out of the brown-to-red mineral particles by precipitation is known as "loess rain" in China, and "blood rain" in Mediterranean Europe and further north. These terms also draw attention to the contrast between yellow Chinese dust and the red dust of North Africa.

The concentration of mineral dust in the atmosphere (the atmospheric aerosol loading) is both a function of, and a factor influencing climatic change, as it affects physical and biogeochemical exchanges between atmosphere, land, and ocean. The presence of aerosols influences the chemistry of the troposphere including the proportion of ozone. Aspects of climate affected by atmospheric dust loading include the ability of dust to raise or lower air temperatures depending upon the differential effect of its particle size and chemistry, and upon the extent to which solar radiation is absorbed and scattered, which is an effect significantly modified by the amount, altitude, and thickness of any cloud cover. Deposition of dust may add notable volumes of certain nutrients to the world's oceans, including nitrates, ammonia, phosphates, and oxides of potassium and iron. It is considered that such inputs of iron to oceanic waters stimulate nitrogen fixation by plankton, thus enhancing productivity. The Sahara, commonly regarded as the world's greatest source of wind-transported mineral dust (Goudie and Middleton 2001), has some influence on the nutrient dynamics and biogeochemical cycles of a region stretching from northern Europe to South America (Prospero 1999) (Fig. 19.6). In addition Saharan dust storms are known to transport bacteria and fungal spores that cause deterioration in Caribbean coral reefs (Shinn et al. 2000), events that have also been linked to reduced air quality and cases of asthma and other respiratory problems in residents of parts of the southeastern United States. Bacteria are also present in dust from East Asian sources reaching Korea and Japan. Hara and Zhang (2012) showed that total bacterial cells in dusty conditions were one or two orders of magnitude higher than in non-dusty periods, while the viability of bacteria in dusty conditions was $<40\%$, this value fell to 76% in clean air, suggesting that long-range transport is an efficient means of dispersing both viable and non-viable bacteria up to the global scale.

The global dust cycle is integrally involved in Earth's energy and carbon cycles (Kohfeld and Tegen 2007; Shao et al. 2011) (Fig. 19.7). Such an intimate relationship between aerosols and the global environment, taken together with the probability that human actions in the past century or so have progressively enhanced the atmospheric dust loading, has implications for future climatic change (Harrison et al. 2001). The effect of such changes upon human societies is likely to include some notable health impacts.

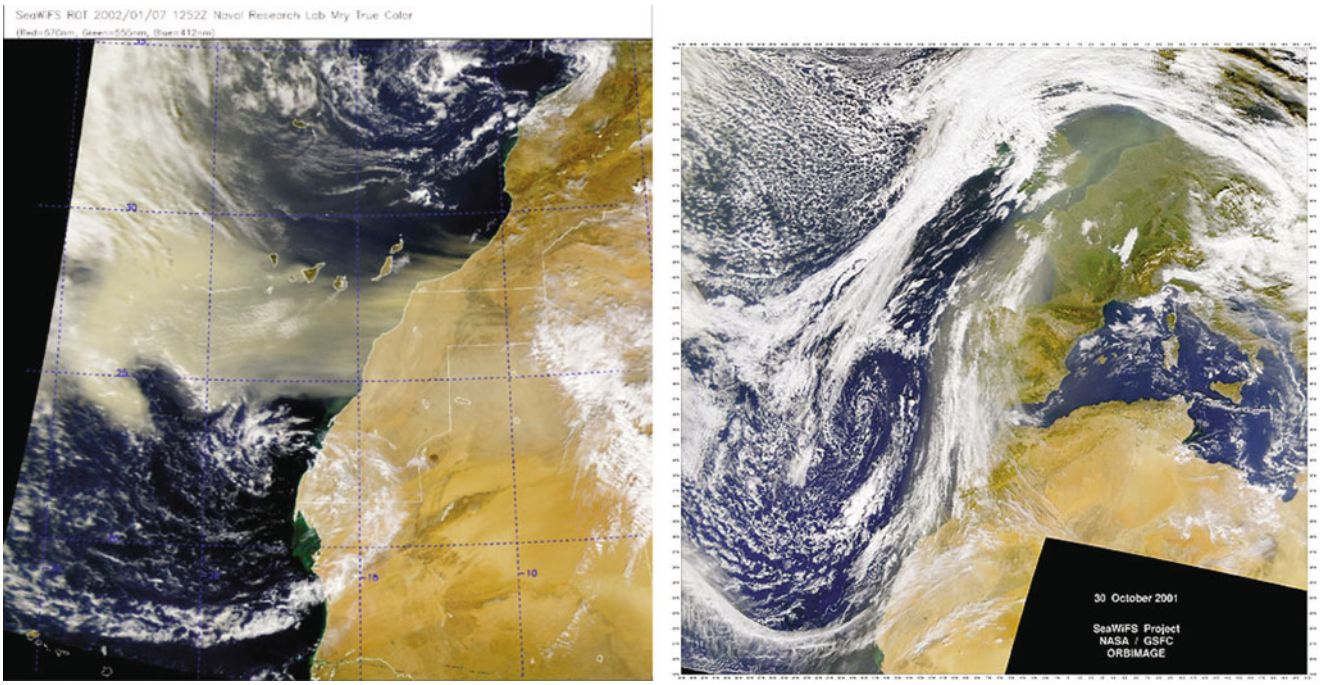


Fig. 19.6 (*Left*): Outbreak of Saharan dust across the North Atlantic, January 2002. High pressure over and north of Morocco, with a depression centered off the West Sahara-Mauritania coast, indicates a strong easterly flow across the African coast and the Canary islands (*center*) and into the mid-Atlantic. A dense pall of dust, about 500 km wide, reduced visibility and enhanced sunset colors for several days in the southern Canaries and deposited red dust. (*Right*): African dust over western Europe, October 2001. High pressure over the Mediterranean basin and an extensive and vigorous depression west of North Africa and Spain (centered close to the island of Madeira) induced strong southerlies from Mauritania to Scandinavia. A high-level dust pall can be clearly seen running from off the Moroccan coast across western Iberia, the Bay of Biscay, western and central France, southern and central Great Britain, the Low Countries, North Germany, and Denmark. Both are NASA SeaWiFS images

19.2 Health-Impacting Minerogenic Aerosols

19.2.1 Dust Storms

The type, size, and extent of dust plumes raised during dust storm events are fundamental factors influencing the degree to which naturally occurring atmospheric dust impacts upon the health of human and animal populations. Dust storms may be generated by local vortexes, generally known as "dust devils" (or willy-willies in Australia). Dust devils are only a few meters in diameter and raise dust to heights of 100–200 m (exceptionally 1,000 m) for periods of a few minutes to a few hours. Similarly low altitude, though more regionally extensive dust-carrying wind systems in northwest Africa, for example, arise from the relatively shallow northeasterly trade winds and by squall lines associated with northward incursions of equatorial air (the West African summer monsoon). These raise dust above the shallow trades and into the troposphere so that it crosses the Atlantic in winter as the dry northeasterly "harmattan." Such source-proximal transport involves a relatively high percentage of the coarser dusts (medium and coarse silts with some very fine sand), which are usually deposited at distances of only hundreds of kilometers downwind. Extensive regional atmospheric turbulence, however, arising from air mass frontal systems associated with the hemispherical wind regimes, notably the upper westerlies, carry the finer dust fractions at high levels within the troposphere. These source-distal events frequently transport terrestrial dust across oceans, including the Atlantic and Pacific, with deposition occurring some 2 or 3 weeks after initial entrainment (Pye 1987).

An example of the relationship between landforms, surface sediments, and soils that are particularly susceptible to dust deflation and specific meteorological situations, on the one hand, and atmospheric dust loadings and their source-proximal and source-distal effects, on the other, are illustrated by a dust storm that occurred in northwest China in May 1993 (Derbyshire et al. 1998). The highest wind velocities and most severe damage and loss of human life were felt in the Hexi Corridor (Fig. 19.8), a WNW-ESE topographical constriction between the mountains bordering the northeastern edge of the Tibetan Plateau (the Qilian Shan; in Chinese *shan* means mountains) and the Mongolian

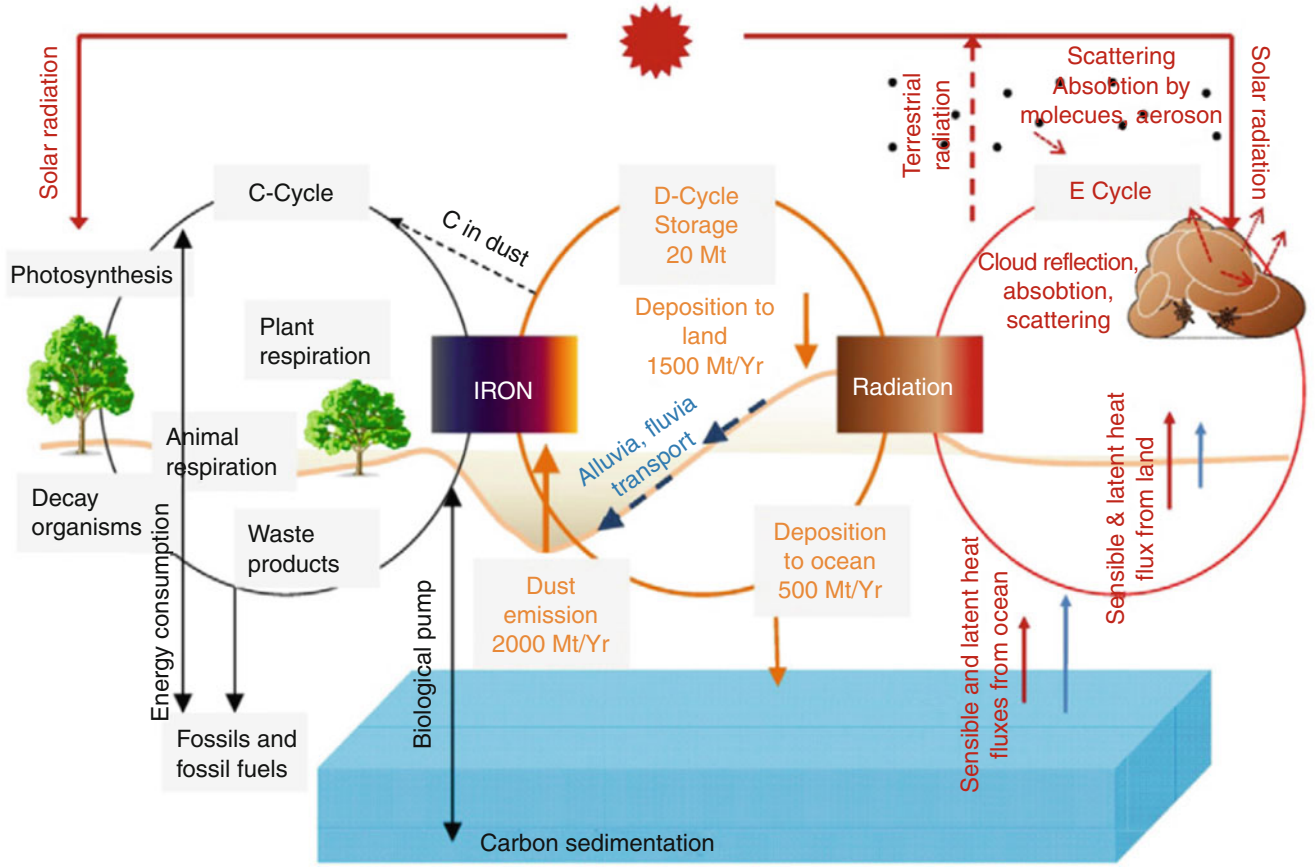


Fig. 19.7 The dust cycle (D-Cycle) in relation to the carbon cycle (C-Cycle) and the energy cycle (E-Cycle) within the Earth's system (From Shao et al. 2011)

Plateau. The wider impact was felt in the provinces of Gansu, Ningxia, Inner Mongolia, Shaanxi, and Hebei, a region equal to the combined area of France and Spain.

The meteorological situation on May 4, 1993, was controlled by a large high-pressure system (the Siberian High) over western Eurasia, with a depression centered on the northern Urals but with a trough extending far to the south. The cold front associated with this trough extended to the northern edge of the Tian Shan range. The constriction between the Tian Shan and the Altai Shan ranges resulted in increasingly convergent, and hence accelerating, air flow toward the east. By the next day, the cold front had reached Dunhuang (Fig. 19.9), where the channeling effect of the western end of the Hexi Corridor sustained a wind velocity of 12 ms. In 3 h the cold front had reached Jiuquan, with velocities of 17 ms, and within 5 h it had reached the narrowest part of the Corridor (between Zhangye and Wuwei) where velocities peaked at 34 ms (Fig. 19.9). These high velocities were sustained across the Tengger Desert, mobilizing sand as well as dust as far as Beiyin. With the opening out eastward of the Hexi Corridor, however, airflow became increasingly divergent and progressively slower; only the finer dust fractions were transported beyond the North China plain.

The effect of this "dark storm" in the proximal area of the Hexi Corridor was extremely serious. Visibility declined below 10 m in full daylight, and the depressed temperatures created severe frosts (minima −6.6 °C) with some local snowfalls. The direct effects included 380 people and 120,000 farm animals killed and damage to about 3,300 km^2 of crops. The particulate aerosol concentrations reached the "extensive dust pall" category (see below, Sect. 19.2.4) in the center and east of the Hexi Corridor, and there was widespread loess rain in the more distal provinces to the east (Shaanxi and Hebei). The coarser suspension load in the lower atmosphere, including coarse silts, reached as far as the northern slopes of the 3,700 m high Qinling Shan, south of the city of Xi'an. Five storms of similar magnitude occurred in the Hexi Corridor between 1952 and 1993. A satellite image of a dust storm that affected the Hexi Corridor and a broad region to the east of it on March 29, 2002, is shown as Fig. 19.10. Comparison of this image with the regional details of the 1993 event (Figs. 19.8 and 19.9) shows it to be very similar in source and extent, if not in destructive power.

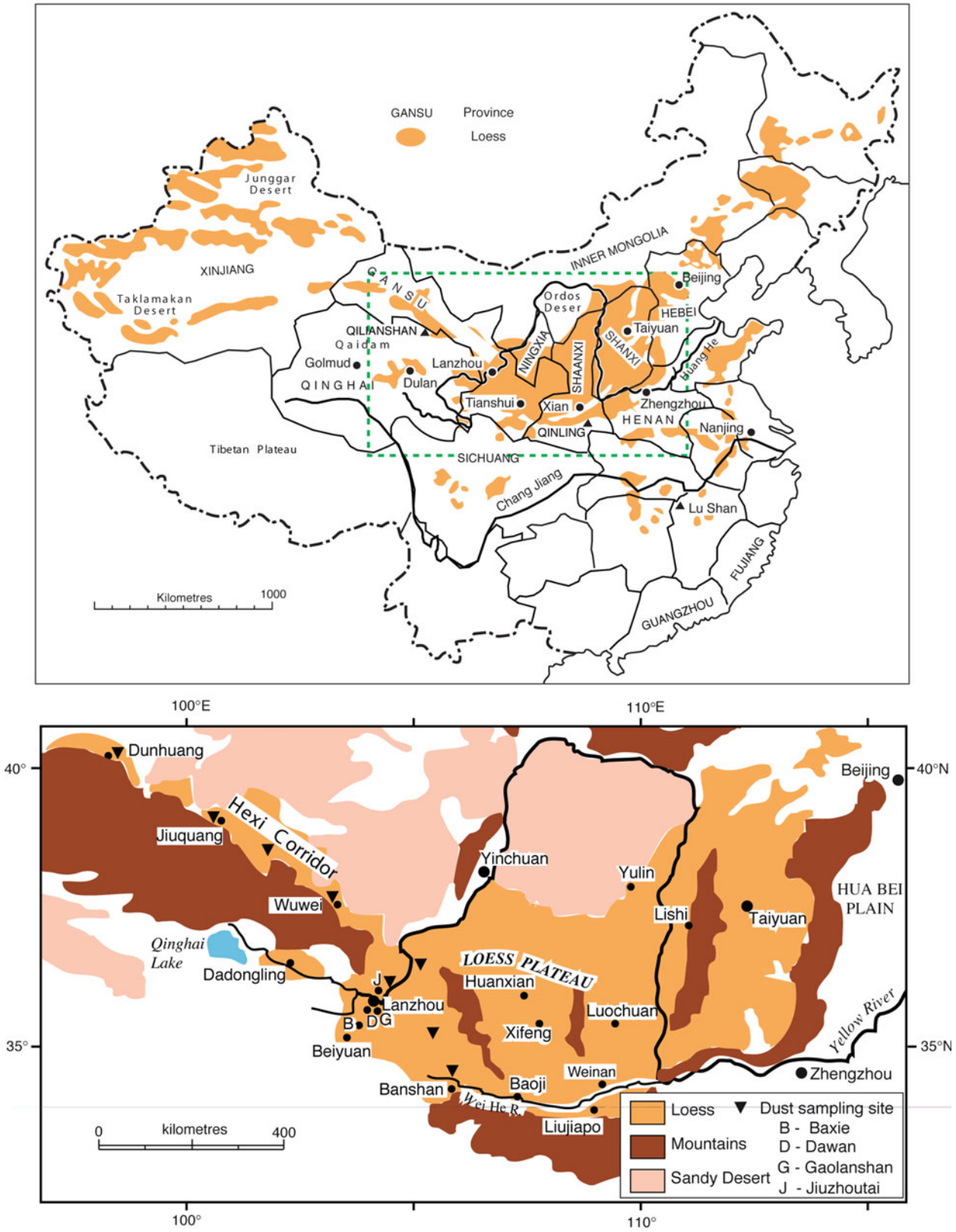


Fig. 19.8 *Upper*: The Loess Plateau of North China in relation to the Hwang He (*Yellow River*) and the principal deserts. The *box* indicates area covered in lower half of the figure. *Lower*: Part of northern China, showing the Loess Plateau and the Hexi Corridor. See text (Re-drawn from Derbyshire et al. 1998)

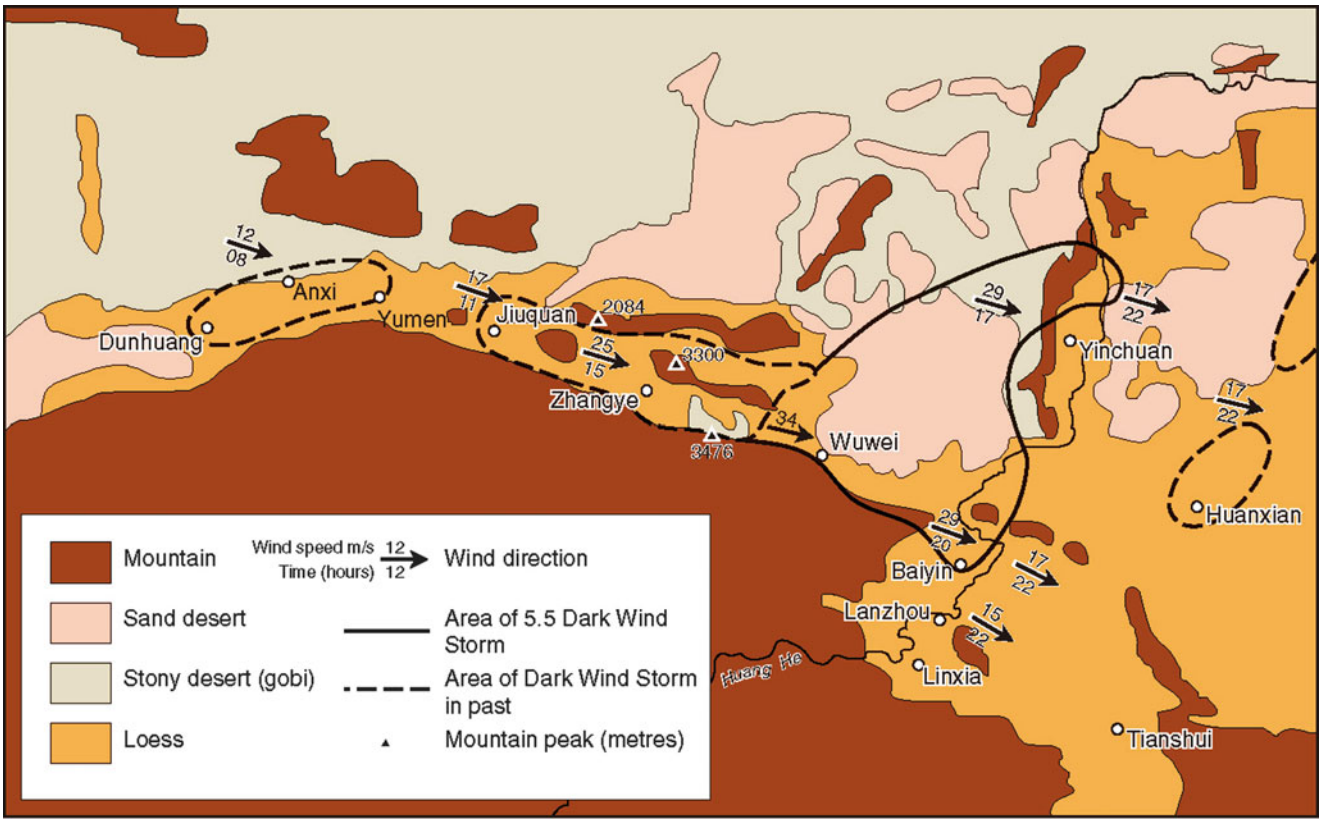


Fig. 19.9 Terrain sediment cover types, dust storm zones, wind velocities, and timing (in hours) of the "dark storm" of May 5, 1993, in the Hexi Corridor and the western Loess Plateau, China. Data provided by the State Meteorological Service of China, and the Meteorological Bureau of Gansu Province (Re-drawn from Derbyshire et al. 1998)

This Hexi Corridor case study is an example of dust impact from a proximal source with the bulk of the visible dust pall consisting of relatively coarse silt particles in the lower few kilometers of the atmosphere. However, the finer components of such palls, traveling at higher levels, are known to be carried great distances across the Pacific Ocean. Such "distal source–high altitude–finer dust" systems have aroused considerable recent interest because they can be tracked using orbital imagery backed up by study of synoptic meteorological charts. One such example of a distal dust source of global significance is the Tarim Basin, a region that probably has the highest mean annual number of "dust days" in China. The Taklimakan Desert, occupying most of the Tarim Basin, is predominantly a sandy desert, but loess accumulations are found along extensive parts of its windward (southern and western) mountain rim, which shows that the Tarim also has a functioning "proximal, low altitude–coarser dust" system (Fig. 19.11).

Steep atmospheric pressure gradients associated with extensive Siberian–Mongolian ridges of high pressure, most notably between late winter and early summer, strengthen the easterlies around the southern flank of the seasonal high pressure cell over the Taklimakan Desert. This air flow is then subject to vigorous uplift as it comes up against the western Kunlun, the Pamir, and the western Tian Shan ranges, all of which have peaks around 7,000 m above sea level. Current opinion (e.g., Sun 2002a) is that the finest fractions of the Tarim dust cloud are uplifted into the upper troposphere to be carried northeastward across Outer Mongolia, eastern Siberia, and across the Pacific Ocean (Fig. 19.11). Fine Chinese airborne dust is commonly recorded in western North America, and it is known to reach the eastern United States from time to time. This airborne dust was recorded over the Atlantic Ocean on April 20, 2001, and has recently been discovered in Greenland ice cores (Bory et al. 2003) and in the French Alps, a distance from source of about 20,000 km (Grousset et al. 2003). Preservation of mineral dust from both of the principal Chinese source regions in the Greenland ice cores and, in 2007, an upper-troposphere dust pall with a source in the Tarim Basin (Taklimakan Desert, western China) that was traced along more than a full circuit of the globe (~450° meridian) in 13 days, with a dust transport mass flux decay of an order of magnitude from 75 to 8 Gg (gigagrams: Uno et al. 2009), confirms that the "distal–high level–fine dust" pathways are persistent, of global scale and probably long-established.

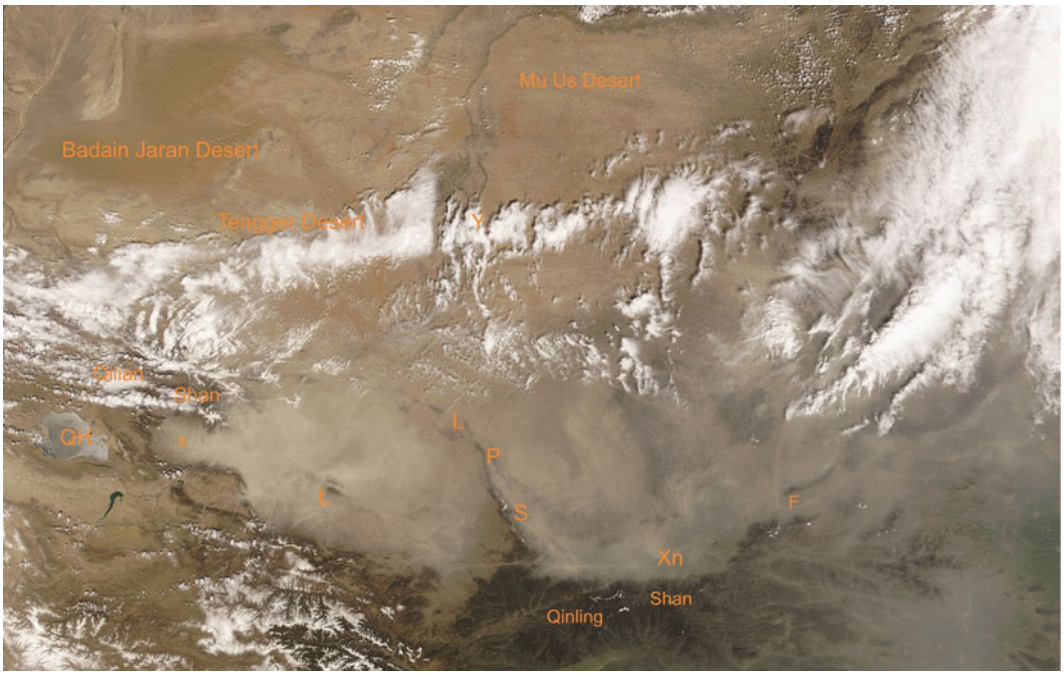


Fig. 19.10 Part of a Terra Satellite image, using the MODIS sensor, taken on March 29, 2002, and showing a dust storm generated by winds from the west-northwest over northwest China. (For comparative location, compare with Figs. 19.8 and 19.9). The air over the Badain Jaran and Tengger deserts is clear of dust, as it is over the Mu Us Desert (to the north of which can be seen the big bend of the Yellow River). However, the dust thickens rapidly with proximity to the cold-weather-front and the alluvial-fan-covered Hexi Corridor on the northern side of the Qilian Shan. The Xining Basin (X = city of Xining, just to the east of Qinghai Lake; QH) generates its own pulse of dust, but not the adjacent Qaidam or Gonghe basins in this case. Lanzhou city (L), near the outlet of the Hexi Corridor, has a thick dust pall over it as well as its own locally generated pollution cloud. The dense dust plume is split by the NNW-SSE aligned Liupan Shan (L-P-S). East of this mountain range, the plume completely covers the twin basins of the Jing and Luo rivers (draining the central and southern part of the Loess Plateau). This part of the plume just covers the city of Xi'an (Xn); its pollution pall is more modest than that of Lanzhou. The southeastern margin of the plume is very sharp as it comes up against the ~3,700 m high Qinling Shan (on which several snow-covered areas can be seen). The plume extends eastward, crossing the sharp bend of the Yellow River at Fenglingdu (75 km west of the Sanmenxia Reservoir on the lower Yellow River). The dust plume over the green farmlands of Henan and Hebei provinces in the southeastern part of the image is much more diffuse, which suggests that it may be the product of a pre-frontal trough

Atmospheric dust from diverse sources may converge on a single location. For example, Lee et al. (2010) present evidence indicating that influx of mineral dust (based on Al, Ca and Fe) over Hong Kong during 76 dust events (most in the winter half year) during the period 1996–2007 could be traced, in addition to the two main regional sources (Mongolian and Taklimakan deserts), to the Thar (NW India), central-west Asian, Arabian and Saharan deserts. Analytical results using TOMS, NASA's Ozone Monitoring Instrument (OMI), US Navy Aerosol Analysis and Prediction System (NAAPS) and backward trajectories using the NOAA HYSPLIT model (Fig. 19.12) show the presence in 76 events of particles from the following locations: Mongolia 68 and Taklimakan 30, with non-East Asian sources numbering 73, of which the Thar number was 65, the Sahara 63 and Arabia 62. Estimated mean dust concentration of non-East Asia PM_{10} was ~10%, with a Thar-Arabian-Saharan contribution of ~8%.

19.2.2 Dust Sources

Interest in the detection, tracking, and measurement of distal (regional and global scale) mineral dust in the atmosphere has been greatly stimulated by the increasing availability of images and other data provided by Earth-orbiting satellites. Aerosol optical thickness (AOT), as estimated using the advanced very high resolution radiometer (AVHRR) of the United States' National Oceanic and Atmospheric Administration (NOAA), is based on backscatter radiation measurements made at an effective wavelength of 0.63 μm. In general, high AOT values indicate high atmospheric dust concentrations. However, because the AOT algorithm requires the surface below the dust plume to have a low and constant albedo, AOT can be estimated in this way only over the oceans. This restriction is generally true of satellite sensors operating in the visible spectrum. The situation was greatly improved by the advent, in 1980, of the total ozone

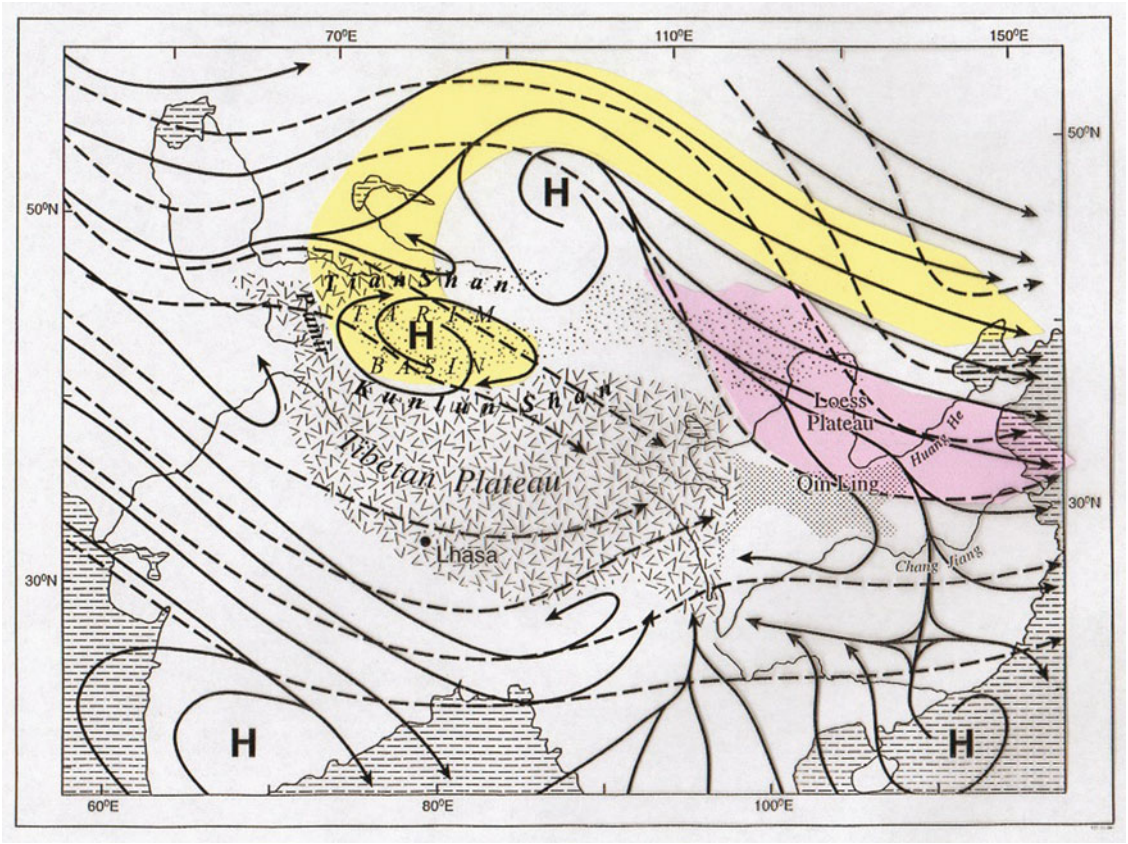


Fig. 19.11 Relationship between generalized winter-spring pressure systems over central and eastern Asia, the major orographic features, the two dominant dust source regions, and the associated principal atmospheric dust pathways. The deserts, alluvial fans, and dry lake basins to the north and west of the Loess Plateau provide both source-proximal, coarser silts to the Loess Plateau, with much finer material being carried at higher atmospheric levels across the east China plain, and the Pacific Ocean and beyond (*pink arrow*). The Tarim Basin surrounded by high mountains (the Tian Shan, the Pamir, and the Kunlun Shan) concentrates fine sediments largely derived from glacial meltwater rivers and alluvial fans. These are re-worked by winds blowing from north of east to be deposited as loessic silts on the northern flanks of the Kunlun Shan. At over 5,000 m above sea level, this may be the highest loess in the world (From Sun 2002b). Finer dusts are lifted above this level and enter the westerly jet stream to be carried great distances, sometimes as far as Europe and beyond (*yellow arrow*)

mapping spectrometer (TOMS). This is used to detect absorbing aerosols based on the spectral contrast at 340 and 380 nm in the upwelling ultraviolet (UV) spectrum. TOMS is sensitive to a range of UV-absorbing aerosols such as mineral dust, volcanic ash, and black carbon from fossil-fuel combustion sources and biomass burning. The UV surface reflectivity is typically low and nearly constant over both land and water, which allows TOMS to detect aerosols over both continents and oceans. The UV spectral contrast is used in a non-quantitative way as an absorbing aerosol index (AAI). The temporal and spatial variability of this TOMS AAI has been matched to types of absorbing aerosols, as well as to known sources such as individual volcanic eruptions, forest fires, and large-scale dust events. The global distribution of the occurrence frequency of relatively high TOMS AAI values (January and July 1980–1992) is shown in Fig. 19.13 (Prospero et al. 2002) (see also Chap. 29, this volume). This image contains one huge, dominant area with high AOT values extending westward of the North African coast and eastward from the Middle East, indicative of dust plumes from the world's premier atmospheric dust source region. The plume off the west coast of South Africa, in contrast, is attributed to biomass burning.

The irregular timing and the variety of sources contributing to dust storms, as well as the technical limitations of the different sensors in use, complicate the determination of the location and extent of individual dust source regions or areas around the globe. The common association of dust-raising conditions with the cloudy conditions generated by pressure troughs and air-mass fronts is a case in point. The TOMS system is most sensitive to aerosols in the middle and upper troposphere and above (distal dust) and least sensitive in the boundary layer where aerosol residence times are shorter (proximal dust). Aerosols below the altitude range of 1,000–1,500 m remain largely undetected. Thus, the source-proximal components of major

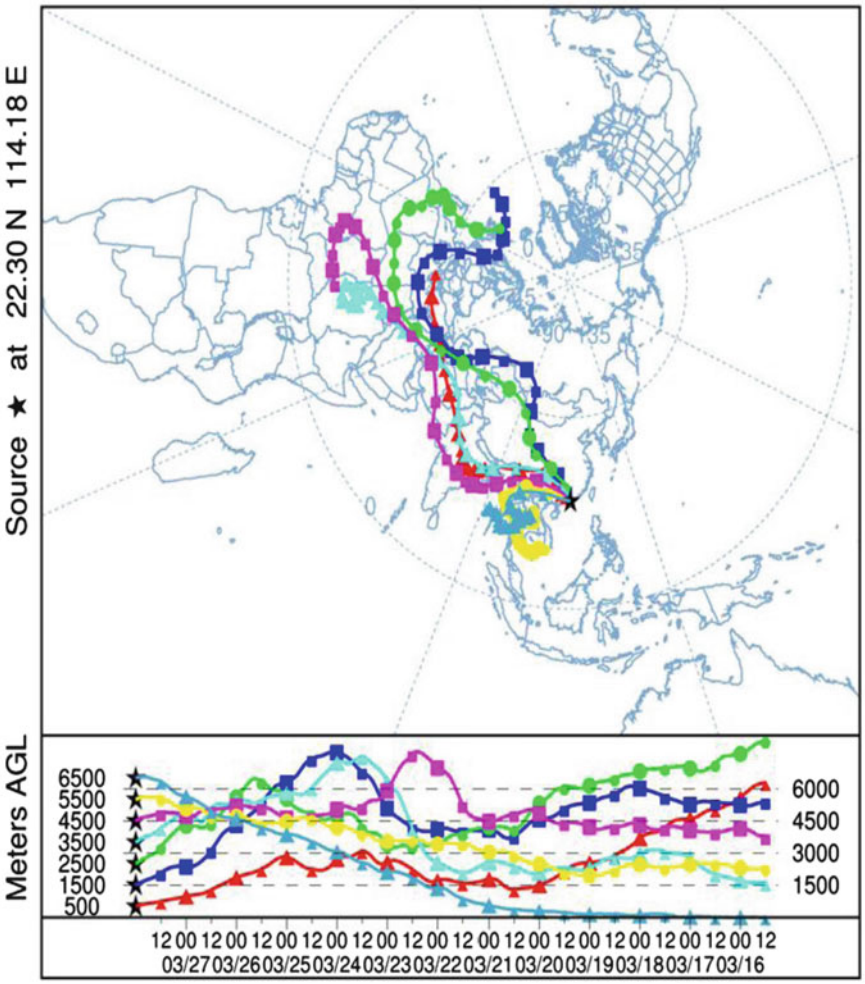


Fig. 19.12 NOAH HYSPLIT model with backward trajectories at 0000 UTC 28 March 2000. Dust of northern China provenance tends to have lower altitude trajectories (barometric level ~900 hPa), with higher arrival trajectories (barometric level ~700–500 hPa) associated with the longer distance sources, Saharan, Arabian and Thar (From Lee et al. 2010)

destructive dust storms may not be detected on some visual imagery.

Despite such difficulties, the use of global distributions based on the month in the year that best represents the long-term (13 year) frequency of dust storm occurrence as indicated by TOMS AAI has yielded a map of major global dust sources that closely matches the information available from other types of observation (Prospero et al. 2002). The result (Fig. 19.14) indicates sources in all continents except Europe and Antarctica. Most of the major sources are in surface depressions or adjacent to mountain fronts.

For example, the Ahaggar and Tibesti mountains in the Sahara are surrounded by what may be the greatest single regional dust source on Earth. There is a strong link between dust sources and extensive alluvial deposits, as well as ephemeral, saline, and dried out lakes throughout North Africa, the Middle East, the northwestern Indian subcontinent, Middle Asia (from the Caspian Sea to Kazakhstan), and across northwest China, as shown above (Sect. 19.2.1). Sand dune deserts, as such, are not important consistent sources, although their sporadic drainage systems frequently provide abundant fine particles for deflation. There are many smaller, but important, sources outside these major regions. These include the Basin and Range province of the southwestern United States and northern Mexico; the Lake Eyre and Great Artesian Basin in Australia; Patagonia, the Andean footslopes in central Argentina, and the Altiplano of Bolivia and northern Argentina; and southern Africa. Potential secondary sources also include the major loess deposits, notably in parts of northern China. The "mountain deserts" of Iran and Pakistan, at the western end of the Himalayan tract, constitute an important regional dust source, notably in summer, which involves channeling of dust by down-valley (katabatic) winds (Fig. 19.2).

The human impact, varying in both type and intensity of activity as well as in length of its history, further complicates assessment of "natural" dust sources. The major concentrations of fine-grained, poorly vegetated deposits that constitute important dust sources are those associated with floodplains, alluvial fans and lake depressions, and sites that are fed by seasonal perennial freshwater flows that also attract human communities and

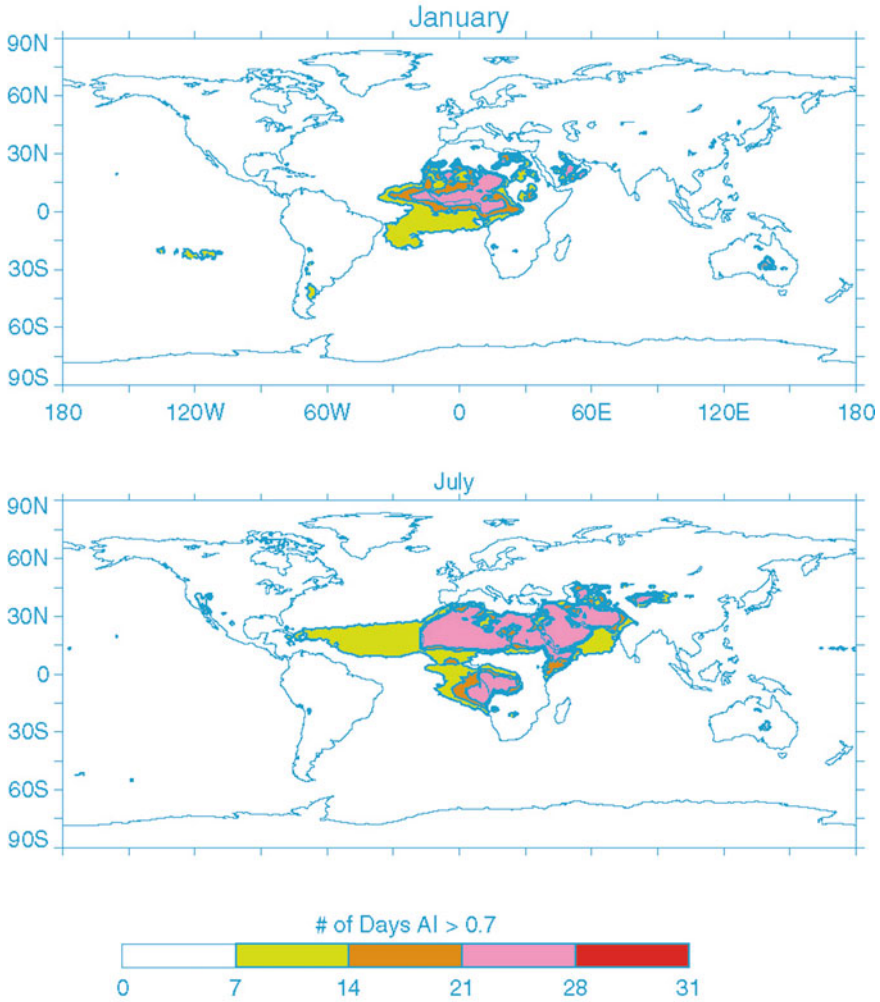


Fig. 19.13 Global distribution of dust and smoke. Monthly frequency of TOMS absorbing aerosol product for January (*top*) and July from 1980 to 1992. Scale: number of days per month when the AAI equaled or exceeded 0.7. The large, dark area in southern Africa in July is a product of biomass burning, and there is also evidence of biomass burning in January just north of the Equator in Africa. Part of the plume over the Equatorial Atlantic is smoke. All other distributions shown are due to the presence of dust (After Prospero et al. 2002, with kind permission of the first author)

their animals. The dust sources in the Middle East include the Tigris-Euphrates basin where agriculture has been widespread on this rich alluvium for thousands of years. There is also a long history of human use and interference with such water sources in the drylands of central Asia. Many small states and cities in western China have collapsed as a result of a failure of water supply due either to overuse or destruction of dams in warfare as well as severe periodic drought (Derbyshire et al. 2000). The present-day use of dung, wood, and, to a lesser extent, coal, as fuel sources in the drylands of western China is an important factor affecting the extent and composition of airborne dust.

The colonization, by sophisticated agricultural people, of the Loess Plateau of northern China, a mass of wind-deposited silt with an area >400,000 km^2 and an average thickness of 100 m, has had a notable impact. Locally dense populations practicing hand agriculture and their grazing animals have played a major role in accelerating river erosion and slope failure. Some commentators take the view that the Loess Plateau and the Mongolian steppe lands to the north of it can be regarded as a secondary source of atmospheric dust in present climatic conditions (Fig. 19.15), although this view is not universally held (see Sect. 19.2.3).

Northern India, another region with a dense human population, injects the products of the burning of dung, wood, and fossil fuels into the atmosphere to an extent that makes it difficult to estimate the natural component in atmospheric dust palls. The definition of the main dust sources is also complicated by widespread biomass burning in central and southern Africa. Land degradation arising from resettlement of quite large populations since the middle of the twentieth century, and subsequent land clearing, agriculture, and/or animal grazing, has created new sources of dust in the Etosha Pan in Namibia (southern Africa; Bryant 2003). This is also true of the Mongolian Plateau in eastern Asia. The recent history of human interference by damming rivers draining into the Caspian and Aral seas provides another such example. The diminution of the Aral Sea has become a classic case of the human generation of an atmospheric dust source, but smaller examples are known, such as Owens Lake in California in the United States (see Sect. 19.3.2).

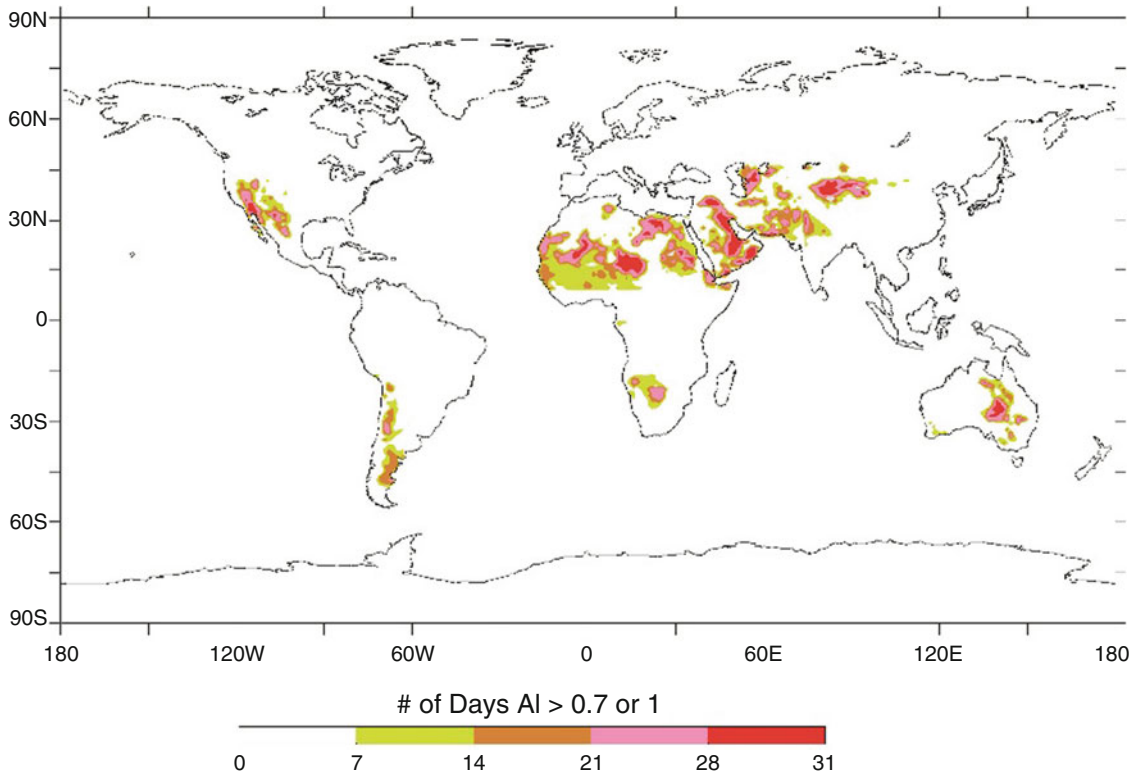


Fig. 19.14 Global distribution of TOMS dust sources. This is a composite of selected monthly mean TOMS AAI frequency of occurrence distributions for specific regions using those months that best illustrate the configuration of specific dust sources. The distributions were computed using a threshold of 1.0 in the "global dust belt" (west African Saharan coast, through the Middle East, and central Asia to the Yellow Sea), and 0.7 elsewhere (After Prospero et al. 2002, with kind permission of the first author)

A number of studies have attempted to quantify global dust emissions, but estimates vary substantially, which reflects the evidently wide gap between modeling-based estimates and those based on the numerous available (but short-term) measurements; for selected global and regional dust flux values see Engelbrecht and Derbyshire 2010 and Shao et al. 2011. Equally diverse are estimates of the effect of land use and land use changes on the global atmospheric dust loading, which may be less than 20% (Prospero et al. 2002).

19.2.3 Dust Storm Frequencies

Dust storm frequency is usually measured by the annual number of "dust storm days," defined as a reduction of visibility by dust to less than 1 km for all or part of a day. Allowing for the fact that the number of ground observations is low, the location of measuring sites is extremely uneven, and the measurement record very short for many sites, the range of dust storm days per year values is wide with figures reaching about 80 in southwest Asia, >60 in Turkmenistan and the Karakum, >45 in Kazakhstan, >30 in the Tarim Basin, the Hexi Corridor, and the Loess Plateau (China), ~30 in parts of North Africa, ~20 in the Northwestern Indian subcontinent, and >15 in central Australia (Middleton et al. 1986).

Measured dust storm frequencies are open to different interpretations, however, even in some densely populated regions of the world. For example, the rise in frequency in North China in the second half of the twentieth century referred to above (Fig. 19.15) is currently under review in the context of global warming. It is argued that, while increased desertification might be expected to result in more frequent dust storms, the recent decline in the number of dust storms in the Beijing region, indicated by some records, is consistent with the current warming trend, although the effects on total seasonal dust volumes of any such decline may be offset by an increase in the vigor of individual dust events. Certainly, higher mean temperatures over North China would be consistent with a generally weaker winter-spring monsoon and fewer outbreaks of cold, dense air from the northwest. Thus, although there is some observational evidence suggesting a negative correlation between the number of spring dust storms and mean temperatures over this large region, the case for decline is inconclusive. This may reflect the complex cause-and effect-relationships involved in dust storm analysis. Another more

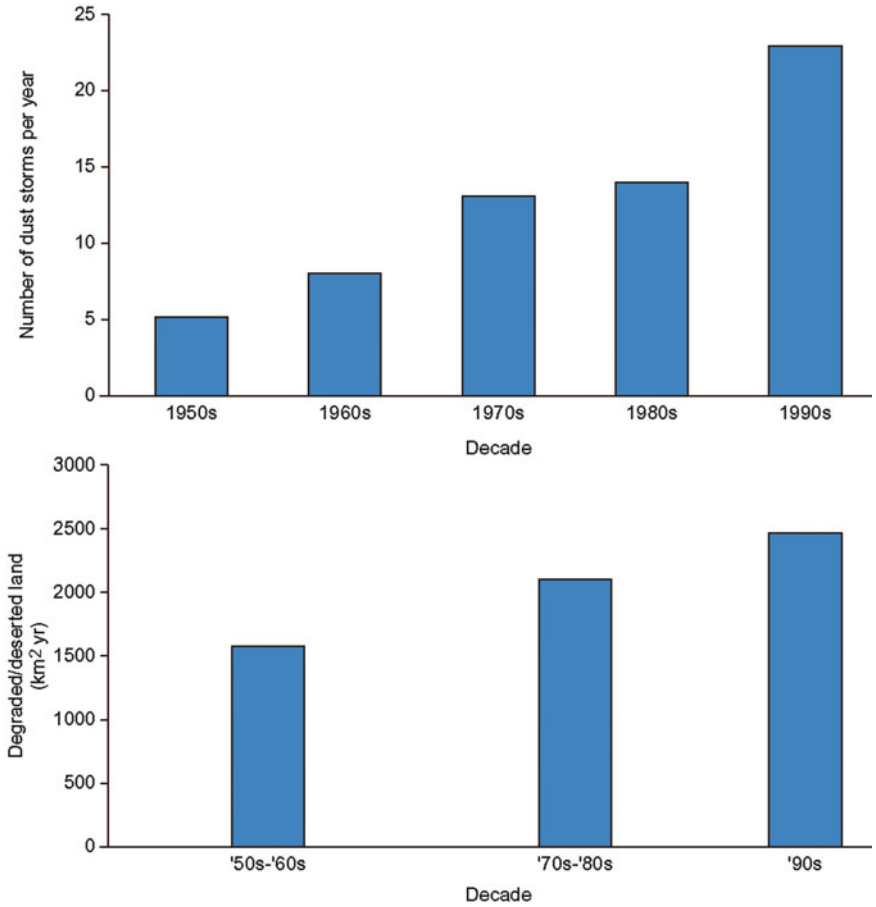


Fig. 19.15 Graphs showing (*top*) the number of dust storms per year by decade, from the 1950s to the 1990s, in Beijing and (*bottom*) the annual rate of land degradation (in km^2 year) in Inner Mongolia (lying northwest (windward) of Beijing). Data kindly supplied by Dr. Xingmin Meng from a Chinese-language Web site

skeptical point of view is that governmental and public concern about dust storms is now much greater than it was in the social climate of the 1950s, 1960s, and 1970s.

19.2.4 Ambient Dust: Continental Concentrations

Dispersion of mineral aerosols during and following dust storms yields ambient atmospheric dust in the form of sets of discrete plumes and extensive regional sheets or palls. Discrete plumes, thought to derive from point sources, give rise to very high dust concentrations that include the coarsest particle fractions; such plumes persist for relatively short periods (hours to days). With increasing distance from source, atmospheric dust disperses as regional palls of finer particles in which the respirable fractions are dominant; these may persist for several days or longer.

Numerous measurements of ambient dust concentrations have been made, although most cover short periods only. In addition, the terrestrial sites concerned are very unevenly distributed. Both of these facts make it difficult to assess with any precision the terrestrial "dust climate" at a regional scale, as exemplified by the current controversial situation in North China mentioned above. Direct and regular measurement of dust concentration is rare or absent in many parts of the world. Most national meteorological services measure or estimate visibility. However, because visibility is a composite product of both atmospheric humidity and concentration of aerosols (both natural and anthropogenic), visibility estimates provide only a rough guide to dust concentrations (see below). Moreover, although satellite observations are very informative with respect to sources and transport paths in the troposphere, many dust events are not detected in this way (Sect. 19.2.2). Further, the common assumption of global uniformity in the optical properties of dust is manifestly unrealistic. In regions in which frequent high concentrations of minerogenic dust constitute a health hazard for large human populations, such as northern China, existing data gathering is clearly inadequate, as indicated in the recent call within China for a nationwide state-owned network for the monitoring and analysis of atmospheric dust. Many of the existing measurements of atmospheric dust in continental locations are orders of magnitude greater than those obtained from measurements over the oceans. Maximum values of mineral dust cited in the literature for continental areas are around 10^5 μg m^{-3} (Pye 1987), although concentrations may occasionally exceed this value in some parts of the world. The "normal background"

lower atmosphere dust concentration in northwest China is 0.083 mg m^{-3}, which reaches values of about 4 mg m^{-3} in "ordinary" dust storms (Table 19.1). However, concentrations of 69 mg m^{-3} (April 1998) and 21.61 mg m^{-3} (April 2003) have recently been recorded, and an extreme value of 1,016 mg m^{-3} occurred in May 1993 (Derbyshire et al. 1998).

Human settlement of continental drylands has undoubtedly served to enhance the frequency, magnitude, and impact on health of dust-entraining events and ambient dust levels. Activities such as arable farming, intensive grazing, industry, urbanization, and road and rail construction are frequently concentrated within natural dust source environments such as alluvial fans, river floodplains and terraces, and lake-basin margins. Loess and silts deposited by rivers and lakes provide some of the most fertile and readily cultivated soils on Earth, as well as a widely used building material (known in parts of the Americas as adobe). Recent extension of agricultural activities along desert margins in several continents, and notably in Asia, has caused varying degrees of land degradation (often generalized as desertification). Hand cultivation and shallow plowing of such deposits certainly stimulates local dust palls, and dust concentrations in many adobe dwellings are often some orders of magnitude higher than those found in normal background conditions in regions such as Ladakh and northern China. In such situations, it is often difficult to discriminate between natural and anthropogenic dust, and so to attribute with any assurance human health effects exclusively to natural versus "occupational" dust events.

Table 19.1 Categories of dust concentration used to describe Chinese dust storm events

Dust concentration (mg m^3)	Description
0.083	Normal background
0.356	"Detachment mode"
1.206	Extensive dust pall
3.955	Ordinary dust storm

From Chinese Meteorological Bureau

19.3 Pathological Effects of Aerosol Dust

Inhalation of mineral aerosol particles, followed by deposition in human pulmonary alveoli, varies with a number of factors, but particle size and composition and certain lung functions are particularly important. Most coarse particles in minerogenic dust (diameter <100 μm) are abundant close to the dust sources, and they are deposited relatively quickly by both dry and wet depositional processes. Thick dust palls characteristic of locations relatively close to dust sources pose a number of hazards to human health and welfare, which include transport accidents, destruction of crops, and eye irritation. When inhaled, many of the larger dust particles are eventually rejected by expectoration. However, inhalation of large dust particles (>10 μm) may constitute a health risk if the mineralogy is toxic, regardless of where the grains lodge in the respiratory system. Of the finer dust particles (diameter <10 μm) that remain in suspension in the atmosphere for much longer periods (the respirable fraction), most between 10 and 5 μm become trapped in the upper respiratory tract and are ultimately removed by coughing. Particles finer than 5 μm frequently penetrate more deeply into the lungs to cause silicosis (Pendergrass 1958), asbestosis, and other lung conditions. Recent studies suggest that about 75% of dust found in some Chinese post-mortem lung tissue is finer than 3 μm. Atmospheric dust finer than 2.5 μm is considered to be of particular importance with respect to community health, as in the PM standard of the United States' Environmental Protection Agency.[1] Ambient dust may also absorb harmful gases, disease-generating bacteria, and even carcinogenic hydrocarbon compounds. Recent work in China, for example, has shown that the denser the ambient dust, the higher the rates of chronic respiratory disease and associated death rates. Respiratory disease may also exacerbate cardiac problems (see also Chap. 25, this volume).

19.3.1 Pneumoconioses

The pneumoconioses, lung diseases that include silicosis and asbestosis, are a result of prolonged inhalation of fine minerogenic aerosol dust. The condition is best documented from studies of workers in certain industries in which high mineral concentrations are generated (occupational pneumoconiosis). Much less attention has been accorded to cases of pneumoconiosis arising from non-occupational exposure to ambient mineral dust; one recent exception is the "cleanup" campaign at Libby, Montana in the United States (see below). Studies based on occupational cases show both pneumoconioses to be insidious in the early stages but then progressively noticeable when exercising, thus the symptoms are sometimes attributed to a patient's aging (Wagner 1997). Many symptoms are nonspecific in the absence of radiography, and this may constitute an important factor influencing diagnosis in some developing countries. Radiographic diagnosis of silicosis is made with confidence only after the appearance of silicotic nodules 2–5 mm in size. Continued dust exposure leads to an increase in nodular size and number so that they eventually cover much of the lung, and the nodules sometimes coalesce to form conglomerate shadows often called progressive massive fibrosis (Saiyed 1999).

[1] The PM (particulate matter) standard is based on the total mass of particles measuring 2.5 μm or less observed in a 24 h period.

Some types of inhaled particulates are degraded by macrophages, but many are highly resistant to this process and persist in the lung cavity and lymph nodes. Some resistant particulates appear to cause no problems, but others stimulate fibroblastic cells to deposit collagen. In the case of asbestos, for example, detectable fibrosis appears only after a threshold number of particles have been retained (Bar-Ziv and Goldberg 1974). Silica is a highly fibrogenic agent in lung tissue, and the reaction is very different from the granulomatous reaction to many other nondegradable grains; the fibrotic reaction has been associated with release of polysilicic acids. Fine-grained, sharply angular quartz grains are widely considered to enhance this process, although the precise nature of the pulmonary response to crystalline silica remains rather poorly known (Saiyed 1999). Continued exposure to silica is thought to lead to increased rates of infection with pulmonary tuberculosis, a notable public health problem in many developing countries, with non-tuberculosis mycobacterial infection (involving intercellular bacterial parasites) also occurring. Many people with silicosis have been shown to be susceptible to tuberculosis (Snider 1978), although present constraints on diagnosis, especially in poor and remote regions of the world, carry with them a continuing risk of confusing silicosis-related massive fibrosis with tuberculosis. Nevertheless, evidence exists that patients with silicosis carry a greater susceptibility to tuberculosis, and the World Health Organization (1997) has now listed crystalline silica as a human carcinogen. Silicosis has a number of deleterious effects upon the immune system. One important effect is a reduced ability of the macrophages to inhibit growth of tubercle bacilli. Some rheumatic, as well as chronic renal diseases also show higher than average incidence in individuals exposed to silica, and it is likely that such increased susceptibility of subjects to a suite of mycobacterial diseases is to some degree due to impaired function of macrophages in silicotic lungs (Snider 1978).

A number of pathological conditions are associated with inhalation of asbestiform minerals. Asbestos is found in a wide variety of geological environments. For example, chrysotile is known to occur in hydrothermally altered ultramafic or carbonate rocks. Crocidolite is abundant in some metamorphosed iron formations, and may also occur as an authigenic mineral and as a hydrothermal alteration product in some carbonatite complexes. Although not covered by the term asbestos, the mineral erionite, a fibrous zeolite (group of hydrous aluminosilicates), is also known to cause asbestosis and related conditions as noted below (Sect. 19.3.3). Natural release of asbestiform minerals from the host rock occurs by the processes of weathering and erosion, and the fibers are frequently concentrated by surface wash. In seasonally dry climates, such concentrations of fibers dry out and become susceptible to deflation. The health effects of asbestos inhalation include asbestosis, mesothelioma, and lung cancer. Some asbestos fibers penetrate body tissue and remain in the lungs, lung lining, and abdominal cavity. Radiographically visible fibrosis may take 15–20 years to appear following initial exposure (Wagner 1997) (see also Chap. 24 of this volume and Derbyshire et al. 2012).

19.3.2 Case Studies of Non-occupational Silicosis

Non-industrial deposition of silica in human lung tissue was first reported in three inhabitants of the Sahara Desert half a century ago. The autopsy results showed a high content of fine (<3 μm) silica dust, but there was no sign of typical silicotic lesions (Policard and Collet 1952). Other findings from different parts of North Africa include radiological evidence of multiple micronodules in reticular disposition scattered throughout the lungs, and this is considered to be consistent with silicosis. A radiographic survey of 18 asymptomatic Bedouin females in the Negev Desert by Hirsch et al. (1974) found positive indicators in all patients who were aged between 26 and 70 years, with nine cases in people older than 50. Histological examination showed varying amounts of dust-laden macrophages in a perivascular and peribronchial distribution, mainly in the middle and lower lungs. No typical silicotic nodules with collagenization were found, and a 3–5 year radiological follow-up study showed no evidence of progression. The fact that older patients made up the largest proportion of the sample group, and the lack of any progression toward formation of fibrotic conglomeration, suggested a benign condition called simple siliceous pneumoconiosis, which was possibly attributable to long periods of work within the confined environment of the home tent. In a larger study involving 54 cases, siliceous particles were found both free and in macrophages, and the incidence of fibrosis was shown to be age related with a progression more noted in women (13 out of 22) than in men (only 4 out of 32) (Bar-Ziv and Goldberg 1974).

Such “desert lung syndrome” has a long history and was even found in ancient Egyptian mummies (Tapp et al. 1975). In the past half century it has also been recorded in Pakistani farmers, Californian farm workers, Ladakh villagers, people in the Thar Desert of Rajasthan, northwest India, and residents of northern China as well as the cases mentioned above from the greater Saharan region.

A survey of two villages situated at altitudes between 3,200 and 3,500 m in the western Himalayas, situated only 15 km from Leh (the capital city of Ladakh), was undertaken by Norboo et al. (1991). There are no mines or industries in

Ladakh, but dust storms are frequent between late winter and summer and there is characteristic local variability in their incidence in this high mountain environment. Radiographic evidence was taken from an equal number (23) of men and women between the ages of 50 and 62 years. Of these, 8 men and 16 women showed varying grades of silicosis, with important differences resulting from the higher dust concentrations at the lower village (at 3,200 m) compared to the higher one (at 3,500 m). Three cases of progressive massive fibrosis were found in the lower village, which suggested the likelihood that silicosis here causes appreciable morbidity. Later augmentation with necropsy lung tissue samples revealed heavy dust deposition with abundant hard, 1- to 3 mm nodules and a lymph node largely replaced by hyaline collagenous nodules, a classic feature of silicosis. More than 20% of the mineral dust extracted from the lung tissue consisted of quartz, bulk chemical analyses yielding 54% elemental silica, and 19.2% aluminum. A larger study (Saiyed et al. 1991), involving a total of 449 patients aged over 50 years (245 women, 204 men) from three different villages in the Leh vicinity, showed typical cases of pneumoconiosis associated with progressive massive fibrosis and egg-shell calcification of the bronchial glands (indicative of high concentrations of free silica) in 101 cases (22.5% of the population sampled). A close correlation was found between frequency of dust storms and number of cases of pneumoconiosis: the village with low dust storm frequency recorded only a 2.0% incidence, the one with moderate frequency showed that 20.1% of the sample population was affected, and the village with severe dust storm incidence revealed a 45.3% incidence of pneumoconiosis. The existence of such a high proportion of pneumoconiosis cases in the populations of such remote villages, with no possibility of exposure to the products of mining or industrial activity, is striking evidence of the role of minerogenic dust. Although there is little doubt that the burning of brushwood and dung in the adobe dwellings places the women at higher risk of developing pneumoconiosis, as clearly indicated by Norboo et al. (1991), Saiyed et al. (1991) found no clear differences in incidence between the sexes. Mineral dust found on the upper surfaces of wooden roof beams in houses built of loess in Ladakh is all finer than 15 μm, more than 25% by weight being <1 μm; the silica content is $>60\%$ (Fig. 19.4b).

Several studies, both published and unpublished, have implied that the large number of people subject to the frequent dust storms characteristic of north China are potential silicosis cases. In one investigation 395 subjects (294 men and 101 women) from two communes in the middle of the Hexi Corridor, Gansu Province, were studied (mean dust concentration: 8.25–22.0 mg m^3), and 88 people (46 men and 42 women) were randomly chosen from a third commune with low dust storm exposure as a control group (mean dust concentration: 1.06–2.25 mg m^3). The incidence of silicosis was 7.09%, with no cases in the control group, but this rose to 21% in subjects over 40 years of age. There was no significant difference in incidence between the sexes, but comparative necropsy of the lungs of camels showed them to contain evidence of silicosis (Xu et al. 1993). In another, larger but unpublished radiological survey, involving 9591 residents in Gansu Province, a prevalence of 1.03% was found, which rose to 10% in subjects over 70 years old (Changqi Zou, personal communication).

In the past four decades, about half (36,000 km^2) of the former bed of the Aral Sea (western Uzbek Republic, middle Asia) has been exposed, providing a new and frequent source of fine dust (Fig. 19.16). This situation has been exacerbated by the diversion, for irrigation purposes, of most of the waters of the two main rivers (Syr Darya and Amu Darya) that drain into the Aral Sea, adding to the regional desiccation. The fine-grained sediments (silts and clays) on the seabed are rich in agricultural chemicals, and they are readily deflated. Despite some reports of increasing respiratory illness in children, including some mention of interstitial lung disease in this region, there is little authoritative information about the link between the desiccation of the Aral Sea region and human health (Wiggs et al. 2003).

Owens Lake (California in the United States) was a shallow but perennial water body for most of the last million years, but diversion of water for use by the city of Los Angeles began in 1913 and, by 1926, the lake had dried up. The dry lakebed is now probably the greatest single source of respirable mineral dust particles in the United States. Palls of dust <10 μm in aerodynamic diameter occur on about 10 days/year across a wide area. With arsenic levels sometimes as high as 400 ng m^3 within air samples (Reid et al. 1994), and tests indicating high solubility of the arsenic in simulated lung fluids (G. Plumlee, personal communication), these events are viewed as a health hazard for residents of Owens valley (Reheis 1997).

19.3.3 Case Studies of Non-industrial Asbestosis

Asbestosis arises from inhalation of asbestos fibers, although conclusive risk assessment has been hampered because the microscopically detectable fibers make up only an insignificant proportion of the total dust burden in lung tissue (Eitner 1988). Most studies of this interstitial lung disease have been concerned with the impact of asbestos on the health of workers in a wide range of occupations, including mining, manufacturing, and construction, as well as in users of the thousands of commercial products that contain asbestos primarily because of its insulating qualities. However, cases of non-occupational asbestosis have been reported in several

Fig. 19.16 Orbital image of the Aral Sea area taken on June 30, 2001, from Space Station Alpha (Earth Sciences and Image Analysis Laboratory, Johnson Space Center). A major dust storm can be seen, driven by strong westerly winds. The sharp northern margin of the dust pall coincides with the Syr Darya River. This is beyond the area of exposed sea floor sediments, where soil moisture and vegetation cover impede deflation

countries in Europe and around the Mediterranean, including Czechoslovakia, Austria, Bulgaria, Greece, and Turkey.

In central Turkey, inhalation of agricultural soils rich in tremolite (a common fibrous amphibole found in contact-metamorphosed impure calcareous rocks) and erionite (most commonly arising from alteration of volcanic rocks) is responsible for an endemic malignant pleural mesothelioma. Incidence of this disease is specific to certain villages around which the soils contain one or both of these minerals.

Incidence of pleural plaques, associated with mesothelioma, was also found in residents of northern Corsica who had no history of occupational contact with asbestos. The percentage of 1,721 subjects shown by radiographs to have bilateral pleural plaques was 3.7% for those born in north-eastern Corsica compared to only 1.2% for those born in the northwest. The rocks of the northeast are rich in serpentine, asbestos, and chrysotile, but this region is separated from the northwest part of the island by a mountain barrier. A clear excess of subjects with bilateral plaques born in villages close to asbestos outcrops was shown (94.6% for affected subjects born in the northeast compared to only 5.4% of subjects born in unexposed villages; Boutin et al. 1986). Preliminary data indicated high levels of chrysotile fibers in the atmosphere, suggesting that incidence of the disease arises directly from inhalation, in an environment in which asbestos exposures and patients with plaques are juxtaposed.

In 1999, public concern led to investigation of a vermiculite mine in Libby, Montana, USA, following its closure in 1990 after more than a century of operation. It had been found that the vermiculite, a micaceous mineral widely used in the insulation of buildings, was contaminated with the tremolite-actinolite form of asbestos. In the alkalic intrusive complex at Libby, the amphiboles are a product of hydrothermal alteration of pyroxenites which also occur as hydrothermal veins cutting across the igneous rocks (G. Plumlee, personal communication). Investigations designed to determine the extent of the impact upon human health arising from occupational links as well as any non-occupational effects arising from activities such as gardening and use of unpaved roads included testing of more than 7,000 people over 18 years old in the years 2000 and 2001. This involved interviews, medical history, chest x-ray, and lung function (spirometry) tests. The results showed that radiographic pleural and interstitial abnormalities were present in 51% of former mine workers. The risk of such abnormalities increased with age and with increasing length of residence in the Libby area. The odds of finding pleural abnormalities were stated to be 1.7–4.4 times greater (depending on age) in the case of former mine workers compared to residents with no mine connection, although the incidence of abnormalities in the latter group (3.8%) was higher than for groups within the United States with no known asbestos exposure (range 0.2–2.3%; United States Environmental Protection Agency 2003).

19.3.4 Tuberculosis

It has been suggested that the incidence of pulmonary tuberculosis in dryland environments may be linked to non-occupational silicosis. Sunlight and aridity are antipathetic to the tubercle bacilli and droplet transmission of pulmonary tuberculosis is favored by lack of sunlight, higher humidity, and overcrowding. However, data from the Thar Desert, India, presented by Mathur and Choudhary (1997), show a

prevalence of tuberculosis in desert areas of Rajasthan about 25% higher than the non-desert parts. The presence of radiographically determined evidence of non-occupational silicosis in the desert people offers some support for the hypothesis that silicosis may be an important factor in the higher prevalence of tuberculosis in this desert.

19.4 Conclusions

Aerosol mineral dusts affect human health as a result of inhalation and retention of the finest fractions derived directly from source sediments and indirectly from disturbance of surface layers of loess, a geological formation consisting primarily of wind-lain minerogenic dust. The geological and meteorological study of dust sources, sinks, transport, and geochemistry provides a foundation for improved understanding of the extent and magnitude of the impacts of natural minerogenic aerosols on human health.

The pathological effects of prolonged exposure to natural aerosol dust have been recognized in a general way since ancient times, but the number of modern studies of the pneumoconioses outside occupation-specific contexts remains relatively small. The specific health effects of direct inhalation of high concentrations of fine minerogenic dusts, generated by natural deflation from loose, poorly bound soil surfaces, including those exposed by accelerated erosion of weak geological formations such as loess, thus remain rather poorly known and relatively little researched. Knowledge of many of the suspected linkages involved is incomplete, and so is inferential to varying degrees. The magnitude of the world's population affected by inhalation of fine mineral aerosols can only be estimated, although it is likely to number millions of people in the middle latitude desert zone especially across Eurasia between the eastern Mediterranean and the Yellow Sea. Given the progressive improvement and sensitivity of remotely sensed information derived from the several types of orbiting satellite platforms, there is a need for greater investment in improved "ground truth" systems. Obtaining the necessary data on the nature of the "dust climate" and degrees of dust exposure in susceptible environments requires the application of appropriate geological and meteorological methods of monitoring and analyzing dust. This should include regular, standardized measurement and collection and analysis of dust concentrations in the lower atmosphere as a routine component of the meteorological observation systems already operated by most countries. Systematic research programs designed to quantify the respiratory health status of people in the same environments but with contrasting dust exposure potential, and taking full account of other risk factors including those of anthropogenic origin (occupational conditions, cigarette smoking, lifestyle, etc.), will be needed to complement the environmental monitoring.

The impact of trace elements on human health by way of the indirect pathway through soils and groundwater, as found in the loess and loess-like sedimentary accumulations and associated soil types within and adjacent to the great dryland zones of the world, has received much more attention than the direct and indirect pathways considered here, as shown elsewhere in this volume.

Finally, some account of way of life must be considered as a factor in any assessment of the health impact of respirable mineral dust because it directly affects dust generation, re-suspension, and inhalation in many of the world's drylands. Loess and loessic alluvium are abundant and easily applied to building materials used widely in Asia; the predominantly flat-roofed dwellings require only small amounts of wood to complete them. Traditionally, many such houses use small interior kitchens for cooking in winter, with open fires and some primitive chimneys. Although the situation is now slowly changing, domestic burning of dried cattle-dung, wood, and (rarely) low-grade coal is still common, for example, in Ladakh and the Hexi Corridor in northwestern China. To the smoky atmosphere in such confined environments is added fine, re-suspended loessic dust raised by sweeping the dried loess floors. This is mixed with varying concentrations of cigarette smoke. Such high dust concentrations in the home place females at relatively higher risk than males. This may be further enhanced by additional exposure to field dust in areas, such as the tributary valleys of the Indus (northernmost Pakistan), in which females also play a primary role in cultivating the fine silty soils. Such complexity renders the design of a set of strategies for amelioration, if not prevention, a formidable task. It is yet another, perhaps unsung, addition to the challenges posed particularly by many countries in the developing world, with a bearing on the lives of many tens of millions of people.

See Also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements) • Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 20 (The Ecology of Soil-Borne Human Pathogens) • Chapter 24 (Environmental Medicine) • Chapter 25 (Environmental Pathology)

Further Reading

Bar-Ziv J, Goldberg GM (1974) Simple siliceous pneumoconiosis in Negev Bedouins. Arch Environ Health 29:121

Bory AJM, Biscaye PE, Grousset FE (2003) Two distinct seasonal Asian source regions for mineral dust deposited in Greenland (NorthGRIP). Geophys Res Lett 30:1167. doi:10.1029/2002GL016446, 2003

Boutin C, Viallat JR, Steinbauer DG, Massey DG, Mouries JC (1986) Bilateral pleural plaques in Corsica: a non-occupational asbestos exposure marker. Eur J Resp Dis 69:4–9

Bryant RG (2003) Monitoring hydrological controls on dust emissions: preliminary observations from Etosha Pan, Namibia. Geogr J 169:131–141

Clarke J (2003) The occurrence and significance of biogenic opal in the regolith. Earth Sci Rev 60:175–194

Derbyshire E (2001) Geological hazards in loess terrain, with particular reference to the loess regions of China. Earth Sci Rev 54:231–260

Derbyshire E (ed) (2003) Loess and the dust indicators and records of terrestrial and marine palaeoenvironments (DIRTMAP) database. Quat Sci Rev 22(18–19):1813–1819

Derbyshire E, Meng XM, Kemp RA (1998) Provenance, transport and characteristics of modern Aeolian dust in western Gansu Province, China, and interpretation of the Quaternary loess record. J Arid Environ 39:497–516

Derbyshire E, Meng XM, Dijkstra TA (eds) (2000) Landslides in the thick loess terrain of northwest China. Wiley, New York

Derbyshire E, Horwell CJ, Jones TP, Tetley TD (2012) Airborne particles. In: Plant JA, Voulvoulis N, Ragnarsdottir KV (eds) Pollutants, human health and the environment: a risk based approach. Wiley-Blackwell, Chichester, pp 255–286, Chapter 10

Eitner F (1988) Substantiating hazardous exposure to asbestos by examination of pulmonary dust. Int Arch Occup Environ Health 61:163–166

Engelbrecht JP, Derbyshire E (2010) Airborne mineral dust. Elements 6:241–246

Goudie AS, Middleton NJ (2001) Saharan dust storms: nature and consequences. Earth Sci Rev 56:179–204

Grousset FE, Ginoux P, Bory A, Biscaye PE (2003) Case study of a Chinese dust plume reaching the French Alps. Geophys Res Lett 30:6. doi:10.1029/2002GL016833

Hara K, Zhang D (2012) Bacterial abundance and viability in long-range transported dust. Atmos Environ 47:20–25

Harrison SP, Kohfeld KE, Roelandt C, Claquin T (2001) The role of dust in climate changes today, at the last glacial maximum and in the future. Earth Sci Rev 54:43–80

Hirsch M, Bar-Ziv J, Lehmann E, Goldberg GM (1974) Simple siliceous pneumoconiosis of Bedouin females in the Negev desert. Clin Radiol 25:507–510

Horwell CJ, Le Blond JS, Michnowicz SAK, Cressey G (2010) Cristobalite in a rhyolitic lava dome: evolution of ash hazard. Bull Volcanol 72:249–253

Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (2001) Climate change 2001: the scientific basis. Cambridge University Press, Cambridge

Kohfeld KE, Harrison SP (2001) DIRTMAP: the geological record of dust. Earth Sci Rev 54:81–114

Kohfeld K, Tegen I (2007) Record of mineral aerosols and their role in the Earth system. Treatise Geochem 4:1–26

Lee YC, Yang X, Wenig M (2010) Transport of dusts from East Asian and non-East Asian sources to Hong Kong during dust storm related events 1996–2007. Atmos Environ 44:3728–3738

Mathur ML, Choudhary RC (1997) Desert lung in rural dwellers of the Thar Desert, India. J Arid Environ 35:559–562

Middleton NJ, Goudie AS, Wells GL (1986) The frequency and source areas of dust storms. In: Nickling WG (ed) Aeolian geomorphology. Allen & Unwin, New York

Norboo T, Angchuk PT, Yahya M, Kamat SR, Pooley FD, Corrin B, Kerr IH, Bruce N, Ball KP (1991) Silicosis in a Himalayan village population: role of environmental dust. Thorax 46:341–343

Pendergrass EP (1958) Silicosis and a few other pneumoconioses: observations on certain aspects of the problems with emphasis on the role of the radiologist. Am J Roentgenol 80:1–41

Policard A, Collet A (1952) Deposition of silicosis dust in the lungs of the inhabitants of the Saharan regions. Arch Indust Hyg Occupat Med 5:527–534

Prospero JM (1999) Long-term measurements of the transport of African mineral dust to the Southeastern United States: implications for regional air quality. J Geophys Res 104:(15)917–(15)927

Prospero JM, Ginoux P, Torres O, Nicholson SE, Gill TE (2002) Environmental characterization of global sources of atmospheric soil dust identified with the NIMBUS-7 TOMS absorbing aerosol product. Rev Geophys 40(1):1002. doi:10.1029/2000RG000095

Pye K (1987) Aeolian dust and dust deposits. Academic, London

Pye K, Zhou LP (1989) Late Pleistocene and Holocene Aeolian dust deposition in North China and the Northwest Pacific Ocean. Palaeogeogr Palaeoclimatol Palaeoecol 73:11–23

Reheis MC (1997) Dust deposition downwind of Owens (Dry) lake, 1991–1994: preliminary findings. J Geophys Res 102 (D22):25,999–26,008

Reid JS, Flocchini RG, Cahill TA, Ruth RS, Salgado DP (1994) Local meteorological, transport, and source aerosol characteristics of late autumn Owens Lake (dry) dust storms. Atmos Environ 28:1699–1706

Saiyed HN (1999) Silicosis–an uncommonly diagnosed common occupational disease. Indian Counc Med Res 29:1–17

Saiyed HN, Sharma YK, Sadhu HG, Norboo T, Patel PD, Patel TS, Venkaiah K, Kashyap SK (1991) Non-occupational pneumoconiosis at high altitude villages in central Ladakh. Br J Indust Med 48:825–829

Shao Y, Wyrwoll K-H, Chappell A, Huang J, Lin Z, McTainsh GH, Mikami M, Tanaka TY, Wang X, Yoon S (2011) Dust cycle: an emerging core theme in Earth system science. Aeolian Res 2:181–204

Shinn EA, Smith GW, Prospero JM, Betzer P, Hayes ML, Garrison V, Barber RT (2000) African dust and the demise of Caribbean coral reefs. Geophys Res Lett 27:3029–3032

Skinner HCW, Ross M, Frondel C (1988) Asbestos and other fibrous materials. Oxford University Press, New York

Snider DE (1978) The relationship between tuberculosis and silicosis. Am Rev Resp Dis 118:455–460

Sun J (2002a) Provenance of loess material and formation of loess deposits on the Chinese Loess Plateau. Earth Planet Sci Lett 203:845–859

Sun J (2002b) Source regions and formation of the loess sediments on the high mountain regions of Northwestern China. Quat Res 58:341–351

Tapp E, Curry A, Anfield C (1975) Sand pneumoconiosis in an Egyptian mummy. Br Med J 2(5965):276

United States Environmental Protection Agency (2003). http://www.epa.gov/region8/superfund/libby/lbybkgd.html

Uno I, Eguchi K, Yumimoto K, Takemura T, Shimizu A, Uematsu M, Liu Z, Wang Z, Hara Y, Sugimoto N (2009) Asian dust transported on full circuit around the globe. Nat Geosci 2:557–560

Wagner GR (1997) Asbestosis and silicosis. Lancet 349:1311–1315

Wang Z, Ueda H, Huang M (2000) A deflation module for use in modelling long-range transport of yellow sand over East Asia. J Geophys Res 105((D)):26,947–26,959

Wiggs GFS, O'Hara SI, Wegerdt J, van der Meers J, Small I, Hubbard R (2003) The dynamics and characteristics of Aeolian dust in dryland central Asia: possible impacts on human exposure and respiratory health in the Aral Sea basin. Geogr J 169:142–157

World Health Organization (1997) Monograph on the evaluation of carcinogenic risks to humans of silica, some silicates, coal dust and para-aramid fibrils, IARC monographs. 68. World Health Organization, Geneva

Xu XZ, Cai XG, Men XS (1993) A study of siliceous pneumoconiosis in a desert area of Sunan County, Gansu Province, China. Biomed Environ Sci 6:217–222

The Ecology of Soil-Borne Human Pathogens

20

Mark W. Bultman, Frederick S. Fisher, and Demosthenes Pappagianis

Contents

M.W. Bultman (✉)
Geology, Minerals, Energy and Geophysics Science Center, U.S. Geological Survey,
Tucson, AZ, USA
e-mail: mbultman@usgs.gov

F.S. Fisher
Department of Geosciences, University of Arizona, Tucson, AZ, USA

D. Pappagianis
School of Medicine, Medical Microbiology & Immunology, University of California–Davis, Davis, CA, USA

20.1 Introduction

The surface of the Earth, with the exception of the oceans and polar ice caps, is in large part covered with a marvelously complex layer of material called soil, from which we derive a host of useful products including fiber, fuels, building materials, animal forage, many mineral commodities, natural medicines (including antibiotics), and most of our food supply. Soil is teeming with life and is home for a huge array of living organisms. The vast majority of these living organisms are microbes that are ubiquitous on Earth (Table 20.1). They occur in all soils, salt and fresh water, the harsh climates of the Arctic and Antarctic, adjacent to deep-sea hydrothermal vents associated with spreading zones between tectonic plates, throughout the atmosphere, and deep below the surface of the Earth in oil wells where they have been isolated from the surface environment for millions of years (Staley 2002, p. 13).

Some soil-dwelling microbes are pathogenic for humans, including protozoa, fungi, bacteria, and also viruses and the less well understood prions, both of which require a plant or animal host for their survival. Helminths (which are in the mesofauna size class) are included in this chapter because of their importance as human pathogens and because of the numbers of pathogenic viruses and bacteria associated with them that can be introduced into the soil environment. Over 400 genera of bacteria have been identified with possibly as many as 10,000 species and, with the exception of viruses, they are in most cases more abundant than any other organism in soils. The number of bacteria that can be cultured in the laboratory is probably less than 1%; thus their actual diversity is probably much greater (Paul and Clark 1996; Coyne 1999). Fortunately, relatively few of this vast population of microbes are pathogenic for humans. For example, of the approximately 100,000 species of fungi currently recognized, only about 300 are known to cause human disease (McGinnis 1998).

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_20,

Table 20.1 Typical numbers of soil organisms in healthy ecosystems (NRCS 1999)

	Agricultural soils	Prairie soils	Forest soils
Bacteria Per teaspoon of soil (1 g)	100 million to 1 billion	100 million to 1 billion	100 million to 1 billion
Fungi per teaspoon of soil (1 g dry)	Several yards. (Dominated by vesicular-arbuscular mycorrhizal (VAM) fungi)	Tens to hundreds of yards. (Dominated by vesicular-arbuscular mycorrhizal (VAM) fungi)	Several hundred yards in deciduous forests. One to forty miles in coniferous forests (dominated by ectomycorrhizal fungi)
Protozoa per teaspoon of soil (1 g dry)	Several thousand flagellates and amoebae, one hundred to several hundred ciliates	Several thousand flagellates and amoebae, one hundred to several hundred ciliates	Several hundred thousand amoebae, fewer flagellates
Nematodes per teaspoon of soil (1 g dry)	Ten to twenty bacterial-feeders. A few fungal-feeders. Few predatory nematodes	Tens to several hundred	Several hundred bacterial- and fungal-feeders. Many predatory nematodes
Arthropods per cubic foot of soil	Up to 100	Five hundred to two thousand	Ten to twenty-five thousand. Many more species than in agricultural soils
Earthworms per cubic foot of soil	Five to thirty. More in soils with high organic matter	Ten to fifty. Arid or semi-arid areas may have none	Ten to fifty in deciduous woodlands. Very few in coniferous forests

Nonetheless, soil-borne human pathogens have extracted an unbelievable toll in disfigurement, suffering, blindness, death, and medical costs from the human race throughout history, and they will continue to do so for the foreseeable future. Examples are Ascariasis (roundworm) 60,000 deaths in 1993; Schistosomiasis, 200,000 deaths in 1993; and *Clostridium tetani*, killing 450,000 newborns and about 50,000 mothers each year (World Health Organization 1996). Almost three million deaths a year, mostly in developing countries, are attributed to diarrheal diseases, and many are contracted from microbes introduced into the soil via fecal waste and then ingested (NIAID 2000).

20.2 Soil Fundamentals

Soil may be defined as that part of the regolith that is capable of supporting plant life. The regolith is the portion of unconsolidated rock material that overlies bedrock and forms the surface of most land. The upper boundary of soils is either air or shallow water, and horizontally, soil boundaries are bodies of deep water, rock outcrops, or permanent ice fields. The lower limit of soil layers is the underlying bedrock where biological activity is severely restricted. Microbes are abundant in the upper parts of soil layers, where organic material is more likely to be present, and decrease in numbers with depth. However, some microbes may occur in the deepest soil layers and in places within the fractured underlying bedrock. The physical properties of soils are described in depth by Brian Alloway in Bioavailability of Elements in Soil in this volume. In this chapter only additional material important to the understanding of soil dwelling microbes is presented.

20.2.1 Important Physical and Chemical Properties of Soil

Physical properties of soils—texture, porosity, permeability, water-holding capacity, and temperature—are of considerable importance in determining the characteristics and microhabitat utilization of a given soil by microbes. Texture is the term used to describe the relative proportions of sand, silt, and clay-sized particles in a given soil (Fig. 20.1). Texture is important to microbes for several reasons: (1) it is the major factor controlling water holding capacity of a soil, (2) it determines the amount of pore space and the character (size and distribution) of the pore space, (3) it affects the rate of chemical reactions, (4) it is a major control on root penetration by plants, and (5) it controls soil aeration. All of these factors influence the types, distribution, and abundance of different microbes in any given soil profile.

Porosity is the percentage of the volume of a soil not occupied by solids; i.e., the interstices, isolated or interconnected, between individual solid soil particles (either mineral or organic). The availability of water and air are controlled largely by the amount, size, and interconnections between interstices, which are also the habitat for microbes.

Permeability is the ability of a soil to transmit water. Soils that contain a high proportion of sand will have larger continuous pores and will rapidly transmit water and air. In comparison, clay soils, which often have a high porosity because of the small size of individual particles, may have low permeability and transmit water slowly because of poor connectivity between soil interstices and swelling of individual clay particles. Soil water-holding capacity and soil temperature are discussed in the next section.

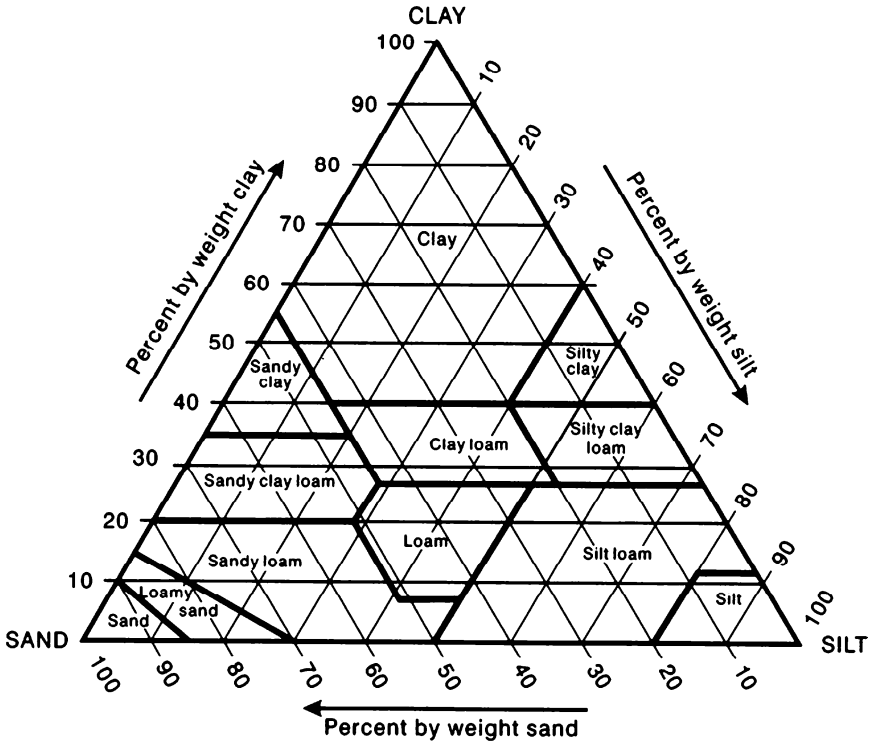


Fig. 20.1 U.S. Department of Agriculture soil texture triangle. Sand, silt, and clay are defined by the size of their constituent particles. Sand is composed of particles with sizes ranging from 2.0 to 0.05 mm. Silt is composed of particles from 0.05 to 0.002 mm. Clay is composed of particles less than 0.002 mm in size

20.2.2 Soil-Forming Factors

Five major factors interact to form soils: parent material, climate, soil organisms, topography, and time.

20.2.2.1 Parent Material

Parent material may be residual or transported igneous, metamorphic, or sedimentary bedrock debris. The mineralogy of parent material is very important in determining the type and amount of clay minerals developed in a given soil profile, which in turn has a profound effect on the types and distribution of soil microbes. Texture and the degree of consolidation of the parent material directly control the movement of water and air within the soil, the rooting ability of plants, and consequently the mobility of many microbes.

20.2.2.2 Climate

Climate characteristics, temperature and moisture in particular, influence the kinds and amount of soil microorganisms, the rate of decomposition of organic material, weathering rates of mineral matter, rates of formation of secondary minerals, biological activity, and the removal, movement, and deposition of materials between different soil layers.

Soil temperatures vary with latitude, altitude, slope aspect, vegetation cover (shading), soil color, and moisture content. Typically the surface temperature of soils ranges from below freezing to as much as 60°C and displays daily and annual cycles. Soil temperature has an important influence on biological processes such as the germination of seeds, nutrient absorption, root growth, and microbial activity, which generally is enhanced at higher temperatures if adequate moisture is available.

Precipitation may result in water on the land surface, water within the soil horizons, or deeper below the surface in ground water. Water that infiltrates the soil will surround soil particles, fill pore spaces, and may eventually move downward into the water table. A quantitative measure of water availability is the soil water potential, which is defined as the amount of work that must be done per unit quantity of pure water in order to transport reversibly and isothermally an infinitesimal quantity of water from a pool of pure water, at a specified elevation and at atmospheric pressure, to the soil water (at the point under consideration) (Soil Science Society of America 1998).

Informally, soil water potential is the amount of energy that must be expended to extract water from soil. It may be expressed mathematically as:

$$\Psi_{\text{soil}} = \psi_{\text{g}} + \psi_{\text{m}} + \psi_{\text{s}}$$

where Ψ_{soil} is the soil water potential, (ψ_{g}) is the gravitational potential, (ψ_{m}) is the matric potential, and (ψ_{s}) is the solute (or osmotic) potential. Water potentials are expressed in megapascals (MPa). One megapascal equals 1×10^6 Pa. One atmosphere equals 0.1013 MPa. The matric potential is related to the attraction of water molecules to solid surfaces.

In unsaturated soils it is always negative and becomes more negative as the surface area of the soil increases. Solute potential is associated with the solutes in the soil and is always a negative value. It becomes more negative as soil solutes increase and the soils become more saline. The gravitational potential may be positive or negative depending on where in the soil it is measured relative to some arbitrary reference level. The reference level is generally set in the soil profile at some point (e.g., at the water table) below the soil profile considered, thus the gravitational potential is usually positive in reference to microbes. If the reference level is the soil surface then the gravitational potential is negative for any point below the surface. The osmotic and matric potential together determine the amount of energy that must be expended by microbes to extract water from soil.

In general, microbial growth rates are greatest for Ψ_{soil} near −0.01 MPa and decrease as soils become drier and have correspondingly larger, negative water potentials. Microbial activity also decreases as soils become waterlogged or saturated, which results in Ψ_{soil} values at or near zero. At field capacity the matric potential of the soil water will generally be in the range of −0.01 to −0.03 MPa and the wilting point will occur near −1.5 MPa. Soils vary widely in the size and shape of their pore space and this variation precludes a simple relation between the water content of any given soil and the associated water potential. However, when a soil is saturated and all pores are completely filled with water, the matric potential is zero.

Because few soils have solid particles that are uniform in size, shape, or composition, texture and structure have a large influence on the matric potential of unsaturated soils. For example, the amount and type of clay in a given soil is extremely important, as some types of clay will swell as water is added and shrink as they dry. Matric potential is complicated by these changes. As water drains from a saturated soil containing some clay, its removal is mostly influenced by the larger particles (most often sand-sized particles) and the matric potential determination given above is relatively straightforward. As drying progresses the behavior of the clay particles becomes more important. The clay particles carry a variable negative electrical charge and thus tend to repel one another, and work is necessary to bring them closer together. Drying also causes the clay to shrink due to the loss of water between layers in the clay structure. The force of repulsion between the particles thus determines the matric potential of the remaining water. Detailed discussions of water potential and its effects on soil microbes may be found in Griffin (1972), Harris (1981), and Brown (1990).

Many soil microbes obtain nutrients from water by diffusion and rely on water for mobility, either by swimming, movement by Brownian forces, or flow of water. Void spaces smaller than the diameter of a given organism and pores from which water has been removed act as barriers. As soils dry, the water films around solid particles become thinner, which impedes bacterial and protozoal mobility, limits nutrient availability, and slows nutrient diffusion through cell membranes. If films become discontinuous, then microbes may be trapped or have to move in much more tortuous paths in the remaining water films. In general, in wet unsaturated soils, all major groups of microbes are active and in competition. In drying soils with progressively lower matric potentials, unicellular organisms become less active and ultimately microbial activity in relatively dry soils is confined largely to filamentous organisms, such as fungi that can, through growth of hyphae, utilize water unavailable to bacteria.

20.2.2.3 Soil Organisms

Many organisms have long been recognized as having an important role in the genesis of soils. Microbes play an important role in many soil processes. They convert plant and animal residue into humus and produce compounds that help bind soil into aggregates. They create new organic compounds that are nutrients for other microbes. Many microbes have mutually beneficial associations with higher plant forms and may fix the nitrogen that is needed by these organisms. They may also inhibit the growth of disease causing soil-borne organisms and mediate many soil related chemical reactions.

Soil microbial populations are generally more abundant in surface horizons than in deeper horizons but are not uniformly distributed laterally or with depth (Coyne 1999, p. 152). Typically, populations occur in localized concentrations associated with various favorable and unfavorable microenvironments throughout the soil profile (a vertical section of the soil which includes all the soil horizons). For example, the presence of plant roots is influential in determining the activity of microbes. The soil area influenced by a plant root is called the rhizosphere. This area (a few millimeters from the root) usually has chemical properties quite different from the bulk of the soil because of the uptake by the plant roots of moisture and nutrients and the secretions and exudates of amino and organic acids, sugars, proteins, and other chemical elements from the roots. Microbes may be up to 100 times more concentrated in the rhizosphere than in soil some distance from roots. Plants and microbes may form symbiotic relationships that are mutually beneficial. Examples are bacteria of the genera *Rhizobium* and *Bradyrhizobium* that convert (fix) atmospheric nitrogen into nitrogen compounds that can be utilized by the plant that in turn reciprocates and supplies carbohydrates to the bacteria. Another example are mycorrhizal fungi that attach to and grow on roots thereby acting as an extension of the root system which greatly facilitates the uptake of water and nutrients by the plant. In exchange, the fungus obtains sugars and other nutrients directly from the plant root.

20.2.2.4 Topography and Time

On the largest scale, climate and vegetation have great relevance to the distribution and formation of various soil types. On a smaller scale, topography modifies the climate and vegetation factors and has a major role in determining the character and development different soil types. The length of time that a given parent material has been exposed to the forces of climate, development of vegetation, and influence of animal life is believed by many soil scientists to be a strong influence on the nature of soil development. And, in general, the degree of development of differentiation between horizons in a soil is related to the age of a soil.

20.3 Soil as a Complex System

The study of complex systems has appeared as a separate scientific discipline only over the last 25 years and debate is ongoing as to its precise definition (Flake 1998; Cowan et al. 1999). However, there seems to be general agreement about some features that all complex systems exhibit:

1. They are collections of many simple nonlinear units that operate in parallel and interact locally with each other to produce behavior that cannot be directly deduced from the behavior of its component units
2. There are numerous units that operate in layers or at different scales
3. Local interactions within layers sometimes produce a global behavior at another scale or layer, which is called emergent behavior
4. The result of emergent behavior is that simple components combine to form more complexity than the sum of its parts (Flake 1998).

Soils are the interface between the lithosphere, atmosphere, hydrosphere, and biosphere. They are composed of an enormous number of individual constituents, both organic and inorganic, that are highly diverse in form, composition, and purpose. All the constituents adapt and interact with each other and respond chemically, biologically, and physically to processes specific to any of the abovementioned Earth domains. Rates of reaction to changes in environmental processes are not constant and rarely reach equilibrium. Many processes (perhaps most?) are nonlinear and thermodynamically irreversible. Soils are also open systems that exchange matter, energy, and organisms across numerous boundaries. This incredibly diverse, heterogeneous, dynamic, reactive, and adaptable nature makes the study and description of any aspect of soils a daunting task. Perhaps the overriding rule about soils is that they are constantly changing and adapting to their environment, and that the aggregate behavior of a soil cannot be predicted by summing up the behavior of its individual parts. These changes may be relatively rapid and readily visible or measurable (e.g., microbial growth, movement of soluble components, changes in water potential, seasonal changes in salinity, etc.) or they may be longer term taking place over years, decades, centuries, or millennia (e.g., weathering of silicate minerals and formation of clays, destruction of agricultural lands due to salt accumulation in irrigated areas, vegetation changes due to global warming or cooling, continental erosion, and deposition processes).

Most soil processes and microbial responses are gradational as are many boundaries between soil types, horizons, and chemical and physical properties. This lack of sharply defined volumes bedevils soil mapping and descriptions because spatial gradients in both process and products are at best poorly represented and most often ignored. One way of dealing with diverse data bounded by gradients is through the use of fuzzy logic; a method that allows the grouping of data into continuous sets in which membership can range between 0 and 1. Classes of data may then be expressed on an intermediate scale and spatial variations accounted for. The use of fuzzy logic (by the authors) to describe the habitat of a soil-borne human pathogen (*Coccidioides*) is described in the case study (Sect. 20.13).

20.4 Pathogen Classification

Classification is necessary to develop and discuss the connections and relationships between geology, soils, and soil-borne human pathogens. Soil fauna have been classified by several criteria including body size, degree of presence in soil, habitat preference, and activity (Wallwork 1970). Soil-borne human pathogens include many biological entities, most of which contain widely diverse members. Thus, it is useful to organize different pathogenic microbes into groups on the basis of the character of their soil residency. Any type of pathogen that is present in the soil, for any reason or time period, is included in the following classification in which two factors are essential: presence in soil (for any reason), and human pathogenicity. The classification terms of permanent, periodic, and transient from Wallwork (1970) were adopted, but redefined. An additional term, incidental, was added and defined.

Included in our classification are several human pathogens that have been classified as water- or food-borne. Many pathogens classified as water-borne are the result of animal or human fecal, urine, or other wastes introduced first into the soil environment and subsequently washed into surface water or incorporated into groundwater. Infection then follows by contact with, or consumption of, contaminated water. Similar circumstances take place with some food-borne diseases. The pathogen is again introduced into the soil via defecation or through some type of contaminated waste material and then may be consumed on unwashed raw fruit or vegetables.

Many pathogens have complex life cycles that may involve hosts in which they live and reproduce, biological vectors (insects, animals) and physical vectors (wind, water) for transport, and reservoirs to exist in during adverse environmental conditions. Soils provide these features to a wide variety of microbes and thus, the following classification to illustrate the importance of soil attributes and processes in the understanding the pathology of numerous diseases is offered.

In our classification some microbes may be classified as both transient and incidental. An example is *Giardia lamblia*, which can be introduced naturally into the soil through animal feces but can also be introduced anthropogenically via sewage systems, allowing for its dual classification. Four soil-borne pathogen residencies are defined below.

1. Permanent: Pathogenic organisms that are permanent soil inhabitants and can complete their entire life cycle within a soil environment. Examples are the bacteria *Clostridium botulinum*, *C. tetani*, *Burkholderia pseudomallei*, and *Listeria monocytogenes*. Also included are dimorphic organisms if one of their morphologic forms is capable of living and reproducing completely within the soil. Examples of permanent dimorphic soil pathogens are the fungi *Coccidioides* and *Histoplasma capsulatum*.
2. Periodic: Pathogenic organisms that require part of their life cycle to be completed within a soil environment on a regular, recurring basis. Examples are spores of *Bacillus anthracis* and the eggs laid in the soil by tick vectors that contain the bacterium *Rickettsia rickettsii*. Additional examples are eggs of the helminths *Ancylostoma duodenale* and *Necator americanus* (hookworms).
3. Transient: Pathogenic organisms that may naturally occur in soil, but the soil environment is not necessary for the completion of the organism's life cycle. Examples are cysts of a protozoan parasite *G. lamblia* and viruses in the genus *Hantavirus* that are introduced into soil environments worldwide via urine and feces of rodent vectors. Also included are *Leptospira*, a bacterium shed in urine of animals on soil, skin, and water and spores of the bacterium *Coxiella burnetii*.
4. Incidental: Pathogenic organisms introduced into the soil via anthropogenic means such as in sewage sludge, waste water, septic systems, unsanitary living conditions, biologically toxic spills, dumping of biohazardous waste materials, and release of biological warfare agents. Examples of viruses are enterovirus poliovirus (etiological agent of polio), enterovirus Coxsackie A and -B, and enterovirus hepatitis A. Length of survival time and virulence depend on numerous physical and chemical factors of the soil and the effluent, and it can range from hours to years.

20.5 Gateways for Infection

A soil-borne pathogen must come in physical contact with and establish itself in a human to cause disease. There are many ways in which this can be accomplished. First, soil-borne pathogens need to start this process from the soil. From the pathogen residency classification given above, the pathogen may be in the soil on a permanent, periodic, transient, or incidental basis. Permanent and periodic soil-borne organisms occur in the soil through natural pathways. Their presence in the soil is a normal part of the life cycle of a particular organism. Transient soil-borne organisms generally are incorporated into the soil via the excrement of wild or domestic animals. Residency in the soil is not necessary for their survival, but they are deposited there by natural means. Man places incidental organisms in the soil, generally in the form of solid human waste.

Probably the most common method of introducing soil-borne pathogens into the human body is through ingestion. Ingestion of soil-borne microbes is generally accidental, but in some cultures, geophagia is practiced. This subject is covered extensively in Geophagy and the Involuntary Ingestion of Soil, this volume.

In many countries, human waste is a valuable commodity that cannot be wasted. The practice of spreading human waste, or night soil, is found in many cultures and is an efficient means of introducing many human pathogens into the soil. The concentration of viruses in feces can be very high: e.g., enterovirus, 10^6 virus particles per gram; hepatitis A, 10^9 virus particles per gram; and rotavirus, 10^9 virus particles per gram (Sobsey and Shields 1987). In the United States and other industrialized nations, the application of human waste products in the form of treated sewage sludge is also practiced, but regulated. In the United States 24% of sludge is applied on or just below the land surface as fertilizer (Bertucci et al. 1987). Federal pollution control acts regulate the microbial content of the sludge and access to treated land by people, grazing animals, and use for crops grown for human consumption. In the United States, 90–99% of viruses are removed by primary and secondary sewage treatment (Bertucci et al. 1987). Raw sewage in the United States can contain concentrations as high as 100,000 infectious units per liter (Sobsey and Shields 1987). A 99% reduction still leaves 1,000 infectious units per liter. In countries where no such standard exists, pathogens from human waste have a much easier time entering soil and the environment. Many bacteria are also known to survive the sewage treatment process. If pathogens are ingested that became soil-borne as human waste, the gateway of infection will be referred to as the fecal-soil-oral gateway.

Other oral (ingested) gateways for soil-borne pathogens require an intermediate host. For example, the trematode *Schistosoma mansoni* has a transient soil residency. In soil its eggs hatch into larvae, which can infect an intermediate host. This intermediate host releases infective larvae into its environment, water. Consumption of infected water results in the pathogen moving to the human host. The cestode *Taenia saginata* also has a transient soil residency. Cattle ingest soil infected with this pathogen and develop cysts in their muscles. When humans consume incompletely cooked beef, they can become infected with the pathogen.

A respiratory gateway for soil-borne pathogens occurs when soil-borne pathogens are inhaled as airborne dust. Airborne dust is soil in motion. Human activity or natural forces can cause dust emission for every type of soil given the proper environmental conditions. Each year, several million tons of airborne soil makes its way from Africa to the Americas, Europe, and the Middle East as dust. Asian dust crosses the Pacific Ocean and dust from the southwestern United States can make its way to Canada (Raloff 2001). Every place on Earth receives dust from both local and distant sources and along with the minerals that make up the inorganic part of the dust ride fungi, bacteria, and viruses. African dust in the Caribbean has been shown to contain the fungus *Aspergillus sedowii*. Between 1973 and 1996 the Queen Elizabeth Hospital in Barbados documented a 17-fold increase in asthma attacks (NASA 2001). This time frame corresponds to an increased period of dust production in Africa due to drought conditions.

Microbes rarely penetrate the intact skin. It has been speculated that the bacterium *Francisella tularensis*, the etiological agent of tularemia, may be capable of penetrating unbroken skin, but there is little support for this. It may, however, enter through the thinner epithelium of the conjunctiva. The nematode *Strongyloides stercoralis* is capable of burrowing its way through healthy skin. The bacterium *C. tetani* enters the body if contaminated soil makes contact with a break in the skin.

20.6 Soil-Borne Human Pathogenic Helminths and Microbes

There are enormous numbers of helminths and microbes in the soil but only a small number of these are pathogenic to humans. Helminths are multicellular parasitic worms with complex reproductive systems and life cycles. Microbes are microscopic or submicroscopic organisms. For our purposes these include protozoa, fungi, bacteria, viruses, and, possibly, the agents (prions) of transmissible spongiform encephalopathies (TSE).

20.6.1 The Importance of Soil-Borne Human Pathogenic Helminths and Microbes

Diseases where the responsible pathogen spends some or part of its life in the soil are a major concern. In 1995, the World Health Organization (WHO) estimated that there were 3.7 million deaths worldwide from food-, water-, and soil-borne pathogens. More than 2.4 million of these deaths were children under the age of 5. In that same year, there were over four billion (4×10^9) new cases of these diseases (World Health Organization 1996). A soil-borne bacterium, *C. tetani*, was responsible for almost half a million deaths to newborns and 50,000 mothers each year due to tetanus (World Health Organization 1996). Billions of infections and over a million deaths occur each year from soil-borne helminths and protozoa.

Some soil-borne human pathogens are called frank pathogens because they are capable of infecting anyone. An example is *C. tetani*. Most soil-inhabiting pathogens are opportunistic pathogens; their main targets of opportunity are individuals with a suppressed immune system. These may include young children, the malnourished, HIV-positive individuals, individuals who have had transplant surgery, and the elderly.

20.6.2 The Distribution of Helminths and Microbes in Soils

The distribution of helminths in soils fluctuates greatly with season, climate, and amount of organic matter in the soil. Helminths typically prefer warm, moist soils with plentiful organic material. During favorable conditions, most helminths are found in the upper 10–15 cm of the soil profile. They may move vertically in the soil profile in response to seasonal weather changes.

As discussed in the introduction, microbes are found virtually everywhere on the planet. They cannot move very far on their own, but their small size allows them to be distributed globally by wind, water, animals, and humans.

On a small scale, the type of microbes in the soil are determined by the types of soils and the local climate. *Penicillium* is a fungus that is found in both warm and cold soils, whereas *Aspergillus* (a fungus) grows better in warm soils. *Fusarium*, a fungus that causes banana wilt, does not thrive in soils with the clay mineral smectite (Paul and Clark 1996). *Coccidioides*, a soil-borne fungus and the etiological agent of coccidioidomycosis, is found in dry, alkaline soils with a soil texture that includes large percentages of silt and very fine sand.

Within a given soil, the distribution of microbes on soil surfaces is uneven or irregular. Microbes are found clustered where conditions are favorable for growth and there may be a relatively large distance between the clusters. The determining factors for the locations of these clumps include the size and distribution of pore spaces in the soil, the soil water potential, the types of gases present in the soil pore spaces, the distribution of organic debris, and the local mineralogy of the soil. The influence of water potential on microbes has been discussed previously. Soil gases tend to be enriched in carbon dioxide and depleted in oxygen due to biological activity. Also, even in well-aerated soils, water may be blocking many pore spaces limiting the diffusion of oxygen in and carbon dioxide out. Some carbon dioxide is dissolved in soil water and produces carbonic acid, which helps to dissolve soil minerals (Coyne 1999). Many microbes have an affinity for clay minerals within soils. Clays have large surface areas, are chemically reactive, and have a net negative charge. They are a source of inorganic nutrients, such as potassium and ammonia, and modify the chemical and physical habitat immediately around them. Also, clays adsorb water making it less available for microbes.

Most soil-dwelling microbes are found in the upper 8 cm of the soil profile, and their numbers decrease significantly below 25 cm depth (Coyne 1999). The main reason is that the soil organic content, including root density, tends to decline with soil depth. Also, in alluvial soils the microbial populations fluctuate with textural changes in the soil profile. Microbes are more numerous in clay layers than in sand or coarser size materials. There is an increase in the numbers of microbes in the unsaturated zone directly above water table (Coyne 1999).

The population of soil-dwelling microbes is also affected by human activities. Global warming (anthropogenic and natural) is changing the characteristics of soils worldwide. Acid rain has changed the pH and mineralogy of soils and has affected the microbial populations of these soils. Clear-cutting can increase microbial populations in soils due to an increased supply of dead organic material (Paul and Clark 1996). Microbial populations are lower in tilled soils and compacted soils. Tilled soils are less moist than non-tilled and compacted soils have reduced pore space and aeration (Coyne 1999).

20.7 Soil-Borne Human Pathogenic Helminths

Human pathogenic helminths are all parasites and have man as their definitive host. Most inhabit the human intestines at some point in their life cycle; however, some are systemic in the lymph system or in other tissue.

Table 20.2 Selected soil-borne human pathogenic helminths

Nematodes
Ancylostoma duodenale and Necator americanus (hookworm)
Ascaris lumbricoides (roundworm)
Enterobius vermicularis (pinworm)
Strongyloides stercoralis (roundworm)
Toxocara canis (roundworm)
Trichuris trichiura (whipworm)
Trematodes
Schistosoma spp. (fluke)
Cestodes
Taenia saginata (tapeworm)
Taenia solium (tapeworm)

The exact taxonomy of the helminths is under continuing discussion and will not be addressed here. They can be grouped into the nematodes (including hookworms, roundworms, whipworms, and pinworms), the trematodes (flukes), and the cestodes (tapeworms). Human diseases caused by cestodes and trematodes are saprozoonoses. These are zoonotic diseases where the transmission of the disease requires a non-animal development site or reservoir. In many cases this site is the soil. Table 20.2 presents a summary of the soil-borne human pathogenic helminths discussed here.

20.7.1 Selected Soil-Borne Human Pathogenic Nematodes

Most human pathogenic helminths are nematodes. There are approximately 10,000 species of nematodes and approximately 1,000 of these are found in soils. Most of the soil-dwelling nematodes are found in the upper 10 cm of the soil profile. Desert soils have the lowest population density of nematodes (about 400,000/m^2) with the highest densities occurring in permanent pastures (up to 10,000,000/m^2). Although there are large numbers of soil-dwelling nematodes, they do not contribute to a large percentage of the biomass (Coyne 1999).

Nematodes are generally microscopic and transparent or translucent, ranging in length from about 0.05–2 mm. They have cylindrical unsegmented bodies with a bilateral symmetry. The body is covered with a tough cuticle. Nematodes have internal organs including a digestive, excretory, nervous, and muscular systems. They develop and grow by molting (shedding the cuticle). In almost all cases individual nematode species have sexual organs and separate sexes. Their life cycle begins with the development of an egg (most lay eggs in soil), followed by egg fertilization, embryonic growth in the egg, hatching and development of larvae, and molting and growth into an adult. Nematodes can produce five to six generations a year.

Nematodes are generally associated with water films in soils. These films partially fill the interstitial spaces between soil particles with water and are held in place by adhesion and cohesion. If soil becomes dry, many nematodes can form cysts or enter a dormant period allowing them to survive. They do well in warm organic-rich soils with a neutral pH, but can tolerate many soils. Most nematodes are predators or saprophytes. In this role they regulate microbial populations in soil by consuming up to 5,000 bacteria per minute (Coyne 1999).

The toll from nematode-caused human disease is staggering. Billions of people, rich and poor, are infected throughout the world each year, which causes much discomfort and suffering. There are over 130,000 deaths worldwide from nematode infections annually. Many infections are the result of a lack of appropriate personal hygiene or from poor sanitation. Most nematodes infect a human host by being ingested and some infect by entering through the skin. Almost all soil-borne human pathogenic nematodes inhabit the intestines. One exception is the nematode that causes trichinosis. In this case the mature nematodes live in the small intestines but, after a short period of time, they release larvae that migrate to striated muscle tissue and form cysts. Nematodes are the etiological agent of numerous human diseases and have complex life cycles involving soil, water, and animals as illustrated in the following discussion.

20.7.1.1 *Ancylostoma duodenale* and *Necator americanus*

Two species of hookworms are capable of causing human intestinal infection, generally called ancylostomiasis. These are *Ancylostoma duodenale* and *Necator americanus*. *A. duodenale* is found in parts of southern Europe, North Africa, northern Asia, and parts of western South America. *N. americanus* is found in Central and South America, southern Asia, Australia, and the Pacific Islands. Worldwide, there are approximately 1.2 billion cases annually of human hookworm infections (Cambridge University Schistosomiasis Research Group 2002). About 100 million of these involve a serious infection that creates a continuous loss of blood leading to chronic anemia. Less severe cases usually include mild diarrhea and cramps. Serious hookworm infection can create major health problems for newborns, children, pregnant women, and the malnourished.

Ancylostomiasis is a disease usually associated with unsanitary conditions. Hookworm eggs pass from the feces of infected humans to the soil. The eggs must be in the terrestrial environment to hatch and the soil residency of *A. duodenale* and *N. americanus* is periodic. Individuals with a major infection can excrete 2,000 eggs per gram of feces (National Institutes of Health 2001). Once exposed to air the eggs will develop rapidly in the upper few centimeters of moist warm soil. They hatch into larvae after a few days and feed on bacteria and organic matter. After about 5 days they molt and form the infectious form of the larvae. During cool damp periods, the larvae may come to the surface and extend their bodies into the air searching for a host. If they come into contact with human skin they attach and burrow in. The larvae are then transported in the blood to the lungs where they burrow into the airspace then migrate or are coughed up in the bronchi and trachea and are swallowed into the gut. Once in the intestine, they attach themselves to the wall of the small intestine and mature to adulthood which causes damage by blood ingestion. Thus, infection in conjunction with poor nutritional status can induce chronic anemia. Female *N. americanus* hookworms can produce 10,000 eggs daily and female *A. duodenale* can produce 20,000 eggs daily (Cambridge University Schistosomiasis Research Group 2002).

20.7.1.2 *Ascaris lumbricoides*

Ascaris lumbricoides is a large roundworm that causes ascariasis, an infection of the small intestines. There are over 1.5 billion new cases of ascariasis annually; about 210 million of them are symptomatic (Cambridge University Schistosomiasis Research Group 2002). Ascariasis is the most common helminthic infection and is distributed worldwide. The highest prevalence is in tropical and subtropical regions and in areas with inadequate sanitation.

A. lumbricoides is the largest nematode (roundworm) parasitizing the human intestine. Adult females can be 20–35 cm long and adult males can be 15–30 cm long (CDC 2002). Although infections may cause stunted growth, acute symptoms are usually not caused by adult worms. High worm burdens may cause abdominal pain and intestinal obstruction. Migrating adult worms may cause symptomatic occlusion of the biliary tract or oral expulsion. During the lung phase of larval migration, pulmonary symptoms can occur (cough, dyspnea, hemoptysis, eosinophilic pneumonitis).

The female parasite may produce 240,000 eggs per day (CDC 2002) which are passed in the feces. Fertile eggs may remain viable in the soil for many years if conditions are optimal. In warm, moist, shaded soil they begin to develop and can become infective after about 18 days. The soil residency of *A. lumbricoides* is classified as periodic. Infection occurs when the infective eggs are swallowed through the ingestion of contaminated raw food, such as fruit or vegetables, or through the incidental ingestion of soil. The eggs then hatch in the small intestine and the resulting larvae migrate to the lungs. In the lungs they molt twice and then migrate up through the air passages of the lungs to the trachea. They then enter the throat and are swallowed, finally ending up in the small intestine where they mature and mate, to complete their life cycle. Usually the roundworms only

feed on the semi-digested contents of the gut. There is some evidence that they may also feed on blood and tissue taken from the intestinal mucous membrane. The worms can live for up to 2 years in the intestine (CDC 2002).

20.7.1.3 *Enterobius vermicularis*

Enterobius vermicularis is a pinworm that causes enterobiasis, often referred to as human pinworm infection. *E. vermicularis* females grow up to 13 mm in length and males grow up to 5 mm. Enterobiasis is a common infection in children worldwide. It is the most common nematode parasite in temperate climates and in areas with modern sanitation. Symptoms are generally mild and vague. Most often anal itching is the only problem. It is estimated that over 200 million people are infected annually (Cambridge University Schistosomiasis Research Group 2002). Enterobiasis is most common in soils associated with poor sanitation where human feces are distributed in yards or fields, generally, as fertilizer.

Adult pinworms live in the human colon. Eggs are deposited by the female in the perianal region and can enter the environment with the feces. Given the proper conditions, the eggs become infective about 4 h after being laid. The eggs are resistant to drying and can remain infective in dust for several days (National Institutes of Health 2001). The soil residency of *E. vermicularis* is classified as incidental. Person-to-person transmission can also occur through handling items contaminated by an infected person. Following ingestion of infective eggs, the larvae hatch in the small intestine and the adults establish themselves in the colon.

20.7.1.4 *Strongyloides stercoralis*

Strongyloides stercoralis is a small roundworm, 2 mm in length, which causes strongyloidiasis. It is parasitic in the mucosa of small intestines. Strongyloidiasis is generally found in tropical and subtropical areas with poor sanitation. However, it does occur in temperate areas, including the southern United States (CDC 2002). In this area, it is often found in rural areas and in lower socioeconomic groups. Strongyloidiasis is often asymptomatic but sometimes causes chronic disease. In individuals who are immunosuppressed, it can be life-threatening.

The life cycle of *S. stercoralis* is complex. Unlike other helminths, its eggs can hatch in and re-infect the host. It is also capable of completing its life cycle and reproducing in the soil and thus is classified as a permanent soil resident. *S. stercoralis* eggs can hatch in the small intestines and produce rhabditiform larvae that can be passed in the feces of an infected host and may enter the soil. If they molt twice, they become a form (filariform) that is infectious to humans. If the rhabditiform larvae molt four times, they become free-living males and females that can mate and produce rhabditiform larvae in the soil (CDC 2002).

The parasitic cycle begins when filariform larvae penetrate human skin where they are then transported in the blood to the lungs and penetrate the airway. They are then swallowed and reach the small intestines where they molt twice and become adult, egg-laying females. The eggs hatch into rhabditiform larvae in the host. These larvae can either be passed in the feces or penetrate the intestinal mucosa which creates an internal autoinfection (CDC 2002).

20.7.1.5 *Toxocara canis*

Toxocara canis is the most common cause of visceral larva migrans. The disease is found worldwide and most often afflicts children aged 1–4 (Pitetti 2001). The definitive host for *T. canis* is the dog. Humans are called paratenic hosts, because the life cycle of *T. canis* cannot be completed in humans.

A heavily infected dog can pass millions of eggs each day in their feces (Pitetti 2001). Humans contract *Toxocara* infections as accidental hosts by ingesting embryonated eggs in contaminated soil (CDC 2002). The soil residency of *T. canis* is transient. The eggs hatch in the small intestine and the resulting larvae invade the mucosa and enter the bloodstream. The larvae can then disseminate to any organ in the body, provoke a granulomatous reaction, and die. Symptoms depend upon the degree of tissue damage and the associated immune system response. In the United States, 2–10% of children test positive for *Toxocara*, and international incidence is probably similar or slightly higher (Pitetti 2001).

20.7.1.6 *Trichuris trichiura*

Trichuris trichiura is the whipworm responsible for trichuriasis (often called whipworm disease). Trichuriasis is an infection of large intestine and is caused by the accidental ingestion of *T. trichiura* eggs. Most infections are associated with tropical areas with poor sanitation and occur in children. The disease occurs worldwide and there are an estimated 800 million people infected (CDC 2002). The soil residency of *T. trichiura* is classified as incidental.

Adult *T. trichiura* worms are approximately 4 cm in length. Female worms can shed 3,000–20,000 eggs per day in the feces. In the preferred environment of warm, moist, shaded soil the eggs embryonate and become infective in 15–30 days. They are often ingested through soil-contaminated food or hands causing the eggs to hatch in the small intestines. This releases larvae that mature to adults and establish themselves in the large intestine. There, the larvae fix themselves into the mucosa to feed. They have a life span of about 1 year (CDC 2002).

20.7.2 Selected Soil-Borne Human Pathogenic Trematodes

Trematodes are soft-bodied invertebrate animals with bilateral symmetry that are also called flukes. They can cause parasitic infections in humans. Trematodes have complex life cycles that always involve an intermediate host that is a mollusk. Trematodes and cestodes are members of a group of animals (the phylum Platyhelminths) that are commonly called flatworms, because most species are flattened dorsoventrally. This shape is due to the fact that they must respire by diffusion and no cell can be too far from the surface. Trematodes and cestodes have only one opening to the gut, which must both take in food and expel waste. The gut is often extensive and branched in order to provide nourishment to the entire animal.

As adults, trematodes are usually found in vertebrate animals including fish, amphibians, reptiles, birds, and mammals. These animals serve as the definitive hosts (a host where the adult parasite is able to reproduce sexually), but other intermediate hosts are usually involved in the trematode life cycle. Trematodes may cause highly severe infections of the lungs, bladder, blood, liver, and most often, the gastrointestinal tract. Several *Schistosoma* species are important human pathogens.

20.7.2.1 *Schistosoma mansoni, Schistosoma japonicum,* and *Schistosoma haematobium*

Schistosomiasis is a trematode infection that affects over 200 million people in 74 tropical countries. Up to 600 million people worldwide are at risk for this disease, which is spread by bathing or wading in infected rivers, lakes, and irrigation systems (World Health Organization 1996). People who are repeatedly infected can face liver, intestinal, lung, and bladder damage.

Schistosoma mansoni, S. japonicum, and *S. haematobium* are the three species that cause the most prevalent form of the disease. Each species occupies a different geographic region and there are slight differences in the clinical presentation of the disease itself. *S. mansoni* is found in parts of South America, the Caribbean, Africa, and the Middle East; *S. haematobium* in Africa and the Middle East; and *S. japonicum* in the Far East. These trematodes are unusual in that they reproduce sexually (one mated pair of *S. japonicum* can produce about 3,000 eggs per day) and that they live in the mesenteric veins that lie outside of the liver. The life expectancy of an adult is from 10 to 25 years. Some of these eggs can pass through the membranes of the bowels and are excreted in feces. The soil-residency of *Schistosoma* spp. is periodic. When fully wetted, the eggs hatch into free-swimming larvae call miracidia. These larvae infect amphibious snails that live in mud on the edges of bodies of water. These larvae are not infectious to humans at this stage of their development.

Once a snail is infected, the larvae reproduce asexually into another free-swimming larval stage and are called cercaria. The fork-tailed cercaria completes development in the snail, migrates to the surface of the snail's soft tissue, and enters the environment (water). If the cercaria comes into contact with the skin of a human host it releases enzymes that soften the skin and allow it to enter the host. Once in the host the cercaria sheds its tail and becomes a schistosomula and finds its way to the mesenteric veins where it matures into an adult worm capable of mating. Diseases caused by *Schistosoma* spp. are saprozoonoses. The disease is a zoonosis, but the eggs require an environmental residence to develop (CDC 2002; University of California 2002b).

20.7.3 Selected Soil-Borne Human Pathogenic Cestodes

Cestodes are often referred to as tapeworms because of the shape of their long ribbon-like body. They resemble a colony of animals in that their bodies are divided into a series of segments each with its own set of internal organs. Adults of the species can reach 100 m in length.

There are several parasitic tapeworm infections where man is the definitive host. Two examples of soil-borne tapeworms discussed below are *Taenia saginata* and *T. solium*; and both have a periodic soil residency. Adult parasitic human pathogenic cestodes live in the intestines. The bodies of cestodes can be divided into three regions: the scolex or head, the neck, and the strobila. The strobila is composed of a series of segments called proglottids and each proglottid has its own complete set of internal reproductive organs. As the organism grows proglottids are added, and the result may be thousands in a mature animal. This pattern of growth forms a long, ribbon-like body referred to as a tapeworm. Mature proglottids containing eggs are shed from the rear of the animal. The eggs, or proglottids, exit the body in feces and enter the environment.

20.7.3.1 *Taenia saginata* (The Beef Tapeworm)

Humans are the only definitive host for *T. saginata*, which causes about 50 million cases of tapeworm infection annually at locations spanning the globe. Infections are generally asymptomatic, but in some cases vitamin deficiency may be the result of excessive absorption of nutrients by the parasite. Occasionally mild symptoms like abdominal pain, digestive disturbances, excessive appetite, or loss of appetite, weakness, and loss of weight may accompany the infection.

Adult tapeworms live in the small intestines. They are generally 5 m or less long, but may reach lengths up to 25 m. Mature worms have over 1,000 proglottids and mature

proglottids contain 80,000–100,000 eggs each. Once mature, the proglottids separate from the tapeworm and can be passed in the feces and may enter the soil. The soil residency is classified as transient. The eggs can survive for months to years in the environment. Cattle become infected by ingesting vegetation contaminated with eggs (or proglottids). The eggs develop in the intestines and release onchospheres that evaginate and invade the intestinal wall. They then migrate to the striated muscles and develop into a cysticercus, which is capable of surviving for several years. Ingesting undercooked meat containing cysticercus infects humans. In the human intestine the cysticercus develops into an adult tapeworm that is capable of surviving for over 30 years (CDC 2002).

20.7.3.2 *Taenia solium* (The Pork Tapeworm)

Taenia solium, the pork tapeworm, can cause both taeniasis and cysticercosis in humans. As with beef tapeworm infections, the disease occurs worldwide with about 50 million annual cases. *T. solium* and *T. saginata* have very similar life cycles, but *T. solium* larvae can infect humans as well as swine. Humans are the definitive host for *T. solium* adults, which can live up to 25 years in the intestine. *T. solium* have less than 1,000 proglottids and are 2–7 m in length. Their proglottids are less active than in *T. saginata* and each contain about 50,000 eggs. When the proglottids mature, eggs or proglottids are shed in the feces and into the soil. The soil residency is classified as transient. If swine ingest them, they mature to onchospheres that move to the muscles and grow into the larval form of *T. solium, Cysticercus cellulosae*. Undercooked pork from infected swine is then infective for taeniasis in humans.

If humans swallow the eggs or proglottids, they can also develop *C. cellulosae* infection which results in cysticercosis, a disease that can be quite severe. It can also develop in humans infected with adult *T. solium* due to autoinfection from proglottids carried to the stomach by reverse peristalsis. In humans, *C. cellulosae* can develop in the striated muscles, the brain, the liver, and in other tissues. When infecting the central nervous system, the pressure produced by growing larvae can cause severe pain, paralysis, optical and/or psychic disturbances, and epileptic convulsions. Mortality due to cysticercosis is estimated at about 50,000 worldwide (CDC 2002).

20.8 Selected Soil-Borne Human Pathogenic Protozoa

Protozoa are single cell eukaryotic organisms that are phagotrophic, which means that they feed by engulfing and ingesting their prey inside a cell membrane. There are over 30,000 species of which only a small number are parasites of man. Most protozoa range in size from 0.01 to 0.1 mm. In soils, they feed on bacteria and algae. The protozoan life cycle ranges from binary fission in a single host to many morphological transformations in a series of hosts. There are no eggs, larva, or adults. There are about 10,000–100,000 protozoa per gram of upper soil surface (Coyne 1999). Archeozoa are similar to protozoa except that they lack mitochondria.

20.8.1 *Cryptosporidium parvum*

Cryptosporidium parvum is a protozoan that causes a self-limiting diarrheal illness called cryptosporidiosis. It can be more serious in infants and in the immunosuppressed. The first documented human case occurred in 1976 and cryptosporidiosis is now considered a worldwide disease. It is especially common in developing countries. No effective specific treatment is known. The incidence of cryptosporidiosis is not known, but an outbreak in Milwaukee in 1993 infected over 400,000 people (Coyne 1999).

Many animal species, including man, act as a reservoir for *C. parvum*. Infected hosts excrete sporulated oocysts in the feces. In this fashion, oocysts may enter the soil as animal or human waste products. Infection begins when oocysts are ingested, most often through contaminated water and food or by direct fecal-oral transmission. After ingestion, the oocysts mature to sporozoites and parasitize the epithelial cells of the gastrointestinal tract. These parasites then undergo asexual and then sexual reproduction. The soil residency of *C. parvum* is transient-incidental. Generally cryptosporidiosis is a water-borne disease, but *C. parvum* only reaches the environment through human and animal waste. The organism must survive in and travel through soils to become water-borne in many cases. Studies have indicated that *C. parvum* can survive in surface water for 6 months and in liquid manure tanks for many months (Cambridge University Schistosomiasis Research Group 2002; CDC 2002; and Health Canada 2002).

20.8.2 *Cyclospora cayetanensis*

Cyclospora cayetanensis causes cyclosporiasis, a diarrheal disease found worldwide. It is most commonly found in tropical and subtropical regions. This disease has a life cycle similar to *C. parvum* with the exception that when passed in the feces, the oocyst is not infective. This means that no direct fecal-oral transmission can occur. Freshly passed oocysts sporulate after spending days or weeks in the environment at temperatures of between 22 and 32°C and become infective. Soil residency is incidental. Little is known about possible animal reservoirs or environmental

survival time for *C. cayetanensis*. The oocysts are thought to be able to survive for long periods of time in the environment if kept moist (CDC 2002).

20.8.3 *Entamoeba histolytica*

Entamoeba histolytica causes amebiasis, a disease characterized by diarrhea, with severe cases including dysentery or a serious invasive liver abscess. Amebiasis is a worldwide disease with an estimated 40,000,000 people infected annually. Around 40,000 die from the disease each year. Man is the definitive host. As with *G. lamblia, E. histolytica* exists in two forms: the active parasite (trophozoite) and the dormant parasite (cyst). The trophozoites live in the intestine and feed on bacteria or on the wall of the intestine. Trophozoites are expelled in feces and die rapidly. However, cysts expelled in the feces are very hardy and can survive days to weeks in the external environment. In areas where sanitation is poor, indirect transmission of the cysts is more common. The soil residency of *E. histolytica* is incidental.

20.8.4 *Balantidium coli*

Balantidium coli is the largest protozoan found in humans. It causes balantidiasis, a disease with a worldwide distribution. It is capable of causing acute hemorrhagic diarrhea and ulceration of the colon. Pigs, large primates, humans, and dogs are the definitive hosts. The life cycle and methods of transmission of the parasite are similar to *E. histolytica*. The soil residency is transient-incidental. Under favorable temperature and humidity conditions, the cysts can survive in soil or water for weeks to months (CDC 2002).

20.8.5 *Giardia lamblia*

Giardiasis is a disease that is especially common among children and in places where sanitation is poor worldwide. About 200 million people in Asia, Africa, and Latin America display symptoms and there are about 500,000 new cases annually (World Health Organization 1996). It is also one of the most common parasitic diseases in developed nations. It is caused by *G. lamblia*, an archeozoan. Clinically, giardiasis presents as non-inflammatory diarrhea and associated abdominal cramps, bloating, fatigue, and weight loss. Infections can be asymptomatic or chronic.

G. lamblia trophozoites (the active stage of organism) live in the large intestine of infected humans or animals. At times they form cysts and millions of these cysts (and trophozoites) are released in the feces and may enter the soil. The soil residency of *G. lamblia* is transient-incidental. The cysts can persist for some time (up to many months) in the environment, which includes soil, food, water, or surfaces that have been contaminated. Infection results from ingestion of the cyst, usually in contaminated water or food (CDC 2002).

20.8.6 *Isospora belli*

Isospora belli is a protozoan that causes isosporiasis, an infection of the small intestine. Isosporiasis is found worldwide, but is most common in tropical and subtropical areas. It can cause chronic diarrhea, abdominal pain, and weight loss and is especially important in immunosuppressed individuals.

Large and football-shaped *I. belli* oocysts are passed in the feces. The soil residency of *I. belli* is transient-incidental. The oocysts contain a sporoblast (rarely two) that splits and develops cyst walls, thereby becoming sporocysts. Infection occurs by ingestion of sporocysts. *I. belli* has a complex asexual and sexually reproductive cycle within its host, both human and animal, which results in the production of oocysts that are then excreted in feces (CDC 2002).

20.8.7 *Toxoplasma gondii*

Toxoplasma gondii causes toxoplasmosis. *T. gondii* infection can produce flu-like symptoms in healthy people and severe disseminated disease in immunosuppressed individuals. It can also cause birth defects in infants when women are exposed during pregnancy. Toxoplasmosis occurs worldwide and is more common in warm climates and at low altitudes. Under some conditions, toxoplasmosis can cause serious pathology, including hepatitis, pneumonia, blindness, and severe neurological disorders.

Cats are the definitive host for *T. gondii*. They generally acquire the infection though consumption of infected rodents. After a cat consumes the tissue containing cysts or oocysts, viable tachyzoites invade the small intestine. These eventually form oocysts that are excreted. The soil residency of *T. gondii* is transient. The oocysts can remain infective in water or soil for about 1 year. Tachyzoites can also form cysts in tissue. Humans can become infected in several ways including the ingestion of cysts through contaminated food or soil or ingestion of undercooked meat (e.g., lamb, pork, or beef) infected with cysts.

Its life cycle includes two phases called the intestinal (or enteroepithelial) and extraintestinal phases. The intestinal phase occurs in cats only (wild as well as domesticated cats) and produces oocysts. The extraintestinal phase occurs

in all infected animals (including cats) and produces tachyzoites and, eventually, bradyzoites or zoitocysts. The disease toxoplasmosis can be transmitted by ingestion of oocysts in cat feces or bradyzoites in raw or undercooked meat (CDC 2002; Health Canada 2002).

20.8.8 *Dientamoeba fragilis*

Dientamoeba fragilis is a protozoan responsible for *D. fragilis* infection. Symptoms occur in only 15–25% of infected individuals and may result in mild, chronic gastrointestinal problems (abdominal pain, gas, diarrhea, etc.). The disease is found worldwide but has a higher prevalence in developing countries with poor sanitation.

D. fragilis trophozoites are one of the smallest human parasites and survive in the human gastrointestinal tract. No cyst stage has been reported, so a fecal-soil-oral infectious gateway is unlikely. The soil residency of *D. fragilis* is unknown. There is evidence that this organism is transmitted among humans in the eggs of human pinworms (*E. vermicularis*). Infection by *D. fragilis* may require infection by *E. vermicularis* (Mack 2001; CDC 2002).

20.9 Selected Soil-Borne Human Pathogenic Fungi

There are over 100,000 species of fungi of which about 300 are known to be pathogenic (McGinnis 1998). Fungi are non-motile eukaryotic organisms with chitin-based cell walls that can be grouped into molds and yeasts. Molds are composed of branching filaments called hyphae that grow by elongation at their tips. Hyphae can be composed of one cell with continuous cytoplasm, called coenocytic hyphae, or are composed of cells separated by walls (septa) in which case they are called septate hyphae. The mass of hyphae of an individual organism is referred to as mycelium. Reproduction is through sexual or asexual spores and fragmentation of hyphae. Single cell non-filamentous fungi are called yeasts. They are generally spherical or ovoid in shape and reproduce by budding. Some fungi are dimorphic in that they can switch between filamentous or yeast growth.

Most fungi are saprophytes and must absorb nutrients from the environment. In this way they help decompose dead plants and animals. Fungi do not contain chlorophyll and are therefore not capable of photosynthesis. Most molds are aerobic and cannot survive in saturated soils. They need to be able to extend hyphae into air spaces that contain oxygen. Many yeasts are facultative anaerobes and some yeasts are capable of surviving in anaerobic environments. In moist soils, the largest fraction of the microbial biomass is made up of fungi. Soil-borne fungi are more tolerant of acidic soils, grow best between 6 and 50°C, and are usually found in the top 15 cm of the soil (Coyne 1999).

The life cycles of the various fungi are generally not as complex or varied as helminths and protozoa. Table 20.3 illustrates selected soil-borne human pathogenic fungi and the diseases they cause. It also contains information on the geographic distribution, soil residency, gateway of infection, and disease incidence for each of the fungi.

20.10 Selected Soil-Borne Human Pathogenic Bacteria

Coyne (1999), presents an excellent introduction to bacteria in the soil that is summarized below. Bacteria are single-cell prokaryotic organisms that have existed on Earth for over three billion years. They are small, generally less than 50 μm in length and 4 μm in width. One bacterium weighs about 10^{-12} g. Their small size affords them a high surface area to volume ratio, which allows them to maximize nutrient uptake through diffusion. They also possess a very high metabolism and the ability to reproduce through binary fission. Bacteria occur in a wide variety of shapes. Aerobic bacteria require oxygen for existence, whereas anaerobic bacteria do not tolerate gaseous oxygen. Some bacteria, called facultative anaerobes, prefer oxygen, but can grow without it. Heterotrophic bacteria use organic compounds in the environment for energy and for synthesis of cellular constituents. Autotrophs make use of energy from light or of reactions of inorganic chemicals to fix carbon dioxide and synthesize organic cellular components (University of California 2002a).

There are up to one billion bacteria in 1 g of soil (Table 20.1). In general, they prefer warm, moist soils. Soil bacteria can be classified as autochthonous or allocthonous. Autochthonous organisms inhabit the bulk of the soil and are specialists at getting the most out of the available nutrients. Allocthonous microbes are more opportunistic. They are generally saprophytic or pathogenic and tend to be found in areas that are rich with nutrients, even if only for a limited time period. Allocthonous microbes maximize growth when conditions are right. They often are found in the rhizosphere and many of them are plant pathogens. Allocthonous microbes can cover 5–10% of root surfaces and there is a steep decrease in microbial populations just 5 mm from the plant root.

There are a large number of soil-borne human pathogenic bacteria that generally have similar life cycles. They grow and reproduce through binary fission in the proper conditions. When the conditions are unfavorable, they may die off or form a spore that can grow and reproduce again when conditions improve. Examples of several soil-borne human pathogenic bacteria are presented in Table 20.4.

Table 20.3 Selected soil-borne human pathogenic fungi and their properties

Pathogen(s) and disease	Distribution and residency	Gateway(s) and incidence[a]	Comments and soil survival time
Coccidioides spp Disease: *coccidioidomycosis*	Southwestern U.S., Mexico, Microfoci in Central and South America Residency: permanent	Respiratory, rarely trauma Incidence: 15/100,000 in Arizona in 1995	Please see the case study in Sect. 20.13 for complete information
Histoplasma capsulatum Disease: *histoplasmosis*	Locally in eastern and central U.S., Microfoci in Central and South America, Africa, India, and southeast Asia Residency: permanent	Respiratory Incidence: About 80% of people living in endemic area have a positive skin-test. Mortality rate is about 10% in HIV-infected persons with disseminated disease	Found in soils contaminated with bird or bat feces
Blastomyces dermatitidis Disease: blastomycosis	South-central, south-eastern and mid-western U.S., Microfoci in Central and South America, and Africa Residency: permanent	Respiratory Incidence: 1–2/100,000 in endemic areas	Found in soils enriched with decomposing organic debris
Aspergillus fumigatus; A. flavus; less commonly A. terreus, A. nidulans, A. niger Disease: aspergillosis	Worldwide; ubiquitous; found in soil, dust, plants, food, and water Residency: permanent	Respiratory, occasionally via contaminated biomedical devices Incidence: 1–2/100,000 is suggested	Found in soils, decomposing plant material, house-hold dust, food, water, and plants
Sporothrix schenckii Disease: sporotrichosis	New World, Africa, and Europe Residency: permanent	Trauma to skin Incidence: disease is uncommon and sporadic	Most common in sphagnum moss, plants, baled hay

References: CDC (2002), DoctorFungus (2002), Health Canada (2002)
[a]Incidence is the annual rate of confirmed infection. An incidence of 15/100,000 means that there were 15 confirmed cases per 100,000 population

Actinomycetes are prokaryotic bacterial organisms that display filamentous growth. They make up 10–50% of the total microbial population in soils (Coyne 1999). Most actinomycetes are aerobic and prefer warm, dry soils. They tend to be spore formers and are adept at surviving droughts. Most are saprophytic. Due to their filamentous growth, they resemble fungi. The filaments of actinomycetes are much smaller than fungal hyphae, 0.5–1.0 μm as opposed to 3–8 μm for fungi. Although actinomycetes are an important component of the soil microbial population, there are few known to be important soil-borne human pathogens. Four pathogenic genera are included in Table 20.4. Many actinomycetes make antibacterial molecules. About 75% of the 5,000 known antibacterial drugs are derived from actinomycetes (Coyne 1999).

20.11 Selected Soil-Borne Human Pathogenic Viruses

There are over 140 types of pathogenic enteric viruses transmitted from humans to the environment in human feces. For some (like the Norwalk virus and rotavirus), immunity is short-term; there is no life-long protection after recovering from an infection. Viruses are the smallest pathogens, most having maximum dimension of less than 30 nm (Coyne 1999). They are acellular organisms, have no cell membrane, and occur in many shapes including cubic, helical, and icosahedral. Most viruses have two basic structural components: a protein coating that can help the virus survive in the environment and a nucleic acid core. Viruses are so small that their genetic material (the nucleic acid core) contains only 10–200 genes (Coyne 1999). There are viruses that infect animals, plants, fungi, protozoans, algae, and bacteria. They are always host specific.

Viruses are parasites that must use the chemical machinery and metabolism of a host cell to reproduce. In a host, viruses attach to a cell, use enzymes to break through the cell wall, and inject their nucleic acid core into the cell. Once in the cell, the genetic material from the virus begins making three types of proteins. It replicates its own genetic material, builds protein coating, and assembles proteins that will help it get out of the cell. These parts come together by chance to form and release a single or many new copies of the original virus. Outside of the host, viruses are inert. They do not grow or reproduce. Human pathogenic viruses with protective coatings can remain infectious in the environment for up to 6 months.

20.11.1 Viruses in Soils

Soil is not a natural reservoir for viruses. Viruses can only persist in soil in a dormant state but may retain their infectivity in this state. Plant viruses rarely survive in soils for long periods; however, some insect viruses remain infective for years (Coyne 1999). Viruses are also known to infect many soil helminths and microbes.

Table 20.4 Selected soil-borne human pathogenic bacteria and their properties

Pathogen(s) and disease	Distribution and residency	Gateway(s) and information on incidence, morbidity and/or mortality (IMM) if available	Comments and soil survival time if available
Actinomadura spp. Disease: maduramycosis, actinomycetoma	Tropical regions, especially Africa, India, South and Central America Residency: permanent?	Skin trauma IMM: Africa may have highest incidence. In the Sudan, 300–400 patients per year are seen. Causes disfigurement, rarely fatal	An aerobic actinomycete that is a soil saprophyte
Bacilllus anthracis Disease: anthrax	South and Central America, Southern and Eastern Europe, Asia, Africa, the Caribbean, and the Middle East Residency: periodic(?)	Respiration, Skin trauma, ingested (gastrointestinal); often infected by above methods while handling contaminated animal products IMM: Unknown to rare	A spore forming aerobic bacterium. Spores can survive in soil environment for many years, possibly for decades. Biological warfare agent
Burkholderia (Pseudomonas) pseudomallei Disease: melioidosis	Worldwide, primarily in tropical and subtropical regions, especially in Southeast Asia and northern Australia, also in South Pacific Africa, India, and Middle East. Isolated cases in Central and South America, Hawaii and Georgia Residency: permanent	Direct contact with contaminated soil and water, inhalation of dust, ingestion of contaminated water, skin trauma, and contact with mucous membranes IMM: Very important cause of morbidity and mortality in Thailand	Possible biological warfare agent. Soil saprophyte
Campylobacter jejuni Disease: diarrhea, gastroenteritis, Guillain-Barre Syndrome	Worldwide Residency: incidental	Fecal-soil- oral. Contaminated water, raw milk, and raw or under-cooked meat, poultry, or shellfish. IMM: *C. jejuni* along with rotaviruses and enterotoxigenic *E. coli*, is a major cause of diarrhea worldwide	No information on survival in soil. Known to survive in 4°C stream water for over 4 months
Clostridium tetani Disease: tetanus	Worldwide, most frequently in densely populated regions in hot damp climates in soils rich in organic material especially manure Residency: permanent? (In proper setting it can complete life cycle in soil.)	Fecal-soil-oral, humans and many animals. Tetanus spores introduced into the body through a wound contaminated with soil, street dust or feces, or injected street drugs; also through lacerations, burns and trivial wounds IMM: Annual deaths – newborn = 450,000 maternal = 50,000	Anaerobic (all *Clostridium* spp.) but can form spores that are very resistant to heat, many antiseptics, and chemical agents
Clostridium botulinum Disease: botulism	Worldwide Residency: permanent? (in proper setting can complete life cycle in soil)	Fecal-soil-oral, humans and many animals. Ingestion of contaminated food containing toxin. IMM: Generally rare, associated with confined outbreaks	Obligate anaerobic bacterium that can live in oxygen-free pockets in soil as vegetative cells or spore. Present in the soil and water. Spores can be found on food that comes into contact with infected soil or water
Clostridium spp. (other than 2 listed above) Disease: gas gangrene	Worldwide Residency: permanent? (in proper setting can complete life cycle in soil)	Fecal-soil-oral, humans and many animals. Skin trauma (major or minor), burns, deep puncture wounds, ear infections, animal bites. Spores introduced into the body through a wound contaminated with soil, street dust or feces, or injected street drugs; also through lacerations, burns and trivial wounds IMM: Fairly common before general use of antibiotics to treat injuries. Can still pose threat to immunosuppressed	Before antibiotic treatments, about 5% of battlefield injuries were complicated by this bacterium

Coxiella burnetti Disease: *Q fever*	Worldwide with the exception of New Zealand and Antarctica Residency: transient	Inhalation of infected aerosol, often produced from animal products and especially during parturition; also shed in urine and feces IMM Morbidity from 5% in urban to 30% in rural areas worldwide	Highly infective, but unable to grow outside of host (commonly goats, sheep, and cattle). Has a spore like form that is very resistant to heat and desiccation and can last for months outside of host in soils
Escherichia coli several pathogenic strains Disease diarrhea	Worldwide Residency: incidental	Fecal-soil-oral, ingestion of contaminated food IMM: Major cause of traveler's diarrhea of which there are some five million cases per year worldwide	Can survive for months in cool, dark, nutrient rich soils
Francisella tularensis Disease: tularemia	Many areas of U.S. with most cases in Arkansas, Oklahoma, and Missouri; increasing numbers of cases in the Scandinavian countries, eastern Europe, and Siberia; also in Middle East and Japan; rare in the U.K., Africa, and Central and South America Residency: transient	Enters soil through tick feces and possibly other sources. Humans acquire through contact of infected soil with broken skin (might be able to penetrate unbroken skin) and with mucous membranes. Also tick and insect bites, inhalation, and ingestion IMM: Worldwide incidence not known. In the U.S. there are now less than 200 cases per year	A zoonosis; one of the most infectious agents known; highly infectious in both skin and aerosol routes; often found in rural areas; possible biological warfare agent. Known to survive in water and moist soil for weeks
Leptospira spp. Disease: leptospirosis	Worldwide, but more common in temperate or tropical climates Residency: transient	Ingestion and skin contact, especially mucosal surfaces; contact with water, food, or soil contaminated with urine from infected animals IMM: About 200 cases annually in U.S. Considered to be the most widespread zoonotic disease in the world	Outbreaks associated with heavy rainfall and flooding. Known to survive many weeks in contaminated soil
Listeria monocytogenes Disease: listeriosis	Worldwide?; Residency: permanent	Ingestion of contaminated (often uncooked) food contaminated by infected soil and water IMM: 2,500 serious cases per year in U.S. of which 500 are fatal	Found in soil, water, and fecal material of domestic animals. Can grow at temperatures found in refrigerators
Nocardia spp., *Rhodococcus* spp. Disease: Nocardiosis	Worldwide, some species more likely in tropics Residency: permanent?	Cutaneous disease from skin trauma contaminated with soil, pulmonary and disseminated infections from inhalation IMM: In the U.S., there are an estimated 500–1,000 new cases of Norcardia infection annually	Aerobic actinomycete found in soil and water. *Nocardia asteroides* is tolerant of 40–50°C
Rickettsia rickettsii and other *Rickettsia* spp. Disease: Rocky Mountain spotted fever, other fevers (African tick bite fever, Queensland tick typhus) and spotted fevers (Mediterranean, Japanese)	Worldwide, individual species are geographically contained by their mammalian reservoir. Residency: periodic	Zoonosis-spread by bite of tick or by contamination of the skin with tick blood or feces. Rodents are the main mammalian reservoir IMM: 3–5% of individuals who become ill with Rocky Mountain spotted fever still die from the infection. U.S. has 250–1,200 cases of Rocky Mountain spotted fever annually	Vector in North America is the wood tick., American dog tick, or Lone Star tick. Female ticks transfer the bacterium to their eggs that are infective as they mature and hatch in the soil (unlike many other tick vector diseases, i.e. Lyme disease). Feces of infected ticks quickly lose their infectivity on drying

(continued)

Table 20.4 (continued)

Pathogen(s) and disease	Distribution and residency	Gateway(s) and information on incidence, morbidity and/or mortality (IMM) if available	Comments and soil survival time if available
Salmonella spp. Disease: primarily salmonellosis (diarrhea), typhoid fever and paratyphoid fever	Worldwide, primarily a food-borne disease Residency: incidental	Fecal-soil-oral, ingestion of contaminated (often uncooked) food contaminated by infected soil and water; shed in human and animal feces IMM: 2–4 million cases in U.S. annually	Can survive in sludges and soils for many months given proper conditions. In sludge applied to arid soils survival may be 6–7 weeks
Shigella spp. Disease: diarrhea, dysentery	Worldwide Residency: incidental	Fecal-soil-oral, ingestion of contaminated (often uncooked) food contaminated by infected soil and water; shed in human feces IMM: 300,000 cases annually in U.S.	May survive a few weeks in water below 10°C. Soil survival unknown
Streptomyces spp. Disease: skin infection	Africa, India, Latin America Residency: permanent?	Skin trauma contaminated with soil IMM: Invasive infection is extremely rare	An aerobic actinomycete and soil saprophyte
Thermoactinomyces spp. Disease: hypersensitivity pneumonitis, mushroom worker's lung, farmer's lung	Probably worldwide Residency: permanent?	Inhalation IMM: Farmer's lung can occur in 2–10% of farm workers but is regionally variable	Actinomycete found in soil, contaminated compost piles, silos. Tolerant of 45–60°C heat
Yersinia spp. Disease: diarrhea	Worldwide Residency: incidental	Fecal-soil-oral, ingestion of contaminated (often uncooked) food contaminated by infected soil and water; shed in human and animal feces	Disease most often occurs in infants and small children. *Y. enterocolitica* is known to survive in soil for 540 days

References: Ania and Asenjo (2002), Canadian Centre for Occupational Health and Safety (1999), Carey and others (2001), CDC (2002), Health Canada (2002), Rusin and others (2000), World Health Organization (1996)

There are a number of factors that influence the ability of a virus to survive and to move in unsaturated soils. Because viruses are added to soil as anthropogenic waste, the objective is to keep the viruses from the water table where generally cool water temperatures can keep the virus alive for long periods of time. Factors that affect survival include: temperature, soil moisture, soil microbial activity, soil type, virus type, soil organic matter, and adsorption of the virus, generally to clay minerals (Sobsey and Shields 1987). Viruses generally survive longer in cooler, wetter, pH neutral soils with low microbial activity. Humic and fulvic organic material may cause reversible loss of infectivity, but, some other organic materials may complex with the virus and protect it from inactivation by preventing adsorption to soil particles (Sobsey and Shields 1987). Specific species of viruses have different survivability. Also, viruses tend to survive longer when clustered together.

Soil texture is also very important to virus survival in unsaturated soils. Clay minerals, which are found in fine-grained soils, prolong the virus's ability to survive through adsorption of the virus to these minerals. Adsorption by clay minerals can prolong virus survival because adsorption affords protection against inactivation (Gerba 1987). They are essentially removed from the water film and physically protected by the clay minerals, but, they can desorb back into water with heavy rain and then be moved passively in the soil. Matson et al. (1987) showed that viruses have been recovered at distances from septic tanks of over 90 m horizontally and up to 67 m vertically in the soil, summarized the survival of viruses in soils. These results indicated that viruses have been shown to survive from 11 to 180 days in soils, and the length of survival depended on the type of soil, humidity, the soil moisture, and the soil temperature.

Sim and Chrysikopoulos (2000) presented an excellent review of virus sorption in an unsaturated soil. The ability of a soil to adsorb viruses is strongly correlated with the degree of soil moisture. Decreasing the moisture content enhances virus sorption onto the solid matrix by forcing viruses to move into a thin film of water surrounding soil particles. Virus adsorption at the liquid–solid interface is mainly from electrostatic double-layer interaction and van der Waals forces. Also, there is enhanced removal of viruses at low soil moisture because viruses are sorbed on the air-water interface as well as the water-mineral (solid) interface. The sorbtion at the air-liquid interface may be greater than the air-solid interface. Sorption at the air-liquid interface is primarily controlled by virus surface hydrophobicity, solution ionic strength, and particle charge. Even though viruses sorbed at the liquid–solid interface can remain infective, viruses sorbed at the air-liquid are deformed by interfacial tension to the degree that the protein coat of the virus is disrupted and the virus is inactivated.

Although the adsorption of viruses to clay minerals and water films in soils can prevent their movement and even kill them, the potential desorption of the viruses from clay minerals means that viruses can reach the ground and surface waters that may be used by man. This, in addition to viruses that may be directly ingested from contaminated soils, indicates that soil-borne viruses are important human pathogens. Table 20.5 lists the names of some viruses as well as the diseases they cause.

20.12 Transmissible Spongiform Encephalopathies

Transmissible spongiform encephalopathies (TSE) are fatal, degenerative diseases of animals and humans characterized by abnormal limb movements, progressive dementia, and the development of sponge-like holes in brain tissues. The accumulation of an abnormal protease-resistant protein in the brain is associated with all TSE. The nature of the causative agent of TSE is still being debated. One theory is that the agent is a biologically active, self-replicating, infectious protein called a prion, which accumulates in and destroys brain tissue. A second theory is that the agent resembles a virus in that it exists as different strains and causes infective, transmissible diseases and possesses nucleic acids, which carry genetic information (Council for Agricultural Science and Technology 2000; Rabenau et al. 2001; Bartz et al. 2000). TSE induced diseases of humans include: Creutzfeldt-Jakob disease, fatal familial insomnia, Gerstmann-Straussler-Scheinker disease, new variant Creutzfeldt-Jakob disease, and Kuru. Major examples in animals include scrapie in sheep, chronic wasting disease in deer and elk, bovine spongiform encephalopathy in cattle (mad cow disease), and TSE in cats, monkeys, and mink. Transmission of TSE is believed to be mostly by ingestion of infected animal parts or transplanted by use of contaminated medical instruments.

The relationship of TSE to the soil environment is unknown. However, some questions require consideration. First, can TSE agents, specifically bovine spongiform encephalopathy (BSE), be introduced into the soil by the natural death of an infected animal or possibly by anthropogenic activities? How long could the TSE agents exist in the soil? Concern has been raised in the UK by scientific advisors that of the nearly 500,000 cattle that were culled, killed, and buried as a result of foot-and-mouth disease, some were likely to be also infected with bovine spongiform encephalopathy and could spread the agent via soil and groundwater. If this were the case then BSE would be classified as an incidental soil pathogen. TSE agents are found in brain and nervous tissue connected to the brain, and also in bone marrow (National Cattleman's Beef

Table 20.5 Selected soil-borne human pathogenic viruses and their properties

Pathogen(s) and disease	Distribution and residency	Gateway(s) and information on incidence, morbidity and/or mortality (IMM) if available	Comments and soil survival time if available
Adenovirus spp. Disease: respiratory illness, conjunctivitis, diarrhea	Worldwide, specific viral species are found in specific locations Residency: incidental	Fecal-soil-oral IMM: Almost everyone is infected at some point. Diarrheal disease blamed for at least six million deaths per year worldwide	Soil association with solid human waste. Most infections are mild. Serious infection possible in immunosuppressed
Arenavirus spp.*:* Lassa Fever virus*, A.* Junin*, A.* Machupo*, A.* Sabia*, A.* Guanarito and others Disease: hemorrhagic fever (general), Lassa fever, Argentine, Bolivian, Brazilian and Venezuelan hemorrhagic fevers	Worldwide. Each species of virus has a different associated rodent(s) governing its geographic location. Residency: transient	Respiratory and others IMM: The number of Lassa virus infections per year in West Africa is estimated at 100,000–300,000, with approximately 5,000 deaths. May also have a nematode vector	Human zoonosis with the definitive reservoir being rodents. Virus can enter soil from rodent urine, feces and saliva. Disturbance of feces, infected soil or nesting materials can aerosolize. Also spread by contact with contaminated surface
Astrovirus spp. Disease: diarrhea and gastroenteritis, mostly in children and immunosuppressed individuals	Worldwide Residency: incidental	Fecal-soil-oral IMM: Diarrheal disease blamed for at least six million deaths per year worldwide	Soil association with solid human waste. Most infections are mild. Serious infection possible in immunosuppressed
Caliciviruses spp. including *Hepatitis* E virus Disease: diarrhea	Worldwide Residency: incidental	Fecal-soil-oral IMM: Diarrheal disease blamed for at least six million deaths per year worldwide	Childhood diarrhea. Most adults may be immune. Hepatitis E is very dangerous in developing countries and has a 20% mortality rate for pregnant women
Hantavirus spp. including Sin Nombre, Puumala, Thailand, Prospect Hill, Khabarovsk, Thottapalayam, Tula, New York, Black Creek Canal, El Moro Canyon, Bayou and others Disease: HPS-Hantavirus Pulmonary Syndrome	Worldwide, specific virus species inhabit specific hosts in specific locations. Residency: transient	Respiratory and others IMM: Probably thousands of cases annually worldwide but quite variable	Human zoonosis with the definitive reservoir being rodents. Each species has a different associated rodent(s). Virus can enter soil from rodent urine, feces and saliva. Disturbance of feces, infected soil or nesting materials can aerosolize. HPS has a 40–60% fatality rate
Enterovirus poliovirus Disease: Polio	Worldwide Residency: incidental	Fecal-soil-oral, respiratory IMM: Disease eliminated in many parts of the world	Soil association with solid human waste. Known to survive 91 days in unsaturated sand and humid conditions. Known to survive 180 days in saturated sand and compost
Enterovirus hepatitis A Disease: hepatitis	Worldwide Residency: incidental	Fecal-soil-oral IMM: In countries with poor sanitation, most children infected by age 9	Soil association with solid human waste. Known to survive 91 days sand and humid conditions.
Enterovirus coxsackievirus-A Disease: diarrhea, hand-foot-and-mouth disease, respiratory infection	Worldwide Residency: incidental	Fecal-soil-oral IMM: Diarrheal disease blamed for at least six million deaths per year worldwide	Soil association with solid human waste. Known to survive 180 days in saturated sand and compost
Enterovirus coxsackievirus-B Disease: pleurodynia, aseptic meningitis, pericarditis myocarditis	Worldwide Residency: incidental	Fecal-soil-oral	Soil association with solid human waste. Known to survive 180 days in saturated sand and compost
Enterovirus echovirus Disease: diarrhea, aseptic meningitis	Worldwide Residency: incidental	Fecal-soil-oral IMM: Diarrheal disease blamed for at least six million deaths per year worldwide	Soil association with solid human waste. Can survive 3–33 weeks depending on soil environment
Norwalk virus Disease: acute viral gastroenteritis, diarrhea	Worldwide Residency: incidental	Fecal-soil-oral IMM: Diarrheal disease blamed for at least six million deaths per year worldwide	Soil association with solid human waste. Very little is known about this virus

Orthopoxvirus variola Disease: smallpox	Worldwide Residency: incidental?	Mostly direct human to human transmission, but some respiratory environmental transmission known. Humans are the only known reservoir IMM: Last case acquired outside of a laboratory was in Somalia in 1977	Variola virus is unlikely to survive for more than 48 h in environment. Virus recovered in scabs on infected corpses after 13 years
Rotavirus spp. Disease: diarrhea, gastroenteritis	Worldwide Residency: incidental	Fecal-soil-oral, respiratory(?) IMM: Kills 600,000 children worldwide annually. Causes 2.7 million cases of gastroenteritis in children under 5 each year in U.S.	Soil association with solid human waste. Major cause of death in the 3rd World

References: CDC (2002); Health Canada (2002), Rusin and others (2000), Toranzos and Marcos (2000), World Health Organization (1996)

Fig. 20.2 Areas endemic for *Coccidioides* (After Valley Fever Center for Excellence 2002)

Association and Cattleman's Beef Board 2001). This fact may be relevant to the possible natural transmission of the disease among cattle. Bone chewing is a relatively common trait of cattle worldwide and is believed to be related to dietary phosphorus deficiency, which is often associated with soils deficient in phosphorus. Decaying carcasses and bones are part of the surficial soil horizons and host a large array of microbes. If BSE can be transmitted among cattle by ingestion of contaminated bone material then it would also be classified as a transient soil pathogen.

20.13 *Coccidioides* Case Study

20.13.1 Habitat of *Coccidioides*

Coccidioides is a dimorphic soil-inhabiting fungus, an important human pathogen, and the etiological agent of coccidioidomycosis (Valley Fever).

Coccidioides grows in the upper (5–20 cm) horizons of soils in endemic areas (Fig. 20.2). This saprophytic phase of the fungus is characterized by branching, segmented hyphae that form a network of mycelium. As the fungus matures, arthroconidia, 2–5 μm in size, are formed as barrel-shaped, rectangular segments of the hyphae that can be easily separated by soil disturbance (natural or anthropogenic) and consequently be dispersed by the wind. Arthroconidia are also very buoyant and may be readily moved by sheet-wash water during rainstorms only to be concentrated in fine sedimentary material some distance from the initial growth site. Under suitable environmental conditions the arthroconidia can germinate to form new hyphae and mycelium, which can repeat the cycle. If the airborne arthroconidia are inhaled by an appropriate host (humans, animals, even reptiles), then the parasitic phase of *Coccidioides* is initiated (Fig. 20.3). In tissue the arthroconidia transform into spherules 10–80 μm in diameter that, when mature, are internally divided into endospores that are about 3–5 μm in diameter. The mature spherules then rupture and the endospores are released into the surrounding tissue and spread the infection locally, or at times, into other organs by disseminating outside of the respiratory system. The epidemiology and human impacts of the disease are discussed in Pappagianis (1980, 1988, 1999) and Galgiani (1993, 1999).

The parasitic phase may end with the death of the host or by the expulsion of spherules outside of living tissue by sputum, pus, exudates, or degradation of an infected carcass. If this occurs in an acceptable environment then the spherules (or endospores) will germinate and hyphae and mycelium will be developed and the saprophytic phase will again be initiated (Fiese 1958).

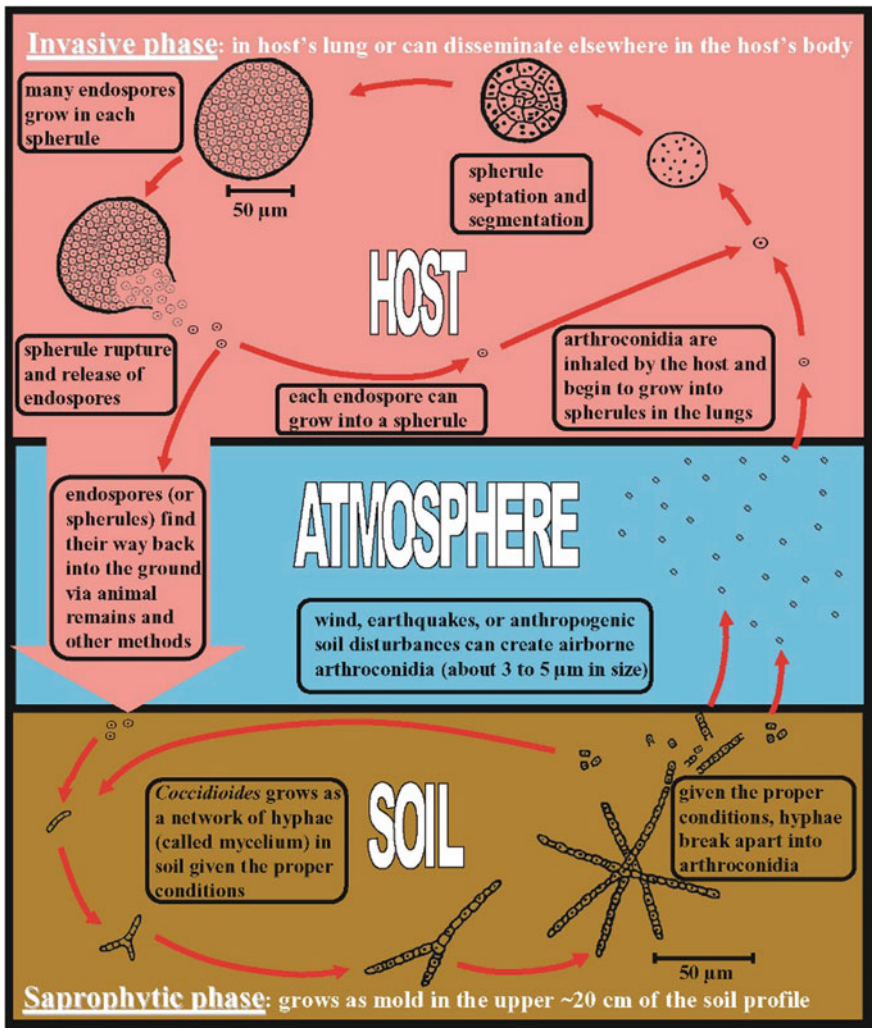


Fig. 20.3 The life cycle of *Coccidioides*

20.13.2 Distribution and Endemic Areas

In the United States *Coccidioides* is endemic in parts of Arizona, California, New Mexico, Nevada, Texas, and Utah (Fig. 20.2). An outbreak at Dinosaur National Monument in northeastern Utah in 2001 (Fig. 20.2) lies outside of the generally recognized endemic area. The site of this outbreak may represent a unique location where several things have come together, which include a favorable microclimate, to allow *Coccidioides* to survive in the soil. Outside of the United States it is endemic in parts of Argentina, Brazil, Colombia, Guatemala, Honduras, Mexico, Nicaragua, Paraguay, and Venezuela. With some exceptions endemic areas are generally arid to semi-arid with low to moderate rainfall (5–20 in.), mild winters, and long hot seasons (Fisher et al. 2000). In 1993 the CDC declared that coccidioidomycosis was epidemic in parts of California (Kern County) and also issued a warning to physicians nationwide to watch for the disease in patients who may have become infected while traveling in endemic areas. The CDC also listed coccidioidomycosis as an example of one of the important disease threats to the United States and has called for expanded studies of the disease (Bryan et al. 1994).

20.13.3 Habitat Criteria Essential for the Growth and Survival of *Coccidioides*

Laboratory and site-specific field studies have shown that many physical, chemical, climatic, and biological factors influence the growth of *Coccidioides* in the soil and the consequent development and deployment of arthroconidia. Many of the following factors are closely interwoven, and the influence on the presence or growth of *Coccidioides* by any combination of, or single factor, is an intricate balance that varies both in time (season) and in response to environmental changes at any given location.

Oxygen, carbon, nitrogen, phosphorus, sulfur, iron, and other trace elements along with water are necessary for the survival of *Coccidioides*. Furthermore, these raw materials

must be available in a physical and chemical environment suitable for *Coccidioides* to satisfy its specific biological functions required for life. Based on measurements and observations gathered from known sites where *Coccidioides* is present in the soil and also on laboratory experiments where *Coccidioides* is grown under controlled conditions, several general conclusions can be made about the habitat parameters required for its life processes and also those parameters that, while not essential for the survival of *Coccidioides*, are favorable for its existence.

20.13.3.1 Important Criteria

1. Most known occurrences are in hyperthermic or thermic aridisols or entisols with mean annual soil temperatures ranging from 15°C to over 22°C.
2. The presence of soils with textures that provide adequate pore space in the upper (20 cm) parts of the profile, for moisture, oxygen, and growing room is very important. Soils in known occurrence sites are mostly fine sand to silt (0.002- to 0.2-mm particle size) with less than 10% clay-sized (<0.002 mm) material (Fig. 20.1). Small amounts of clay foster water holding capacity, but large amounts of clay may be detrimental for *Coccidioides* growth. Smectite (a type of clay mineral) soils may be detrimental because their shrink and swell properties may provide room and water for bacterial growth that would compete with *Coccidioides*. Also, and perhaps, more important, they contain exchangeable cations that lower pH thereby enhancing bacterial growth at the expense of the growth of fungi.
3. The presence of some organic material is needed for carbon and nitrogen but in most known occurrences it is generally sparse, less than 2%. Large amounts of organic compounds may be detrimental because they would foster the growth of bacteria and other fungal species that would compete with *Coccidioides*.
4. Moisture is essential. Rainfall in endemic areas is generally seasonal with some areas receiving most of their precipitation in the winter months while precipitation in other areas may be split between winter rains and summer monsoons. In all cases, annual precipitation ranges from less than 250–410 mm.

20.13.3.2 Favorable Criteria

1. Many *Coccidioides* growth sites have soils with elevated salinity. High soluble salts may act as an inhibitor of microbial competitors. Measured values of soluble salts in soils from known occurrence sites are sodium, 8–75% greater in positive soils than in negative soils; calcium, 2–5 times greater in positive than in negative soils; potassium, 2–5 times greater in positive than in negative soils; sulfates, 2–5 times greater in positive than in negative soils; borates, 3–25 times greater in positive than in negative soils; and chlorides, 10–240 times greater in positive than in negative soils (Elconin et al. 1964)
2. Several *Coccidioides* growth sites are in soils derived from marine sedimentary rocks. These rocks often contain elevated amounts of salts, and when weathered, provide material with textures favorable for *Coccidioides* growth. Also the elevated salinity of these derived soils inhibits microbial competition.
3. The presence of borates in the soil profile may act as antiseptics for bacteria that are competitive with *Coccidioides*.
4. Any environmental factor that reduces competition with other fungal, bacterial, and/or plant species is favorable for *Coccidioides* growth.
5. Parent material derived from aeolian deposits is a good source for the development of soil with favorable textures.

Habitat modeling of the saprophytic phase of the *Coccidioides* life cycle is difficult due to the limited number of known growth sites. This confounds the establishment of statistical relationships of the physical, chemical, and biological habitat parameters. Therefore, habitat modeling is accomplished using analysis of the physical properties of known *Coccidioides* sites and a spatial fuzzy system. A spatial fuzzy system is a system of spatial variables where some or all of the spatial variables are described with fuzzy sets. The fuzzy system is capable of translating structured knowledge into a flexible numerical framework and processing it with a series of if-then rules called fuzzy associative memory (FAM) rules.

Fuzzy systems can describe nonlinear numerical processes with linguistic common sense terms and can handle differing precision and accuracy in the data. They produce models that can be repeated and updated easily.

Fuzzy system analysis was applied to each 30 × 30 m spatial cell over the study area, Organ Pipe Cactus National Monument, Arizona. The resulting product is a map (Fig. 20.4) depicting each cell's favorableness for hosting *Coccidioides* based on a scale of 0–1, which we define as its fuzzy habitat suitability index (FHSI). The fuzzy system allows modelers to change and update relationships among the variables as more is learned about *Coccidioides* habitat. An important property of this kind of analysis is that "what if" scenarios can be used to allow dynamic representation of climate related variables and may predict changes in habitat with changing climate.

Long-term climate fluctuations will undoubtedly have an effect on the distribution of *Coccidioides*. Changes in temperature and precipitation over time will directly influence soil characteristics critical to its growth and propagation. *Coccidioides* growth sites are believed to be relatively small and widely distributed throughout its endemic area and the fungus does not readily colonize outside of

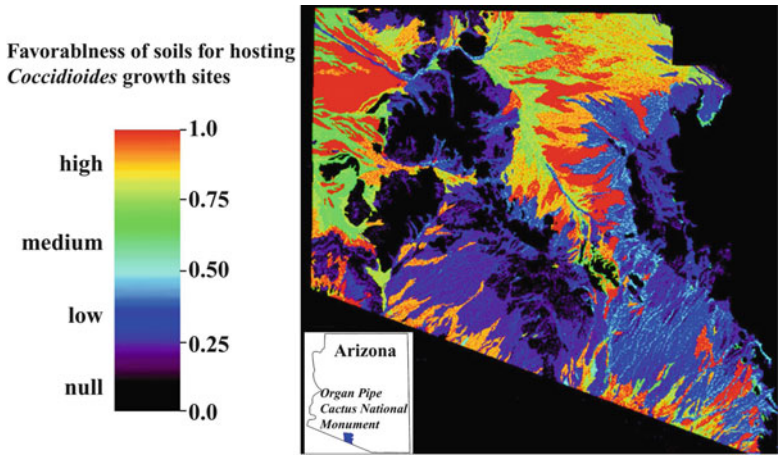


Fig. 20.4 The fuzzy habitat suitability index of *Coccidioides* measured as the favorableness of soils for hosting *Coccidioides*, Organ Pipe Cactus National Monument, Arizona

established growth sites. Laboratory studies show that *Coccidioides* is quite robust with respect to the physical and chemical factors of its habitat, but is very sensitive to competition from other microbes and vegetation. Therefore, climate change models that result in an increase in microbes in the soil profile and in the vegetation of a given area would result in decreased habitat for *Coccidioides* and scenarios that would reduce microbes in the soil profile and the vegetation in a given area may result in an increase in suitable habitat.

20.14 Soils and Emerging Diseases

Emerging infectious diseases may be defined as those that have newly appeared in a population and those whose incidence in humans has increased within the past two decades. A re-emergent disease is the reappearance of a known disease after a decline in incidence (Lederberg et al. 1992, pp. 34, 42). Many soils provide an ideal environment for the emergence of new infectious diseases due to their overall chemical and physical diversity, supply of essential nutrients for microbial growth, and their constantly evolving character in response to various soil forming factors, especially climate.

Disease emergence can be attributed to numerous factors such as the natural evolution and mutation of existing organisms; the spread of known diseases into new populations and/or new geographic areas; increasing human population; ecological changes that increase exposure of people to pathogens carried by insect or animal vectors; environmental changes that increase the exposure of people to contaminated dust, water, and soil; and exposure to as yet unknown pathogens (World Health Organization 2002). The re-emergence of known soil-borne pathogens may occur in response to the breakdown or overtaxing of existing public health infrastructures as a result of refugee circumstances or other types of major demographic changes.

Examples of recently emerging and re-emerging soil-borne pathogens are *Clostridium* spp. bacteria, which cause a variety of diseases and are probably a permanent soil resident, transmitted by the fecal-oral route and through skin trauma; *Listeria monocytogenes*, a bacterium, which causes listeriosis and is a permanent soil resident transmitted by contact with soil contaminated with infective animal feces and also by inhalation of the organism; Sin Nombre virus, a *Hantavirus*, which causes hemorrhagic fever, is a transient soil resident, and is transmitted by inhalation of dust containing aerosolized rodent urine and feces; *Rotavirus* spp., which causes diarrhea and enteritis, an incidental or less commonly, transient soil resident, transmitted by the fecal-soil-oral route also by the fecal-respiratory route; *Coccidioides*, a fungus, which causes coccidioidomycosis, a permanent soil resident, transmitted by inhalation of *Coccidioides* arthroconidia (Bryan et al. 1994); and variant Creutzfeldt-Jakob disease, a TSE caused by prion infection, outbreaks in the UK in the late 1990s, possibly(?) a transient or incidental soil resident (Lederberg et al. 1992, Table 2.1, p. 36; World Health Organization 1998).

Antimicrobial resistance, which is a natural consequence of the adaptation of microbes to exposure of drugs designed to kill them, may also cause re-emergence of infectious diseases. Even though resistance to antimicrobial agents is an irreversible, natural, and evolutionary process, it is exacerbated by several human activities including overuse (in developed countries) and under use (in developing countries) of antibiotic drugs; discharges of wastes from pharmaceutical production plants; disposal into landfills and sewage of wastes from common antibacterial household products (soaps, over-the-counter drugs, cosmetics, cleaning supplies, etc.); introduction into the human food chain by agricultural use of antibiotics for disease and pest control on plants and also by use of antibiotics on many types of livestock for therapeutic reasons and as growth enhancers; disposal of waste products from agricultural operations;

disposal of sewage sludge in landfills and by application directly onto the land surface; and disposal of all types of household and industrial garbage into landfills (Standing Medical Advisory Committee 1998). Antibacterial drugs have received the most attention in regards to antimicrobial resistance; however, resistance is also developing to antiviral and antifungal drugs.

The ability of pathogenic microbes to respond and adapt quickly to new environmental conditions is fundamental to the development of antimicrobial resistance. Both the disposal of waste material in landfills and application of sewage sludge to the land surface create chemical and biological modifications of the natural soil environment in any given place and provide new environments that can foster microbial genetic change. In developed countries modern sanitary landfills and municipal sewage plants are closely regulated and designed to limit the escape of chemical and biological toxins. Homes in these countries, not connected to municipal sewage systems, in most cases utilize septic systems with leach fields that rely on the soils for sewage treatment. Nonetheless some sewage sludge and landfill leachates contain a variety of bacterial, parasitic, and viral pathogens derived from food waste, domestic animal feces, disposable diapers, and garden waste.

In developing countries raw sewage and untreated waste of all types are commonly disposed of directly into soils and at times are added directly to soils as fertilizer. Under these circumstances some bacterial pathogens in sludge and sewage may fail to adapt and die out, whereas others will adapt to the new environment and experience new growth.

Also of concern is the presence of residual amounts of antimicrobial agents (pharmaceuticals, heavy metals, toxic chemicals) that may select for the growth of new bacterial forms that are resistant to various antibiotics. Non-biological dispersal of pathogens from landfills are mainly by water (both surface runoff and groundwater in the vadose zone) and wind. Biological dispersal may be due to birds, rodents, insects, and humans.

20.15 Interconnections: Geology/Soil/Pathogenic Microbes

The importance of the soil environment for hosting human pathogens was recognized over 2,000 years ago by Hippocrates who suggested that a physician, when arriving at an unfamiliar town, should examine the winds, sun aspect, sources of water, and "... the soil too, whether bare and dry or wooded and watered, hollow and hot or high and cold." (Jones 1923). That wisdom has enormous room for development in the twenty-first century.

The study of the ecological habitats of soil microbes (both friendly and pathogenic) has been hampered for centuries by the inability to see, measure, count, and weigh organisms too small for the human eye to distinguish, especially *in situ*. It has also been hampered by the inadequate exchange of ideas and approaches among the wide diversity of scientific disciplines studying soils and microbes. Complete scientific descriptions of soil attributes, profiles, and classifications are rarely, if ever, given in the medical and microbiological literature focused on site-specific occurrences of soil-borne human pathogens. This makes it nearly impossible to conduct followup studies or further experiments in the same area or soil type, or to extrapolate results to other locations and studies. Commonly the only mention of the terrestrial environment of pathogenic organisms is described by phrases such as "soil" or "soil contaminated with bird feces," or "moist anaerobic soil," when describing the organism's habitat. This is not meant to fault prior research but instead to underline the need for multidisciplinary efforts. Many scientists are not knowledgeable about soils or soil attributes that affect microbes. Therefore, important geologic-soil-pathogen-process relationships are overlooked in many studies.

Geological features and processes are inherent in many soil attributes, which are, in turn, important controls over microbial activity and existence. For example, the abundance of ferromagnesian minerals and feldspars in a parent rock will determine the abundance and types of clay minerals formed (given the right climatic conditions) in soils by weathering processes. The presence of clay minerals strongly influences soil water potential, soil aggregation and pore size, microbe movement, virus adsorption, and the types of microbes present in any given soil.

Infection by soil-borne pathogens can be prevented or reduced by disrupting their life cycle. However, to do this a complete understanding of the infectious cycle is necessary to determine where interdiction will be most effective. For example, interdiction of the life cycle of soil-borne enteric pathogens is accomplished by the use of proper disposal and sanitation measures of human wastes. Another example is disruption of the hantavirus cycle of infection by controlling rodents in enclosed areas, thereby reducing or preventing exposure to contaminated aerosols from rodent feces and urine. These and similar examples require basic research into all aspects of the life processes and ecology of soil-borne pathogens and their interaction with the physical, chemical, and biological attributes of their habitat. These studies are best accomplished in the field and on-site using, whenever possible, noninvasive methods, some of which are reviewed by Madsen (1996). In the best circumstances, studies of soil-borne human pathogens would include a soil scientist familiar with field measurements and determinations of soil properties and classification. At the minimum,

soil pathogen collection sites should be precisely located, soil textures should be determined, sand-silt-clay proportions should be estimated, organic content determined, hydrologic setting described, geomorphologic setting determined, pH and salinity (electrical conductivity) measured, and vegetation type and density described. These observations would go a long way to address Hippocrates' counsel to look at the soils.

Infectious diseases are a major cause of human suffering and mortality and account for an estimated 13 million deaths worldwide each year (World Health Organization 1999), and that number is expected to grow. As indicated in previous sections, soil-borne human pathogens are important contributors to those numbers. Drug-resistant microbes are increasing at a dramatic rate and large urbanized areas in developing countries with dismal health care and sanitary facilities are magnets for displaced people. Deteriorating natural environments through urbanization, deforestation, and pollution of soils and waters coupled with the ease of human travel ensures breeding places and rapid transportation for many infectious agents. Increased understanding of the life cycles of pathogenic soil-borne microbes, the ecology of their habitats, and the environmental gateways they utilize for infectious transmissions will help break these cycles of infection. These problems are complex in character, global in distribution, and applicable to every human being. Their solutions are contingent on scientists from many disciplines working together to study the attributes and processes of complex soil ecosystems and communicating their results to public health officials.

See Also the Following Chapters.

Chapter 15 (Bioavailability of Elements in Soil) • Chapter 18 (Geophagy and the Involuntary Ingestion of Soil) • Chapter 19 (Natural Aerosolic Mineral Dusts and Human Health) • Chapter 29 (Investigating Vector-Borne and Zoonotic Diseases with Remote Sensing and GIS)

Further Reading

References cited and recommended readings are marked with an *. For references cited located at Web sites, see Appendix B.

Bertucci JJ, Sedita SJ, Lue-Hing C (1987) Viral aspects of applying sludges to land. In: Rao VC, Melnick JL (eds) Human viruses in sediments, sludges, and soils. CRC Press, Boca Raton

*Brady NC, Weil RR (2002) The nature and properties of soils. Prentice Hall, Upper Saddle River

Bartz, Jason and others (2000) Transmissible spongiform encephalopathies in the United States, Council for Agricultural Science and Technology Report no. 136, p 36

*Brown AD (1990) Microbial water stress physiology, principles and perspectives. Wiley, New York

Bryan RT, Pinner RW, Gaynes RP, Peters CJ, Aguilar JR, Berkelman RL (1994) Addressing emerging infectious disease threats: a prevention strategy for the United States executive summary. CDC MMWR RR-5(43):1–18

Council for Agricultural Science and Technology (2000) Transmissible spongiform encephalopathies in the United States: Task Force Report no. 136, p 36

*Cowan G, Pines D, Meltzer D (1999) Complexity, metaphors, models, and reality: Santa Fe Institute, Studies in the Sciences of Complexity, Proc., vol XIX. Perseus Books, Cambridge, MA

*Coyne MS (1999) Soil microbiology: an exploratory approach. Delmar Publishers, Albany/New York

Elconin AF, Egeberg RO, Egeberg MC (1964) Significance of soil salinity on the ecology of *Coccidioides immitis*. J Bacteriol 87 (3):500–503

Fiese MJ (1958) Coccidioidomycosis. Charles C Thomas, Springfield

Fisher FS, Bultman MW, Pappagianis D (2000) Operational guidelines for geological fieldwork in areas endemic for coccidioidomycosis, U. S. Geological Survey Open File Report 00–348

Flake GW (1998) The computation beauty of nature: computer explorations of fractals, chaos, complex systems, and adaptation. MIT Press, Cambridge, MA

*Galgiani JN (1993) Coccidioidomycosis. West J Med 159:153–171

Galgiani JN (1999) Coccidioidomycosis: a regional disease of national importance, rethinking approaches for control. Ann Int Med 130(4) (part 1):293–300

Gerba CP (1987) Transport and fate of viruses in soils: field studies. In: Rao VC, Melnick JL (eds) Human viruses in sediments, sludges, and soils. CRC Press, Boca Raton, pp 141–154

Goodman RA (ed) (1994) Emerging infectious diseases, update: coccidioidomycosis—California, 1991–1993. CDC MMWR 43 (23):421–423

Griffin DM (1972) Ecology of soil fungi. Syracuse University Press, New York

Harris RF (1981) Effect of water potential on microbial growth and activity. In: Parr JF, Gardner WR, Elliott LF (eds) Water potential relations in soil microbiology. Special publication no. 9. Soil Science Society of America, Madison, pp 23–95

Jones WHS (1923) Hippocrates, airs waters places, vol 1. Harvard University Press, Cambridge, MA

*Kosko B (1992) Neural networks and fuzzy systems. Prentice Hall, Englewood Cliffs

*Lederberg J, Shope RE, Oaks SC Jr (eds) (1992) Emerging infections, microbial threats to health in the United States. National Academy Press, Washington, DC

*Madsen EL (1996) A critical analysis of methods for determining the composition and biogeochemical activities of soil microbial communities in situ. In: Stotzky G, Bollag J-M (eds) Soil biochemistry. Marcel Dekker, New York, pp 287–370

*Maier RM, Pepper IL, Gerba CP (eds) (2000) Environmental microbiology. Academic Press, New York

Matson JV, Lowry CL, Yee MC, Whitworth ME (1987) Physical and chemical characteristics of sediments, sludges, and soils. In: Rao VC, Menick JL (eds) Human viruses in sediments, sludges, and soils. CRC Press, Boca Raton

*Miller RW, Gardiner DT (2001) Soils in our environment. Prentice Hall, Upper Saddle River

NASA (2001) Microbes and the dust they ride in pose potential health risks, Earth Observatory Release no. 01–120. NASA, Washington, DC

Pappagianis D (1980) Epidemiology of coccidioidomycosis. In: Stevens DA (ed) Coccidioidomycosis. Plenum Medical Book Company, New York, pp 63–85

*Pappagianis D (1988) Epidemiology of coccidioidomycosis. Curr Top Med Mycol 2:199–238

Pappagianis D (1999) *Coccidioides immitis* antigen, letter to the editor. J Infect Dis 180:243–244

Pappagianis D, Sun RK, Werner SB, Rutherford GW, Elsea RW, Miller GB Jr, Egleston MD, Hopkins RS (1993) Coccidioidomycosis—United States, 1991–1992. JAMA 269(9):1098

*Paul EA, Clark FE (1996) Soil microbiology and biochemistry. Academic Press, New York

Rabenau HF, Cinatl J, Doerr HW (eds) (2001) Prions: contributions to microbiology, vol 7. Karger, New York

Raloff J (2001) Dust storms ferry toxic agents between countries and even continents. Sci News 160:218–220

Rusin P, Enriquez CE, Johnson D, Gerba CP (2000) Environmentally transmitted pathogens. In: Maier RM, Pepper IL, Gerba CP (eds) Environmental microbiology. Academic, London, pp 447–489

Sim Y, Chrysikopoulos CV (2000) Virus transport in unsaturated porous media. Water Resour Res 36(1):173–179

Sobsey MD, Shields PA (1987) Survival and transport of viruses in soils: model studies. In: Rao VC, Melnick JL (eds) Human viruses in sediments, sludges, and soils. CRC Press, Boca Raton, pp 155–177

Staley JT (2002) A microbiological perspective of biodiversity. In: Staley JT, Reysenbach A-L (eds) Biodiversity of microbial life. Wiley, New York

*Toranzos GA, Marcos RP (2000) Human enteric pathogens and soilborne disease. In: Bollag J-M, Stotzky G (eds) Soil biochemistry, vol 10. Marcel Decker, New York, pp 461–481

Wallwork JA (1970) Ecology of soil animals. McGraw-Hill, London

World Health Organization (1996) Foodborne, waterborne, and soilborne diseases. World Health Report, Fighting Disease Fostering Development, World Health Organization, Geneva

Animals and Medical Geology

21

Bernt Jones

Contents

21.1 Introduction

It has been recognized for a very long time that animals can become sick after grazing in certain areas of the world. Eventually, it was determined that local geochemical anomalies were responsible, but in some cases the exact mechanisms behind the deficiencies and intoxications have only been known for the last 50–70 years. Local geochemical anomalies affect both domestic and wild herbivorous animals living in an area; however, wild animals will be especially susceptible to deficiency or intoxication due to the fact that they normally are totally dependent on feeds growing in their territory.

Problems associated with restricted access to suitable feeds are seen in domestic animals in developing countries or among animals that have been reared under less intensive conditions, and where farmers rely on locally produced feeds. Extreme weather conditions, especially drought, affect the availability of pasture and feed in areas with low groundwater tables and sensitive soils. The availability of nutrients is also affected by the seasonal development of forage plants eaten by the animals. The protein content of mature plants is often low, as is the amount of easily digestible carbohydrates and other important nutrients, but the fiber content is higher than in earlier stages of plant development. This seasonal shortage of available nutrients can be circumvented by migration between different areas and is believed to be one of the driving forces behind the seasonal migrations normally seen in many wild African herbivores. In some areas, this behavior is restricted, as some wildlife reservations are fenced to protect the animals from poachers and grazing competition from domestic animals (Maskall and Thornton 1996). Migration between different grazing areas can also be applied to domestic animals and is a common practice for nomadic animal husbandry in Africa and continental parts of Asia. The nomadic behavior of caribou and semidomestic reindeer in the subarctic is primarily caused

B. Jones (✉)
Department of Clinical Sciences, Faculty of Veterinary Medicine and Animal Science, Swedish University of Agricultural Sciences, Uppsala, Sweden
e-mail: Bernt.Jones@slu.se

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_21,

by climatic factors such as snow and ice, which can limit the availability of feed.

The increasing interest in organic farming in some European countries and North America exposes animals on such farms to higher risks of deficiency or dietary imbalance because of the use of feeds produced locally without conventional fertilizers. Farmers who are aware of these risks can overcome them by growing specific feeding plants selected to accumulate or exclude particular elements known to cause nutritional imbalances.

In the intensive type of animal husbandry seen in many parts of the world today, especially in Europe and North America, the effects of local geochemical anomalies are less apparent, as the content of nutrients in the rations fed to animals is controlled, and appropriate supplements or feedstuffs are utilized to compensate for potential deficiencies or excesses.

Generally, the metabolic effects of deficient or excessive intakes of nutrients due to the local geochemistry are the same in animals as in humans; that is, a change in activity of certain important enzymes is the major effect that compromises the health of animals. The symptoms exhibited by various animal species or humans suffering from the same pathological processes will therefore be identical or similar.

21.1.1 Recognition of Problems

Pathological changes caused by mineral imbalances, deficiency, or intoxication due to local geochemical anomalies are often difficult to detect in the living animal, especially in the free-living wild animal. The diffuse and ambiguous signs of these changes include retarded growth, decreased fertility, and decreased immunological capacity. In many cases, the only signs are suboptimal growth or low reproduction. Diagnosis is further complicated by interactions among the various elements normally present in the soil. These interactions affect nutrient uptake by plants, making it difficult to evaluate results obtained from soil and plant analyses with regard to the nutritional value for the grazing animal or to estimate the animal's intake of required nutrients. Interactions also occur among various elements in the animal itself, both in the gastrointestinal tract and in different tissues of the body (see Fig. 21.1). The gastrointestinal interactions are probably more important in herbivores than in carnivores, as the longer passage time through the tract and its larger volume in herbivores provide greater opportunities for chemical reactions to occur. Often, extended analyses of blood and other tissues such as liver and kidney are needed to arrive at a definite diagnosis (see Sect. 21.3.10).

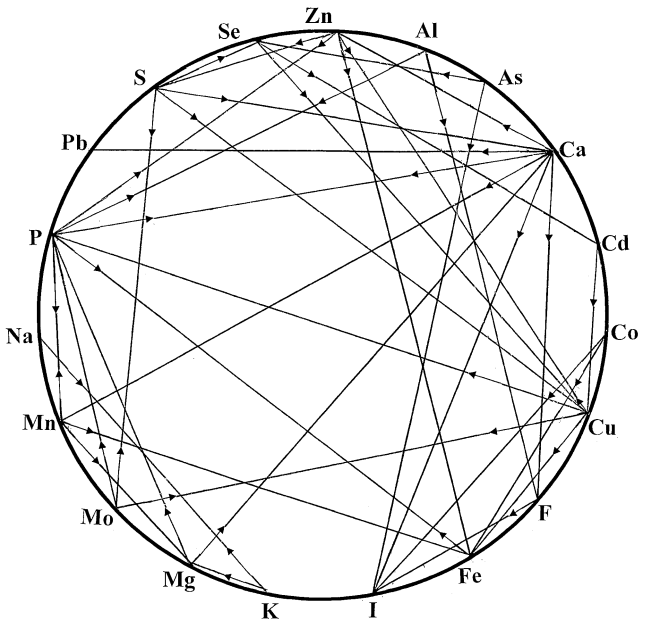


Fig. 21.1 Important interactions between various elements in animals. Note: Interactions can occur in the gastrointestinal tract of the animal or on a cellular level in the tissues. *Source:* Modified after Jacobson et al. 1972)

21.1.2 Mineral and Trace Element Availability

Apart from general factors such as soil composition and pH, the plant species and stage of development will affect root uptake and hence the mineral content found in the plant. Alkaline soils will have more available molybdenum and selenium because these elements are present as anions, whereas most other metals are present as cations and their availability is favored by lower pH. The availability of minerals in soil eaten by animals is often difficult to estimate from common chemical analyses of mineral concentrations, as the actual gut absorption often is very low, depending on the chemical form of the mineral. The different types of interactions presented in Fig. 21.1 will also be of relatively large importance in these situations. Minerals taken up by different plants and present in plant tissues are generally more available as the root uptake requires solubility in soil water or intracellular water. Dust or other soil contaminations on plant surfaces, especially on pilose or sticky structures, will be as unavailable to absorption in the gut as pure soil eaten by an animal. Involuntary soil intake will be much higher in animals grazing hard and close to the ground, such as cattle and sheep, than goats and many wild ruminants that to a great extent feed on shrubs, bushes, and tree leaves (browse). Especially in situations of feed shortage, soil intake can be high (several kilograms per day in cattle).

Table 21.1 Nutrient requirements for maintenance for calcium and phosphorus for some domestic species

Animal	Calcium	Phosphorus	Ref.
Lactating dairy cows	4.8–8.5	3.1–5.3	NRC (2001)
Growing calves	3.2–5.8	2.6–3.4	NRC (2001)
Beef cattle	16–36	13–24	NRC (1996)
Sheep (60 kg)	2.3	2.1	NRC (2007b)
Goat (50 kg)	2.1	1.7	NRC (2007b)
Swine	5.6–10	4.4–7.8	NRC (1998)
Horses (400 kg)	16	11	NRC (2007a)
Dogs	6.6	4.9	NRC (1985)

Note: The amounts are expressed as g/kg d.m. feed per day

Table 21.2 Amounts of manganese required for normal development and health in some domestic animal species and amounts causing adverse effects

Animal	Requirements	Adverse effects	Ref.
Dairy cattle	20	1,000	NRC (2001)
Beef cattle	40	1,000	NRC (1996)
Sheep	16–20	2,000	NRC (2007b)
Goat (50 kg)	14–25	2,000	NRC (2007b)
Horses (400 kg)	5	400	NRC (2007a)
Swine	2–4	400	NRC (1998)
Dogs	1.4[a]		NRC (1985)
Chicken	30–60	2,000	NRC (1994)

Note: The amounts are expressed as mg/kg d.m. feed per day
[a]mg/1,000 kcal ME

21.1.3 Mineral and Trace Element Requirement and Excess

The needs for mineral and trace elements by common domestic animal species have been established in experiments using normal feeding procedures or specialized synthetic or semisynthetic feeds (see Tables 21.1, 21.2, 21.3, 21.4, 21.5 and 21.6). An animal's growth rate and normal development are followed in these types of experiments, as well as the health and reproductive functions of the animal. In many cases, the nutrient requirements are influenced by the selective breeding of animals to provide increased growth rates or improved milk or egg production. The complex interrelationships among minerals and trace elements and other feed components make it difficult to determine if nutrient requirements are being met or if adverse effects due to excess nutrients could be expected when unconventional or extreme feeding regimes are used. Some of the known requirements for various minerals and trace elements are discussed in Sect. 21.3. The mineral and trace element requirements of wild or exotic domestic species are not well known, and neither are the amounts needed to cause deleterious effects in these animals. Extrapolations from known needs of related domestic species should be done with great caution as many wild species have evolved in specific areas and have adapted to the geochemical situations in those particular habitats.

21.1.4 Salt and Mineral Licks

Both domestic and wild animals are regularly seen eating soil or earth. Under conditions of overgrazing or a shortage of feed, soil eating is believed to correct nutritional deficiencies. In certain situations, animals are exploiting natural sources of sodium- or phosphorus-rich minerals for this purpose. Other minerals are actively sought by animals, but the mechanisms behind this behavior are not known. It is a common belief that animals, and humans, will correct deficiencies in their mineral and trace element supply by eating soil (*i.e.*, pica). This behavior is seen in many cases of nutrient deficiencies, but no known mechanism induces or regulates it. More probable in animals is that this knowledge is passed down from mother to offspring as an important factor in the utilization of special sites with favorable mineral compositions. One illustrative example of this is the utilization of minerals found in caves in Kenya that are actively mined by elephants living in the area (Bowell et al. 1996). The elephants use their tusks to mine the veins of calcite–zeolite in the roofs and walls of the caves; some of the rocks are eaten by the animals, but some are left on the cave floor. The environment on the cave floor (*e.g.*, water and debris from animals) induces the formation of salts on the surfaces of the rocks. This cave salt is utilized by local tribesmen for their cattle, and it is also used by other wild animals, such as baboons (*Papio cyancephalus*) and leopards (*Pantera pardus*). Bushbucks (*Tragelphus scriptus*) also seek the formed salt deposits in the caves.

21.2 Species and Breed Differences

For a number of reasons, important differences can be seen between monogastric and ruminant species in the metabolic handling of several elements that cause deficiency or intoxication. One reason is the simple fact that the pH of the forestomachs of ruminant species ranges from 6 to 7, and the pH of the abomasum, the true stomach of ruminants, is 2–3, which renders most metallic elements less available for absorption as compared to most monogastric species, for which the stomach pH varies from 1 to 2. Another important reason for species differences is the simple chemical reactions that can occur in the forestomachs of ruminants. The anaerobic fermentation that takes place in the forestomachs can produce valence stages and complex compounds that are not available for intestinal absorption, as most minerals and trace elements are absorbed in cationic form. For a more comprehensive description of the digestive

Table 21.3 Nutrient requirements for copper and amounts causing toxicity or other adverse effects in domestic animals

Animal	Requirements	Toxic or adverse effects	Ref.
Dairy cattle	10	100	NRC (2001)
Beef cattle	10	100	NRC (1996)
Sheep (60 kg)	8–11	25	NRC (2007b)
Goat (50 kg)	15–25	60	NRC (2007b)
Swine	3–6	250	NRC (1998)
Horses	10	–	NRC (2007a)
Dogs	0.8[a]	–	NRC (1985)
Chicken	6–8	500	NRC (1994)

Note: The amounts are expressed as mg/kg feed d.m.
[a]mg/1,000 kcal ME

Table 21.4 Nutrient requirements for zinc and amounts causing adverse effects in domestic animals

Animal	Requirements	Adverse effects	Ref.
Dairy cattle	23	1,000	NRC (2001)
Beef cattle	30	500	NRC (1996)
Sheep (60 kg)	20–33	300	NRC (2007b)
Goats (50 kg)	12–20	–	NRC (2007b)
Growing swine	50–100	1,000	NRC (1998)
Breeding swine	50	–	NRC (1998)
Horses	40	–	NRC (2007a)
Dogs	9.7[a]		NRC (1985)
Chicken	35–65	1,200	NRC (1994)

Note: The amounts are expressed as mg/kg feed d.m.
[a]mg/1,000 kcal ME

physiology of animals see, for example, Chapter 14 in Sjaastad et al. (2003).

In some cases, breed differences can exist within the same domestic species due to the development of breeding animals having special properties desirable in specific situations. The very fast growth rate seen in many modern domestic animals is a breeding effect that can affect an animal's sensitivity to insufficient or excessive intakes of minerals and trace elements. Because many modern breeds have a high demand for nutrients and therefore a high feed intake, they may not necessarily meet their mineral and trace element requirements. This is especially obvious in very young animals during the suckling period if they feed only on milk that is poor in iron. The available iron is insufficient to meet the requirements for normal hemoglobin formation, thus causing anemia, which, in turn, often also leads to increased susceptibility to infection.

In most cases, attributing pathological conditions to local geochemical anomalies is complicated. The clinical picture and postmortem examination are often equivocal. Support from clinical chemical analyses of blood and other tissues from the living animal and especially chemical analyses from tissues (e.g. liver and kidney) obtained during a postmortem examination are necessary to obtain a definitive diagnosis.

21.3 Specific Elements

This section focuses on effects seen in normal environmental situations in both wild and domestic animals and only briefly discusses effects observed in areas polluted by mining or other industrial activities. The minerals calcium and phosphorus are discussed first, followed by the trace elements in order of increasing atomic weight.

21.3.1 Phosphorus and Calcium

The metabolism of phosphorus and the metabolism of calcium are so interrelated that they are discussed together. Utilization of these two elements is profoundly affected by the amount of vitamin D present in feeds consumed by the animal. Disturbances in the metabolism of phosphorus and calcium that are seen in young growing animals can affect bone formation when cartilage is being gradually replaced by bone tissue containing hydroxyapatite, a calcium phosphate compound that makes bone hard and mineralized. If this process is disturbed, both bone growth and function are affected. A well-known pathology associated with this mineralization is rickets, a disease known in both animals and humans since antiquity. Often, rickets is caused by a vitamin D deficiency. A typical case of rickets shows enlargement of the junction between the ribs and costal cartilages (costochondral junction) as well as the growth plates (metaphyses) of the long bones. The stability of the bone is reduced, resulting in fractures or bending of the legs, thus giving the animal the tell-tale appearance of rickets. In adult animals, where bone growth is completed, a similar defect is called *osteomalacia*.

Direct toxic effects of an excessive intake of calcium are not seen as the balance between calcium and phosphorus is crucial for the biological effects. An excessive intake of calcium, however, severely affects the availability of several other minerals and trace elements, as shown in Fig. 21.1. High phosphorus intake also interferes with the metabolism of many other elements, as seen in the figure. Normal calcium and phosphorus requirements for common domestic species are presented in Table 21.1; these amounts are relevant in situations where the calcium–phosphorus ratio is balanced (ideally, 1–1.2).

Endemic areas deficient in phosphorus are found in South Africa and Latin America; cattle grazing on grassland in these areas often have a very low intake of both phosphorus and calcium. Clinical signs of phosphorus and calcium

Table 21.5 Nutrient requirements for selenium and amounts causing adverse effects in domestic animals

Animal	Requirements	Adverse effects	Ref.
Dairy cattle	0.3	5	NRC (2001)
Beef cattle	0.1	2	NRC (1996)
Sheep (60 kg)	0.1–0.2	5	NRC (2007b)
Goat (50 kg)	0.1–0.2	5	NRC (2007b)
Swine	0.1–0.3	20	NRC (1998)
Horses	0.1	5	NRC (2007a)
Dogs	0.11	—	NRC (1985)
Chicken	0.1–0.15	—	NRC (1994)

Note: The amounts are expressed as mg/kg feed d.m.

Table 21.6 Nutrient requirements for iodine for some domestic species

Animal	Requirements	Adverse effects	Ref.
Dairy cattle	0.5		NRC (2001)
Beef cattle	0.5	50	NRC (1996)
Sheep (60 kg)	0.1–0.8	50	NRC (2007b)
Goat (50 kg)	0.5–0.8		NRC (2007b)
Swine	0.14		NRC (1998)
Horses	0.4		NRC (2007a)
Chicken	0.3		NRC (1994)

Note: The amounts are expressed as mg/kg feed d.m.

deficiency are more apparent in young animals than adults. Subnormal growth, low reproduction, and pica are observed before the bone pathology (osteomalacia) becomes apparent, in both young and adult animals. Deficiencies can also be seen in pigs in the form of poor growth, gait disturbances, lameness, bone deformation, and even fractures. The fractures occur in the vertebrae, which are more prone to demineralization than the long bones of the legs.

A special situation related to excessive calcium intake is calcinosis, which is seen in areas where plants producing substances related to vitamin D are growing. In New Guinea, the European Alps, Argentina, Brazil, and Florida, enzootic calcinosis with calcium deposits in soft tissues has been described. In the Americas solanaceous plants produce compounds that can mimic vitamin D activity in the gastrointestinal tract, thus promoting an excessively high absorption of calcium and phosphates and disturbing the normal homeostatic regulation of these elements. In Argentina and Brazil, this condition (known as *enteque seco*) is caused by ingestion of leaves from the shrub *Solanum malacoxylon*, and in Florida a large ornamental plant (*Cestrum diurnum*) is the cause. In Central Europe, calcinosis is caused by consumption of the pasture grass *Trisetum flavescens*. The increased uptake of calcium results in calcification of soft tissues in the affected animals, and the often massive deposits of calcium seen in the walls of major arteries, especially the aorta, will eventually kill the animal due to disturbed blood circulation and heart failure.

21.3.2 Aluminum

Only trace amounts of aluminum are found in most biological organisms, although this element is abundant in most soils; however, some exceptional tropical plants do accumulate aluminum, which can reach levels of >1 g/kg of plant (dry matter, d.m.). Aluminum can be toxic to plants as well as fish and other aquatic biota, but direct toxic effects of aluminum on livestock are unlikely at natural environmental levels. High levels of ingested aluminum will, however, decrease the availability of nutrients such as iron, phosphorus, and, to some extent, calcium (see Fig. 21.1). Most of this aluminum would be present as external contamination on the plant material eaten by the animals, especially in a dry and dusty environment. The high levels of aluminum and iron seen in some acidic soils can form insoluble phosphate complexes that produce secondary phosphorus deficiency in animals. This could result in grass tetany and osteomalacia in exposed cattle. The amounts needed to cause such problems are estimated to be 0.5–8 g/kg plant material (d.m.).

21.3.3 Fluorine

Some species might need fluorine as an essential trace element, but it is generally known to be toxic to both humans and animals. Small amounts of fluorine absorbed in the gastrointestinal tract will quickly form CaF, which is incorporated into bone and teeth, resulting in more stable and less reactive bone tissue. Signs of fluorine toxicity were described very early (around the year 1,000) in Iceland, where cattle grazed on grass contaminated with fallen volcanic ashes. Areas with volcanic activity are particularly prone to excessive fluorine exposure due to external contamination of pasture plants. Parts of the North African coastal plain have been known for centuries to be areas of endemic fluorine toxicity. Often, drinking water is the source of excessive fluorine intake, and endemic chronic fluorosis has been described in domestic ruminants and horses in Australia, India, Turkey, and several parts of Africa due to high levels of fluorine in their drinking water (see also Chap. 13, this volume).

Poultry is more resistant to the toxic effects of fluorine, and reports suggest that they are able to tolerate twice as much as mammals that only can tolerate 80–100 µg/g d.m. in their feed (Mertz 1987). In water the estimated safe level is < 5 ppm. A high intake of aluminum or calcium (see Fig. 21.1) will protect species from the toxic effects of excessive fluorine in their feed or drinking water.

Clinical signs of fluorine toxicity are evident primarily in the teeth and bone. Teeth still under development are most sensitive to high fluorine intake and become mottled and discolored. More severe indications are modifications of size

and shape of developing teeth and decreased strength resulting in fractures even after limited mechanical stress. Also, bone will show changes in size and shape as a result of fluorine toxicity, especially in growing animals. Lameness can also be seen as a result of bone pathology caused by excessive fluorine intake.

Most plants have low concentrations of fluorine, but external contamination is the primary route of toxic exposures; however, in South Africa, some plants do produce fluoroacetic acid, which can poison the grazing sheep. This fluoroacetic acid is converted *in vivo* to fluorocitrate, which specifically inhibits activity of the enzyme aconitase, which is necessary for the metabolism of citrate in the energy-producing citric acid cycle (or Krebs cycle) vital to all living cells. Fluoroacetate was used earlier as a rodenticide due to its potent toxic effect.

21.3.4 Manganese

Concentrations of manganese in plant material are highly variable, and the effects of a deficiency due to low manganese intake can be seen in ruminants, swine, and poultry fed normal feeding plants. The risk for low manganese intake is greater if the land used for feed production or grazing is excessively limed. Symptoms in affected animals are changes in lipid and carbohydrate metabolism and impaired growth and infertility, depending on the duration of the deficiency and the age of the animal. These symptoms are caused by the decreased activity of manganese-containing enzymes such as arginase, manganese–superoxide dismutase, and pyruvate carboxylase. The requirements of manganese in feeds are generally greater for poultry (see Table 21.2), and feeds based on corn and barley could be deficient. Toxic effects of manganese are unlikely, as this element is considered one of the least toxic; however, continuous grazing on the volcanic soils of Costa Rica can result in a manganese intake of >200 mg/day and has been reported to produce reproductive changes in cattle. Adverse effects of excessive manganese intake are mainly caused by its interactions with iron and other elements (see Fig. 21.1), resulting in secondary iron deficiency anemia that can be seen in both lambs and pigs.

21.3.5 Cobalt

A deficiency of cobalt *per se* is not described in carnivorous animals, but their diet should have this element incorporated into vitamin B_{12} (cyanocobalamin) for normal methionine synthesis and energy metabolism. Two different forms of vitamin B_{12} are responsible for these two metabolic pathways: methylcobalamin for methionine synthesis and adenosylcobalamin for energy metabolism. Methylcobalamin is of importance for both protein and lipid synthesis, and in humans a vitamin B_{12} deficiency typically causes a specific megaloblastic anemia (pernicious anemia) and neurological disorders due to progressive demyelination. Pernicious anemia is not seen in domestic animals in cases of vitamin B_{12} deficiency. Ruminants and other herbivorous species do not need preformed vitamin B_{12} in their feed, as the microbes present in their gastrointestinal tract, especially the rumen, can produce this complex compound if sufficient cobalt is present in the feed. The requirement in all domestic ruminant species is about 0.1 mg/kg feed (d.m.), as the uptake of the vitamin is comparatively low in ruminants and even less in monogastric animals.

Young, growing ruminants grazing cobalt-deficient pastures are most sensitive to deficient levels of the adenosylcobalamin moiety. Lambs especially will show symptoms of inanition in spite of access to grazing, but older animals can also be affected under certain circumstances. Underlying this problem is the fact that adenosylcobalamin converts methylmalonyl-CoA to succinyl-CoA, beginning with propionic acid formed in the forestomachs of ruminants. In monogastric species, this metabolic pathway is of minor significance, but the energy metabolism of ruminants is based on the volatile fatty acids (VFAs) acetate, butyrate, and propionate and not glucose, which is typical for most other species. The clinical effects of cobalt deficiency in ruminants grazing deficient pastures have been recognized for quite some time in many parts of the world. The disease, or rather the apparent symptoms, goes by various local names (*e.g.*, in New Zealand, bush sickness; in Australia, wasting disease; in the United States, salt sick or neck ail; in Brazil, *mal de colete* or *pest de secar*; in Great Britain, pining disease; in Norway, white liver disease), apart from more general terms such as muscular wasting or enzootic marasmus. Sheep grazing in deficient areas are more sensitive than cattle, and young animals are more sensitive than adults. The deficiency is often due to low levels of cobalt in peaty soils or otherwise highly organic soils. Signs of cobalt deficiency in ruminants are retarded development in lambs, slow growth rate, inappetence, emaciation, and weakness.

Eventually, animals suffering a deficiency of cobalt die of emaciation. At an early stage, the symptoms are nonspecific and other reasons for the observed pathology could be suspected. Cobalt deficiency has to be confirmed by determination of methylmalonic acid (MMA) in blood samples or determination of cobalt/vitamin B_{12} in tissue samples, preferably the liver. At autopsy, the carcass is generally pale, and the liver is also characteristically pale with a pathologic texture that gives the deficiency one of its more common names—ovine white liver disease (OWLD) (see Figs. 21.2,

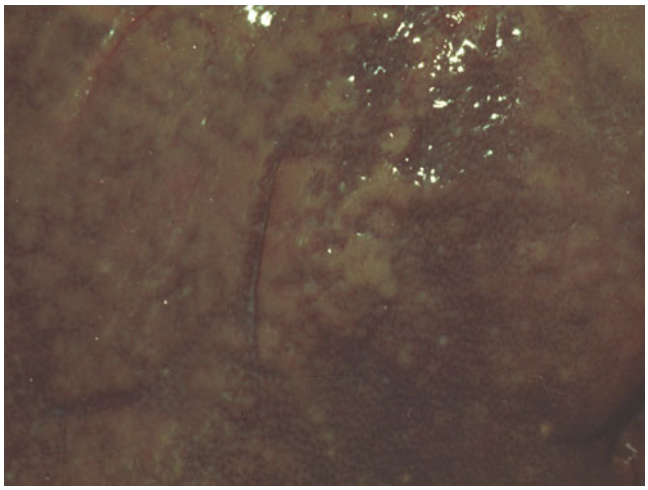


Fig. 21.2 The liver from an animal euthanized due to ovine white liver disease (OWLD); note pale irregular areas protruding above the surface (Photograph courtesy of Martha J. Ulvund, Sandnes, Norway)

Fig. 21.3 Lamb showing nonspecific signs such as serous eye discharge and small crusts on the ears, typical of OWLD. This lamb was only half the size of its healthy flockmates (Photograph courtesy of Martha J. Ulvund, Sandnes, Norway)

21.3 and 21.4). In some situations, OWLD or cobalamin deficiency can be provoked by some additional nutritional factor suspected to be fructan or other water-soluble carbohydrate compounds present in some grasses early in the season. These carbohydrates are fermented to VFAs in the forestomachs which increases the demand for adenosylcobalamin in the energy metabolism of the animal.

Neurological symptoms due to central demyelination that have been reported in vitamin B_{12}-deficient lambs include blindness and locomotor disturbances (Ulvund and Pestalozzi 1990). Recently, similar neurological symptoms observed for decades in Nova Scotia moose have been associated with cobalt deficiency (Frank et al. 2003). Vitamin B_{12} deficiency and neurological problems among elderly humans have attracted considerable interest lately.

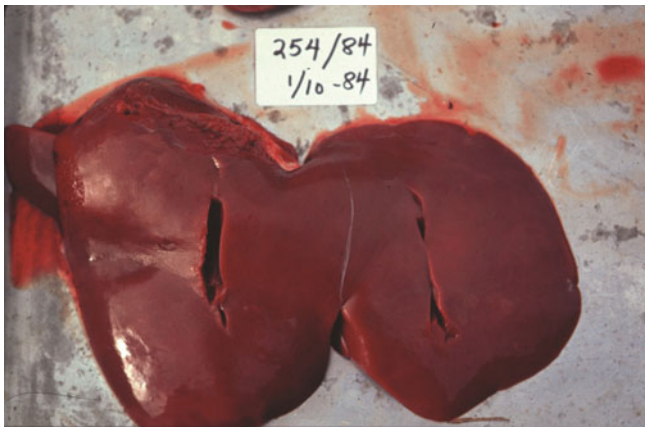


Fig. 21.4 Swollen and soft liver from the animal shown in Fig. 21.3; the color is paler than normal (Photograph courtesy of Martha J. Ulvund, Sandnes, Norway)

21.3.6 Copper

The effects of copper deficiency have been extensively studied in several parts of the world, because it is a widespread problem for sheep and cattle production. Copper deficiencies due to local geochemical anomalies produce significant losses for farmers worldwide. Studies of copper metabolism in various species have shown that there are important differences between ruminants and monogastric animals due to copper–molybdenum sulfate interactions in the forestomachs and biochemical handling of copper in the body after absorption. These biochemical differences are mainly due to the specific binding properties of metallothioneins in the liver and kidneys of ruminant species which differ from those seen in nonruminants. The copper–molybdenum sulfate interactions can make it difficult to evaluate the copper status of an animal because it is necessary to know the availability of all three compounds to make a reliable estimation of biologically active copper. To avoid an imbalance between copper and molybdenum, the ratio in the feeds should be between 1 and 2 or 1 and 4 (this important interaction is further discussed in Sect. 21.3.10). These differences in the binding properties of metallothionein render ruminants very sensitive to chronic copper toxicity even after slight but prolonged excessive intake of copper.

Copper deficiency has been described in cattle and sheep from many areas of the world, such as New Zealand, Australia, Northern Europe, and the United States. In Northern Europe, the problem is associated with grazing on peaty soils which often are also low in several other trace elements. The symptoms of copper deficiency are dominated by anemia and altered iron metabolism, bone disorders, diarrhea, and wasting, but also seen are infertility and depigmentation of wool and hair. In newborn lambs,

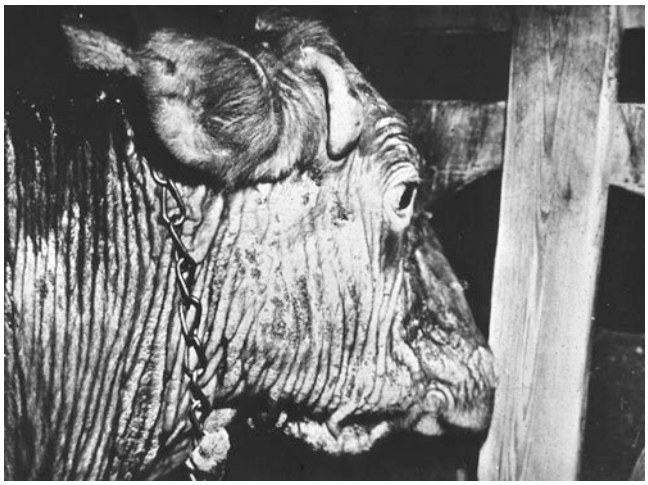


Fig. 21.5 Severe zinc deficiency in a dairy cow causing parakeratosis and pronounced wrinkling of the neck skin

dysfunctioning nerve development (swayback) can be seen when the ewe has been fed insufficient amounts of copper during the gestational period, when the central nervous system of the fetus develops. Swayback can also be seen in goats but rarely, as goats appear to be less sensitive to this type of deficiency. Cattle can sometimes show cardiovascular disorders (falling disease) in cases of copper deficiency, with sudden death after excitement or physical activity.

21.3.6.1 Chronic Intoxication

Chronic copper toxicity seen in grazing ruminants is very often due to very low molybdenum and sulfate intake. A suitable ratio between the content of copper and molybdenum in the feed is 1–2 to 1–4. The figures given in Table 21.3 for copper requirements are valid under the assumption that the ratio is in this interval. Sheep and calves are more sensitive to this intoxication than adult cattle and goats. The excessive copper is bound to metallothioneins and initially accumulated in the liver. Symptoms of chronic copper intoxication are seen only when the animal is stressed by pregnancy and delivery, transportation, infection, or other disease. The stress induces a massive release of copper from metallothioneins in the liver that initiates a hemolytic crisis, with erythrocyte disruption and secondary liver and kidney damage. Only at this stage is the concentration of copper in blood elevated. Often, affected animals do not show any symptoms but are suddenly found dead. If an animal survives the acute hemolytic crisis, jaundice and symptoms of liver and kidney failure will be seen. A typical postmortem picture includes a discolored carcass and kidneys. If chronic copper intoxication is suspected in a flock of sheep, supplementation with molybdenum and sulfate could be used to bind excessive copper and facilitate excretion. In areas with known high levels of copper in feeds, special boluses containing molybdenum are placed in the rumen of grazing sheep and cattle to prevent chronic copper toxicosis. Goats are less sensitive and will seldom show chronic copper toxicosis unless the copper intake is very high or the molybdenum and sulfate intake is very low.

21.3.7 Zinc

Zinc is vital to a very large number of important enzymes, and inadequate amounts of zinc in feeds will produce many different effects on animals. Apart from the interactions shown in Fig. 21.1, the intestinal uptake of zinc will decrease if feed levels of phytate are high or if clay substances are ingested. Clinical apparent zinc deficiency has been reported in various parts of the world (*e.g.*, New Zealand, Australia, northwestern Europe). Early and very obvious signs are skin lesions (parakeratosis and hair loss), seen both in pigs and cattle (see Fig. 21.5). Reduced appetite and retarded growth rate are other signs of deficiency, as are vomiting and scouring in pigs. In calves, a stiff gait and swelling of joints are seen as a result of disturbances in bone mineralization. This effect on bone metabolism is potentiated by a high calcium intake, which is often seen in young growing animals (see Fig. 21.1).

In the Holstein–Friesian breed of cattle, an autosomal recessive trait has been described that interferes with intestinal absorption of zinc. In this condition, which parallels *acrodermatitis enterohepatica* in humans, affected calves show signs of zinc deficiency even though the feed intake of zinc is adequate.

21.3.8 Arsenic

Chronic arsenic poisoning is found in humans and animals in several parts of the world (*e.g.*, Bangladesh, Argentina, Mexico, and the United States). In many cases, water from deep wells with high concentrations of arsenic is responsible for the toxic effects (see also Chap. 12, this volume); however, the use of arsenic compounds in herbicides, anthelmintics, insecticides, and rodenticides presents a greater risk for accidental or inadvertent exposure of animals to toxic amounts of arsenic. The symptoms described in cattle with chronic arsenic toxicity include diarrhea with weight loss, inflammation of the eyes and upper respiratory tract, gait incoordination, and changes in hair coat. More acute cases of arsenic intoxication are rapidly lethal.

21.3.9 Selenium

Diseases caused by a high selenium intake have been recognized for several hundred years in some selenium-rich areas of the world. Diseases caused by selenium deficiency are also well known, although selenium was one of

the last elements to be recognized as an essential trace element, in 1957 (see also Chap. 16, this volume).

21.3.9.1 Selenium Deficiency

Problems associated with a low intake of selenium were recognized much later than the toxic effects of an excess of this element; however, several areas in New Zealand, China, Finland, Sweden, and North America do have insufficient selenium levels in their native feeds. Such a deficiency seems to be related to the growth rate of animals, as young, fast-growing animals such as piglets, foals, lambs, and calves are more sensitive to a low intake of selenium. The selenium-containing enzyme glutathione peroxidase (GSH-Px) has a cytoplasmic localization and acts together with fat-soluble vitamin E, which has a membrane localization in the cell. In fresh feeds, the supply of vitamin E is high but during storage the content decreases; therefore, in situations where stored feeds are used, it is especially important to ensure that the selenium requirements are met. Normally, GSH-Px and vitamin E will detoxify the free radicals produced by the continuous oxidative metabolism present in all cells of an animal. If feeds high in polyunsaturated fatty acids or other oxidative compounds are utilized, the nutritional requirements for selenium will be higher, especially if the feed has been stored for a long time. In pigs, liver degeneration and necrosis are prominent signs, but muscular degeneration is also seen in both skeletal and cardiac muscles. Muscular pathology, often precipitated by even minor physical activity, is the dominating symptom in other species and causes paralysis and acute heart failure. In many cases, the young animals are found dead in the pasture or in their boxes. The typical postmortem picture in these cases features degeneration of muscular tissue, which looks like fish meat. Similar changes can be seen at autopsy in skeletal muscles from slaughtered paralytic animals.

21.3.9.2 Acute Selenium Toxicity and Blind Staggers

If large quantities of selenium are ingested, an acute form of intoxication, called *blind staggers*, can be seen for which blindness and other symptoms from the nervous system are prominent. Other symptoms include diarrhea and prostration. Affected animals often die within hours. Animals with acute intoxication also develop a garlic-like breath odor due to the formation of volatile methylated selenium compounds. This type of acute intoxication can be seen in areas throughout the world where seleniferous plants accumulate selenium from the selenium-rich soil (*e.g.*, Australia, China, North America, Russia, and Western Europe). Normally, grazing animals avoid these plants, but when a shortage of feed occurs the animals will graze on these plants and become intoxicated. If smaller amounts of selenium are ingested, the neurological symptoms are more prominent and, in addition to blindness, paralysis, respiratory problems, abdominal pain, and other signs of pain such as grinding of the teeth and salivating will be seen. Eventually the animals die due to starvation or respiratory failure. Both the blindness and the locomotor problems contribute to the inability to graze and meet the metabolic needs of the sick animal.

21.3.9.3 Chronic Selenium Toxicity and Alkali Disease

Around 1295, in his travels in western China, Marco Polo described what is most probably chronic selenium intoxication in horses. Animals eating certain plants that accumulated selenium showed dramatic changes, such as loss of hoofs (exungulation) and dermatological changes. Similar toxic effects were later described in North America when horses in the U.S. Cavalry were affected. It is even claimed by some that the outcome of the battle at Little Bighorn was determined by the ingestion of seleniferous plants by the horses in General Custer's squadrons which made them sick and unfit for the ensuing battle. Cattle brought by settlers who came to the Great Plains also showed signs of selenosis. These farmers associated the disease with the high salt content of the water and alkali seeps in the area and named the problem alkali disease. It appears, however, that certain seleniferous plants are necessary for the appearance of alkali disease. Such plants can accumulate selenium several hundredfold in their tissues compared to the soil concentration. Seleniferous plants are also found in limited areas of India, Israel, Central Europe, and South America, where chronic selenosis is observed. In other areas of the world, such as Hawaii and Costa Rica, the volcanic soil is rich in selenium but no seleniferous plants grow, and alkali disease is not seen.

21.3.10 Molybdenum

21.3.10.1 Deficiency

The biochemical roles of molybdenum-containing enzymes vary according to the different valence states (+4, +5, or +6) attainable for this element. Deficiencies caused by low molybdenum intake are seen in experimental situations but are rare under natural conditions. The nutritional requirement in ruminants is estimated to be <0.1 mg/kg feed (Mertz 1987). The intricate interrelationships among molybdenum, copper, and sulfate make it difficult, however, to give reliable figures for molybdenum requirements and adverse effects in various animal species (see Sect. 21.3.10.2). It has been suggested, though, that a molybdenum deficiency is the cause of the high incidence of urinary xanthine calculi in New Zealand sheep grazing on pasture low in molybdenum due to the resulting low activity of the molybdenum-containing enzyme xanthine

oxidase. Other situations where molybdenum deficiencies have been suspected to occur spontaneously involve poultry, but these problems have not responded to simple molybdenum supplementation because the etiology is more complex. Most probably the earlier mentioned interactions among molybdenum, copper, and sulfate are involved, but other transition elements may also be involved as contributing factors in these cases.

21.3.10.2 Intoxication

The effects of excessive molybdenum intake (known collectively as molybdenosis) are closely linked to copper deficiency, and most of the symptoms seen in molybdenosis are linked to the decreased activity of copper-containing enzymes. This is especially true in ruminants, where the balance between these two elements and sulfur in the form of sulfate or sulfur-containing amino acids is very important but often difficult to estimate from analytical data on feed composition. The chemical reason for this interaction is the formation of oxythiomolybdate and tetrathiomolybdate (TTM) in the rumen. The TTM will react with copper and other elements in the gastrointestinal tract, making these elements unavailable for absorption. TTM will also be absorbed from the gut to the blood and distributed to various tissues in the body where it binds copper in enzymes and other proteins, thus rendering them dysfunctional. Cattle are most susceptible to molybdenosis, especially dairy cows and young animals. Sheep and goats are less sensitive, and horses as well as swine appear insensitive to high intakes of molybdenum. In the young, growing foal, however, the occurrence of rickets has been reported when the mare has been kept on high-molybdenum pasture. Defective bone metabolism with osteoporosis and joint abnormalities have also been reported as a prominent symptom in molybdenotic cattle and sheep.

Severe molybdenosis in cattle is widespread in the world. Plant uptake of molybdenum is favored by soil with a neutral or high pH. This problem has been known for a long time in England, where it is called *teart*, and in New Zealand, where the problem is known as *peat scours*. When cattle are put on pasture high in molybdenum, after only a few days they may exhibit typical signs of molybdenosis, such as profuse scouring and a harsh and discolored coat (see Fig. 21.6). In milder cases, the diarrhea is less severe and the animals will recover rapidly if put on pasture with low molybdenum content. Less obvious effects are seen with regard to reproduction and thyroid function in cattle, for which a direct effect of molybdenum on the endocrine regulation in the hypothalamus is suspected. Anemia is also seen in cattle and sheep due to decreased hemoglobin synthesis, because the copper-containing enzyme ceruloplasmin is needed for the incorporation of iron into the heme molecule.

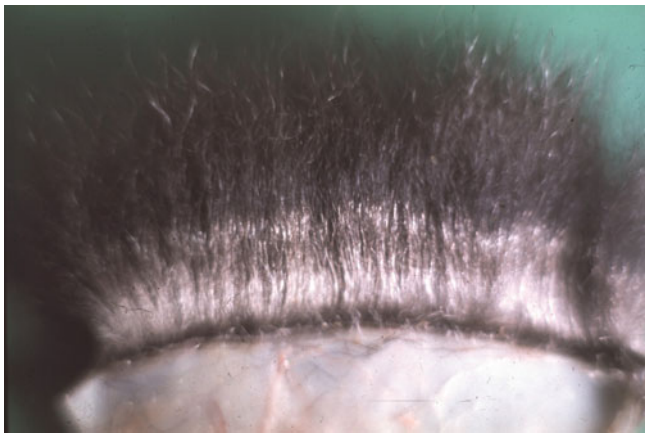


Fig. 21.6 Discoloration of the black wool in a sheep suffering from molybdenosis

21.3.10.3 Molybdenosis in Moose

An example of the complexity of diagnosing trace element imbalances in animals, especially wild animals, is the occurrence of molybdenosis in moose (*Alces alces* L.) in an area in southwest Sweden (Älvsborg county) beginning around 1985 (Frank 1998; Frank et al. 2000). Affected animals that were found alive displayed a number of symptoms, including behavioral disturbances with apathy and pathological locomotion, loss of appetite, diarrhea, emaciation, discoloration and loss of hair, and opacities in the cornea and lens of one or both eyes. Postmortem examinations revealed edema, hyperemia, hemorrhage, and erosive lesions in the mucosa of the gastrointestinal tract. The heart was dilated and flabby, and myocardial lesions were seen macroscopically. Microscopic studies of different tissues showed pronounced hemosiderosis in the liver and spleen. Often, an animal showed some but not all of these symptoms or postmortem signs, making it difficult to classify a specific individual as a victim of this particular disease.

The gastrointestinal changes indicated a viral etiology of the disease, but other findings did not support this, and still no virus had been detected that could be responsible for the disease. It was suggested that starvation due to overpopulation or high age in the local moose population could account for the disease. Some of the chemical findings in samples from sick animals, however, were not compatible with starvation and lack of nutrients, energy, minerals, or trace elements. An indication of the probable etiology was found in measurements of trace elements in livers and kidneys from yearling moose sampled all over Sweden during the normal hunting periods in 1982, 1988, 1992, and 1994. These results showed that from 1982 (before the outbreak of the disease) to 1994 the hepatic concentration of selenium and especially molybdenum increased in the affected area, whereas the levels of copper, cadmium, and some other

elements decreased in the livers of these young animals. These results were different from those for neighboring counties during the same period.

A plausible explanation for this discrepancy is the fact that the Älvsborg county was heavily limed to counteract the effects of acid rain, which was damaging the local forest and aquatic ecosystems. An increase in soil pH after liming increases the availability of molybdenum and selenium but decreases the availability of other metallic trace elements. The majority of symptoms and pathological signs found in the moose can be explained, then, by decreased activity of copper-containing enzymes such as tyrosinase, ceruloplasmin, superoxide dismutase, lysyloxidase, diaminoxidase, and methioninsyntase. Specific analysis of the copper enzyme cytochrome *c* oxidase in myocardial tissue showed lower activity in sick animals than in healthy controls. Determinations of the pancreatic hormone insulin in blood plasma from diseased moose showed concentrations about double those of apparently healthy animals, and still there was evidence of persistent hyperglycemia in these animals, as higher concentrations of furosine and pentosidine were found in kidney tissue compared to the levels determined in healthy animals. Furosine, pentosidine, and some other compounds are formed by nonenzymatic glucation of tissue proteins in the presence of persistent high levels of blood glucose. The hyperinsulinemia and persistent hyperglycemia observed in the moose are similar to non-insulin-dependent diabetes (NIDD, or type 2 diabetes) often seen in humans, especially the elderly.

Some other endocrine disruptions could be suspected, because the direct effects of TTM on the hypothalamus have been demonstrated in experimental animals. Decreased reproductive capacity was suspected but difficult to detect in moose in the field; however, lowered thyroxin levels were measured in blood samples from the wild, free-living moose, which supported the idea of a central endocrine disruption. Similar pathological effects affecting central endocrine functions have recently been reported from clinical and experimental studies in sheep with chronic molybdenosis (Haywood et al. 2004). Similar molybdenum-induced copper deficiency problems were suspected earlier and reported in moose (*Alces alces gigas*) in Alaska (Kubota 1974; Flynn et al. 1977) and in Grant's gazelle (*Gazelle granti*) in the Rift Valley in Kenya (Hedger et al. 1964). The diagnostic biochemical investigations were, however, not complete in these earlier cases.

21.3.11 Iodine

Goiter, enlargement of the thyroid gland due to iodine deficiency, was described in ancient Egypt in humans and animals. It is still probably the most widespread trace element deficiency in world animal husbandry. Typical deficient areas are inland and mountain areas, such as the inner parts of Siberia, Africa, and South America; the slopes of the Himalayas, Alpine valleys, the Pyrenees; and the Andes, where wind and rain carry only limited amounts of iodine from the oceans. Some lowland and coastal areas also have soil in which leaching has caused iodine depletion (see also Chap. 17, this volume). The most sensitive and typical signs of iodine deficiency are enlargement of the thyroid gland and changes in the microscopic morphology (see Fig. 21.7). Lambs and calves are often born with severe goiter when their mothers have been fed insufficient iodine during gestation. In severe cases, these newborn animals can suffocate shortly after birth due to the pressure of the enlarged thyroid gland on the trachea. If they survive the first critical period, they are often so weak that they have severe problems getting up and standing to suckle their mothers to get the antibodies from colostrum and nutrients vital to the start of life.

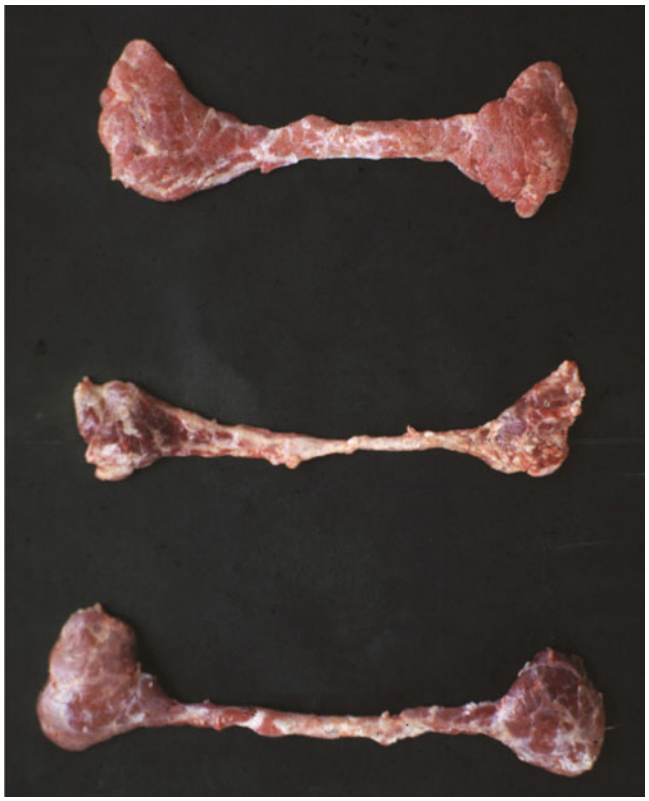


Fig. 21.7 Thyroid glands from three steers fed different amounts of iodine. The gland in the middle is normal sized and comes from an animal with adequate iodine supply. The other two show different degrees of goiter. The *upper gland* shows a moderate increase in size and the *lower gland* a minor increase due to insufficient iodine intake

In adult animals, iodine deficiency causes a generally depressed metabolism due to insufficient production of the thyroid hormone thyroxin. Lethargy, increased fat deposits,

and impaired reproduction are other observed effects. In breeding animals, irregular and suppressed estrous cycles are early signs of iodine deficiency, and the reduced fertility will have obvious effects on the economy of the farmer. If animals get pregnant, the fetus can suffer developmental problems that result in fetal death, abortion, or stillbirth.

Endemic iodine deficiency can be further aggravated by at least two factors operating in modern agriculture. First, plants containing goitrogenic substances are common in both cultivated land and growing on natural pasture. These goitrogenic substances are chemically glucosinolates (*e.g.*, isothiocyanates, thiocyanates, nitriles, and oxazolidinethiones), which block the uptake of iodine in the thyroid. Glucosinolates are present in the green parts and seeds of commonly cultivated crops of Cruciferae plants, such as various types of kale and rape. In temperate zones of the world, cultivation of these glucosinolate-containing plants is common because they contribute valuable proteins to locally produced animal feeds. A second important factor contributing to iodine deficiency is the increasing interest in organic farming, especially in North America and Western Europe, where the limited use of supplements decreases the iodine supply in deficient areas.

Both of these problems can be easily alleviated by using appropriate supplements rich in iodine. The most common remedy is the use of iodized salt (NaCl) as a salt lick or a component of mineral supplements given to the animals. An important detail regarding iodine supplementation via salt licks or otherwise is that KI, often used as an inexpensive remedy, evaporates even at ambient temperature during storage; therefore, old salt licks may not contain the necessary amount of iodine supplementation. It is particularly important to use iodine supplementation for dairy production, as milk and other dairy products are a major iodine source for humans in large parts of the world, especially in areas where marine products such as kelp, fish, or shellfish are not consumed regularly.

Further Reading

Bowell RJ, Warren A, Redmond I (1996) Formation of cave salts and utilization by elephants in the Mount Elgon Region, Kenya. In: Appleton JD, Fuge R, McCal GJH (eds) Environmental geochemistry and health, Special Publication no 113, Geological Society, London, pp 63–79

Chaney SG (2002) Principles of nutrition II: micronutrients. In: Devlin TM (ed) Textbook of biochemistry with clinical applications, 5th edn. Wiley-Liss, Philadelphia

Flynn A, Franzman AW, Arenson PD, Oldemeyer JL (1977) Indications of copper deficiency in a sub-population of Alaskan moose. J Nutr 107:1182–1189

Frank A (1998) Mysterious" moose disease in Sweden: similarities to copper deficiency and/or molybdenosis in cattle and sheep. Biochemical background of clinical signs and organ lesions. Sci Total Environ 209:17–26

Frank A, Danielsson R, Jones B (2000) The mysterious disease in Swedish moose. Concentrations of trace and minor elements in liver, kidneys and ribs, Haematology and clinical chemistry. Comparison with experimental molybdenosis and copper deficiency in the goat. Sci Total Environ 249:107–122

Frank A, MacPartlin J, Danielsson R (2003) Nova Scotia moose mystery—a moose sickness related to cobalt and vitamin B_{12} deficiency. Sci Total Environ 318(1–3):89–100

Haywood S, Dincer Z, Jasani B, Loughran MJ (2004) Molybdenum-associated pituitary endocrinopathy in sheep treated with ammonium tetrathio-molybdate. J Comp Pathol 130:21–31

Hedger RS, Howard DA, Burdin ML (1964) The occurrence in sheep and goats of a disease closely similar to swayback. Vet Rec 76:493–497

Jacobson DR, Hemken RW, Button FS, Hatton RH (1972) Mineral nutrition, calcium, phosphorus, magnesium, and potassium interrelationships. J Dairy Sci 55:935–944

Kubota J (1974) Mineral composition of browse plants for moose. Nat Can 101:291–305

Markert B, Kayser G, Korhammer S, Oehlmann J (2000) Distribution and effects of trace substances in soil, plants and animals. In: Markert B, Friese K (eds) Trace elements in the environment, vol 4. Elsevier, Amsterdam, pp 3–33

Maskall J, Thornton I (1996) The distribution of trace and major elements in Kenyan soil profiles and implications for wildlife nutrition. In: Appleton JD, Fuge R, McCal GJH (eds) Environmental Geochemistry and Health, Special Publication no 113, Geological Society, London, pp 47–62

McDowell LR (1992) Minerals in animal and human nutrition. Academic, San Diego

Mertz W (1987) Trace elements in human and animal nutrition, vol 1 and 2, 5th edn. Academic, San Diego

NRC (1979) Metabolic and biologic effects of environmental pollutants, zinc. University Park Press, Baltimore

NRC (1985) Nutrient requirements of domestic animals: nutrient requirements of dogs, 2nd edn. National Academy of Sciences/National Research Council, Washington, DC

NRC (1994) Nutrient requirements of domestic animals: nutrient requirements of poultry, 9th edn. National Academy of Sciences/National Research Council, Washington, DC

NRC (1996) Nutrient requirements of domestic animals: nutrient requirements of beef cattle, 7th edn. National Academy of Sciences/National Research Council, Washington, DC. ISBN 0-309-06934-3

NRC (1998) Nutrient requirements of domestic animals: nutrient requirements of swine, 10th edn. National Academy of Sciences/National Research Council, Washington, DC

NRC (2001) Nutrient requirements of domestic animals: nutrient requirements of dairy cattle, 7th edn. National Academy of Sciences/National Research Council, Washington, DC. ISBN 9780309515214

NRC (2007a) Nutrient requirements of domestic animals: nutrient requirements of horses, 6th edn. National Academy of Sciences/National Research Council, Washington, DC. ISBN 978-0-309-10212-4

NRC (2007b) Nutrient requirements of small ruminants. Sheep, goats, cervids, and new world camelids. National Research Council of the National Academies, Washington, DC. ISBN 0-309-10213-8

Sjaastad ÖV, Hove K, Sand O (2003) Physiology of domestic animals. Scandinavian Academic Press, Oslo/Norway

Ulvund M, Pestalozzi M (1990) Ovine white-liver disease (OLWD) in Norway: clinical symptoms and preventive measures. Acta Vet Scand 3:53–62

The Impact of Micronutrient Deficiencies in Agricultural Soils and Crops on the Nutritional Health of Humans

22

Marija Knez and Robin D. Graham

Contents

22.1 Introduction – Micronutrient Deficiencies

You can trace every sickness, every disease and every ailment to a mineral deficiency

(Dr. Linus Pauling, Nobel Laureate 1954)

Cereal crops underpin the food supply for peasant farmers in developing countries, a situation that has persisted since the green revolution of the decades of the 1960s and 1970s greatly increased their productivity; indeed yields more than doubled. Ever since, the modern high-yielding, disease-resistant cereals that also show greater tolerance to environmental stresses like drought and heat have dominated the diets of subsistence farmers and urban poor alike. It was known that cereals are less dense in some nutrients important to health and vigour than crops they in part replaced, like pulses, but the consequences of that were not immediately apparent. However, the incidence of iron deficiency in humans grew through the 1980s and by 1990 was becoming a matter of real concern to nutritionists (Graham et al. 2007). Throughout this period the international consortium of agricultural scientists was naturally concentrated on increasing the yield, with little attention being devoted to the issue of nutritional quality (micronutrient density) of produced crops. Exhausted from the years of cultivation, soils slowly become more deficient in most essential nutrients, including micronutrients. Consequently, more and more people in developing country populations, who were entirely dependent on these soils for their food production, began to have inadequate mineral and vitamin intakes, leading to malnutrition and poor health. Later still, we became aware that a large number of people was affected by four other micronutrient deficiencies, those of iodine, vitamin A, zinc and selenium (WHO 1995).

Micronutrient deficiencies still remain a major public health problem in many countries worldwide, with more than two billion people suffering from one or more micronutrient deficiencies (WHO 2012).

M. Knez (✉) • R.D. Graham
School of Biological Sciences, Flinders University of South Australia, Adelaide 5001, Australia
e-mail: knez0004@flinders.edu.au

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_22,

As anticipated, deficiencies of micronutrients are highly prevalent in places where cereals with low nutritional quality are the main diet (Cakmak 2009; Subramanian and Jagan 2008).

The World Health Organization (WHO 2011) has estimated that nearly 3.7 billion people are Fe deficient, with two billion of these so severely deficient in Fe that they can be described as being anemic. In addition, it is estimated that two billion people globally are at high risk of I deficiency (WHO 2009); 35 % of all children (1–5 years old) suffer from Zn or Fe deficiencies, 190 million preschool-aged children and 19 million pregnant women are at risk of subclinical vitamin A deficiency (WHO 2009). In many areas of the world, poor dietary quality and micronutrient deficiencies are more widespread problems than low energy intake (Stewart et al. 2010), with devastating consequences: 19 % of all deaths before the age of 5 years can be attributed to vitamin A, Zn, Fe and/or I deficiency (Prentice et al. 2008).

Micronutrient deficiencies are currently identified as the main contributors to the global burden of disease. The human and economic costs are enormous, falling hardest on the very poor, especially women and children (Bryce et al. 2008). The interaction of poverty, poor health and poor nutrition has a multiplier effect on the general wellbeing of the population and also adds significantly towards keeping a population in a downward trend of poverty and nutritional insecurity (Stewart et al. 2010). Therefore, the problem of resolving the micronutrient deficiencies becomes a major challenge for the entire scientific community.

This chapter addresses the impact of micronutrient deficiencies in agricultural soils and crops on the nutritional health of humans. Initially, the most common mineral deficiencies of soils and crops are discussed. The second part gives an explanation of the biological roles of most important micronutrients in humans, with particular emphasis on the impact of micronutrient deficiencies on human health. Finally, the last part of this chapter discusses the agricultural interventions that could be deployed to decrease the number of people suffering from micronutrient deficiencies, both in developing and developed countries.

22.2 Micronutrient Deficiencies in Soils and Crops

The micronutrients of most practical interest here are iron, zinc, copper, manganese, boron, molybdenum, nickel, cobalt and selenium, there being a second group (chlorine, sodium, aluminium, silicon, vanadium) that may be essential or beneficial in some particular situations. We will concentrate on the first group as those in the second group are only rarely deficient in agriculturally important soils to an extent that is of economic significance.

Unlike humans and animals, plants can synthesize all the organic molecules, such as amino acids and vitamins that they require; that is they need only minerals, air and water. We include boron in our priority group because it is so widely deficient in agricultural soils and so highly effective as a remedial fertilizer. It is important to production in its own right; its requirements for humans become secondary because its essentiality is still under investigation and deemed by some to be not fully proven. The comprehensive work of Sillanpaa (1990), that not only used soil and plant analysis but field experiments with several crops, has given us an appreciation of the extent of these micronutrient deficiencies in agriculturally important soils, especially in developing countries: while he found Zn was deficient in half of the major agricultural soils and boron deficient in a third, the deficiencies of copper, molybdenum and manganese were all of small and similar extent (10–15 %), and all three together adversely affected a total of ~40 % of all soils in the study (Sillanpaa 1990). Most curiously, although Fe deficiency is widespread in humans (on one occasion estimated by WHO in 2005 to affect five billion people, at least to some degree) it is deficient in only about 3 % of all productive soils of the survey. As a broad generalization, a deficiency of at least one micronutrient is probable for most soils and multiple deficiencies are not uncommon because of the similar soil-pH effects on the availability of most of the cations.

Whereas we can generalize that micronutrient concentrations in plants reflect those in the soils they are growing in, this does not take us very far. In reality, plant species differ dramatically in their inherent abilities to extract each nutrient from a given soil. These and varietal differences operate within each target species such that plant breeding within a species can greatly increase the adaptation to a deficient soil (Graham 1984). Whether plant breeding is focused on a specific trait or simply on pursuing potential yield empirically, a nutrient efficiency trait is a common basis for yield improvement, especially providing a capacity to extract nutrient from deeper soil horizons that cannot be economically treated with fertilizers.

Formal studies of micronutrient efficiency traits have shown that copper efficiency appears in cereals to be due to a single major gene (Graham et al. 1987); in Mn deficient soils growing durum wheat breeders' lines, we found the trait involved one gene with major effect operative per genome (Khabaz-Saberi et al. 2002), whereas in Zn-deficient cereal growing soils, the trait was polygenic, yet it can be usefully manipulated in breeding programs using molecular markers that have been developed (Blair et al. 2009; Genc et al. 2009; Lonergan et al. 2009).

Further, these micronutrient efficiency traits are composed of both uptake efficiency traits and seed loading traits (Wissuwa et al. 2008; Genc et al. 2009). Grains with high micronutrient concentrations produce more vigorous seedlings when planted in deficient soils and this early advantage almost invariably manifests itself in significantly higher grain yield. In an example from Bangladesh, micronutrient-enhanced seeds from fertilizer-treated mother plants resulted in 24 % higher grain yields averaged over 4 years (Duxbury et al. 2009).

The chemistry of the micronutrient cations Fe, Cu, Mn, Ni and Co in soils is largely dominated by pH and texture: the lower the pH, the more soluble and available they are to plant roots, but below pH 4.5–5, the effect of H^+ ions on plant membrane integrity counters the advantage of higher solubility of micronutrient cations in soil, so deficiencies in the plant are more likely under extreme soil acidity. The higher the content of clay, the more strongly micronutrient cations are held by soil colloids, countering leaching, but on the other hand, decreasing by adsorption the movement laterally towards the roots for absorption into the plant. All these and many other soil factors make assessing the availability of micronutrients in soil for plant growth quite difficult to the extent that in the writer's opinion, nutrient availability, and fertilizer requirements, are best assessed by analysing the growing plant, especially actively growing leaves, and comparing the results to plant analysis standards that have been set by exhaustive experimentation (see Reuter and Robinson 1997). This approach usually identifies the most limiting nutrient and when treated, other limitations may be explored. At pH 5 and lower, root activity is itself adversely affected so the optimal pH for availability of these nutrients to plants is commonly in the intermediate range around 5.5–7. Acidity in soil can be decreased with application of lime or dolomite. On the other hand if salinity is not severe, many cereal crops can obtain sufficient of these micronutrients from soils in the alkaline range, a common condition in cereal cropping soils of lower rainfall regions.

22.2.1 Remediation of Soils by Fertilizer Use

Micronutrients can be used as fertilizers to increase soil concentrations to a level that results in adequate uptake by plant roots, so removing this constraint to productivity. Micronutrients can be applied separately or combined with organic manures or one or more macronutrients such as nitrogen, phosphorus, potassium, calcium, magnesium and sulphur. Fertilizers may be applied to the soil or in aqueous solution, much diluted and often repeated, to the leaves of the plants themselves.

The use of fertilizers containing some combination of macronutrients and micronutrients is standard practice in highly productive agriculture but most competent studies in subsistence farming systems show they can be equally highly effective. The limitation for poor farmers is the cost. This can be addressed to some extent by first diagnosing the limiting nutrients in order to decrease the overall cost of the optimal fertilizer mix. This is a complicated task for the experienced professional agronomist. Diagnosis of nutrient needs is therefore more often than not prohibitive in cost or availability of expertise to subsistence agriculture. Consequently, the best compromise may be a complete nutrient mix containing macronutrients and micronutrients in ratios derived from an extensive literature of agronomic experience in comparable environments, crops and soils. Small areas should be treated with small amounts of a complete mixture (at small cost) and the results assessed for the economic advantage of scaling up the area treated for the following crop. If promising, professional support should be found to help in fine-tuning the mixture for maximal returns on the investment.

22.3 Biological Functions of Micronutrients

UNICEF (1998) defined micronutrients as nutrients (dominantly vitamins and minerals) that are only needed by the human body in minute amounts. Micronutrients are required by humans for a vast range of physiological functions; they play leading roles in the production of enzymes, hormones and other substances, helping to regulate growth, development and the functioning of the immune and reproductive systems. While it is argued that at least 51 different nutrients are needed in adequate amounts by human beings (Welch 2008), only a few of these elements are needed in such small amounts that they are known as micronutrients (chromium, manganese, iron, cobalt, copper, zinc, selenium, molybdenum, and fluorine). For some other elements, clinical symptoms have been observed with low concentrations in the diet, but the possible essential function is not known. On the other hand, micronutrients defined as nutrients with public health importance include the following: zinc, iodine, iron, selenium, copper, vitamins A, E, C, D, B2, B6, B12 and folate. Humans suffer primarily from deficiencies of iron, zinc, iodine, selenium and Vitamin A (WHO 2009).

The text below provides a concise review of the main biological functions of some of the essential micronutrients in humans. More information on the issue can be found in Chaps. 6 and 7.

22.3.1 Zinc (Zn)

Zinc (Zn) is one of the most important trace elements in the human body, with an average amount in the adult body

between 1.4 and 2.3 g Zn (Calesnick and Dinan 1988). Zinc is found in all body tissues and fluids in relatively high concentrations, with 85 % of the whole body Zn in muscle and bone, 11 % in the skin and the liver and the remaining in all the other tissues (Kawashima et al. 2001).

Zinc has been identified to bind with 925 proteins in humans, and as such, it can be deemed the most important metabolic promoter among the known essential nutrients. It is involved in almost all processes and metabolic pathways, and is necessary for a wide range of biochemical, immunological and clinical functions (Hotz and Brown 2004). Zinc is needed for nucleic acid metabolism and protein synthesis, cellular differentiation and replication as well as glucose metabolism and insulin secretion (Vallee and Falchuk 1993). Additionally, Zn promotes the synthesis of metallothionenin, a protein found in intestinal mucosa, liver, and kidneys that is necessary for normal heavy metal metabolism (Hotz and Brown 2004). Metallothioneins play a key role in Zn-related cell homeostasis due to their high affinity for Zn, which is in turn relevant against oxidative stress and immune responses, including natural killer (NK) cell activity and ageing.

Zinc is essential as a catalytic, structural and regulatory ion. Several genes have been recognised as potential candidates for regulation by Zn: genes involved in the regulation of redox state, fatty acid metabolism, signal transduction and platelet activation.

Zinc serves as a structural component that allows for the coordinate binding of amino acids, mainly cysteine and histidine residues in the protein chain, to form a finger like structure.

Zinc is an important element in preventing free radical formation, in protecting biological structures from damage and in regulating the immune functions. More specifically, Zn is an essential element for thymic functions by means of a Zn-dependent thymic hormone called thymulin required for T-cell maturation and differentiation (Mocchegiani et al. 2000).

22.3.2 Iron (Fe)

The human body contains approximately 3–5 g of Fe (45–55 mg/kg of body weight in adult women and men, respectively). The majority of body Fe (~60–70 %) is utilized within hemoglobin in circulating red blood cells (Andrews 1999). Approximately 20–30 % of body Fe is stored in hepatocytes and in reticuloendothelial macrophages, to a large extent within ferritin and its degradation product hemosiderin. The remaining body Fe is primarily localized in myoglobin, cytochromes, and Fe containing enzymes. A healthy individual absorbs daily 1–2 mg of Fe from the diet.

Iron is a key element in the creation of hemoglobin and myoglobin in humans; necessary for various biochemical pathways and enzyme systems including energy metabolism, cell division, neurotransmitter production, collagen formation and immune system function (Edison et al. 2008). Iron plays an essential role in oxygen transport, oxidative metabolism, and cellular growth. On the other hand, too much of it can be toxic to cells, as it can catalyse the creation of reactive oxygen species through the Fenton reaction. Therefore, it is essential that Fe uptake and storage at both cellular and whole body levels are precisely regulated. For the maintenance of body Fe homeostasis, there must be effective communication between the main sites of Fe utilization (e.g., the erythroid marrow), storage (e.g., the liver and reticuloendothelial system) and the primary site of absorption in the small intestine (Steele et al. 2005). Due to the different Fe requirements within the body, a number of 'regulators' of Fe homeostasis has been hypothesized: dietary regulator (or mucosal block regulator), stores regulator, erythropoietic regulator, inflammatory regulator, etc. (Edison et al. 2008). However, the latest data are showing that the various regulators may not be necessarily different and could represent differential responses mediated by the same molecules (Hentze et al. 2004). A molecule that could have a central regulatory role in Fe metabolism, secreted by the liver and excreted by the kidneys is hepcidin. It is a small peptide, which acts as a functional target for all other regulators (Edison et al. 2008).

In order to better understand how systemic Fe homeostasis is maintained, it is necessary to look at the movement of Fe among various tissues and organs of the human body (Fig. 22.1).

The bloodstream carries Fe bound to transferrin around the body; most of it is integrated into hemoglobin by developing erythrocytes in the bone marrow. Old or damaged erythrocytes are removed from the bloodstream by the macrophages, and the Fe is recycled back to plasma transferrin. Iron is needed in all tissues for their metabolic needs, and as it is not actively excreted, the amount of Fe in the body must be controlled at the point of absorption in the small intestine.

Dietary Fe is absorbed by the body through the small intestine in quantities equal to the amounts of lost Fe from the body, so establishing the body's Fe homeostasis. Iron flux from intestinal enterocyte to the bloodstream is modulated by a liver-derived peptide, hepcidin. Hepcidin expression is influenced by systemic stimuli such as the rate of erythropoiesis, Fe stores, hypoxia, oxidative stress and inflammation.

Hepcidin is the major regulatory point of Fe homeostasis and its expression is determined by the complex interplay of various factors, and depending on the specific situation, one of several stimuli will predominate (Prasad 1991).

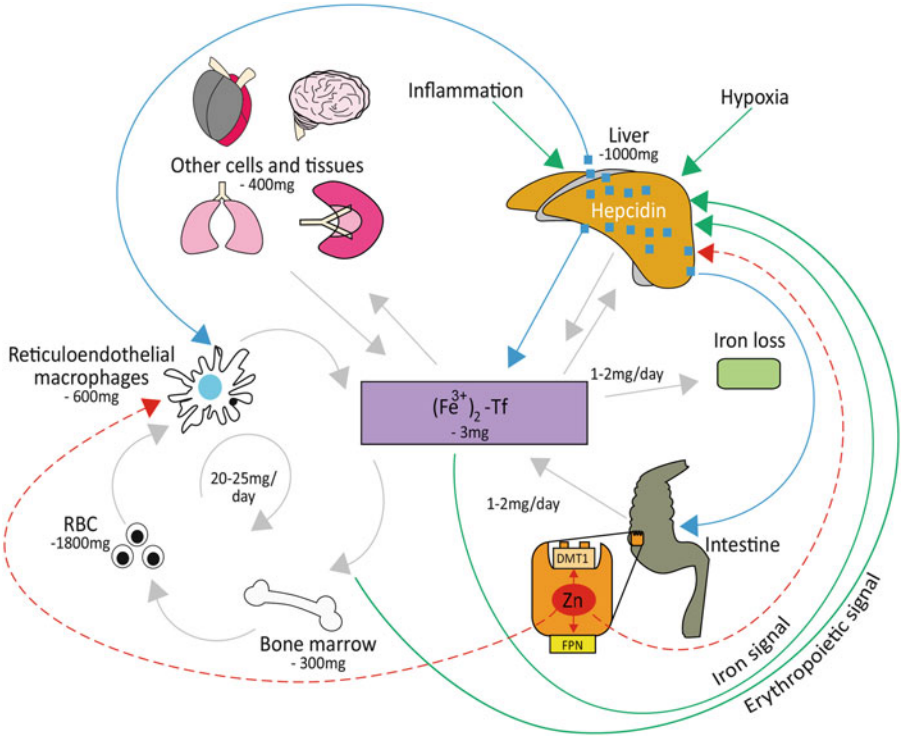


Fig. 22.1 Main routes of iron transport between various tissues and organs including the likely role of zinc in iron homeostasis. *Tf* transferrin, *RBC* red blood cells, *DMT1* divalent metal transporter 1, *FPN* – ferroportin

Stimuli can signal through multiple pathways to regulate hepcidin expression, and the interaction between positive and negative stimuli is critical in determining the net hepcidin level (Darshan and Anderson 2009). Since hepcidin expression is mostly restricted to the liver, it is highly likely that the hepatocyte is the site of action of the regulatory stimulus. Current data would suggest that Fe levels as such do not play a primary role in this process, but rather that an additional signal is involved. Recent research, summarized in Graham et al. (2012) provides evidence that Zn concentrations in the body may have a crucial role in Fe homeostasis (more about Fe uptake in the human body, Fe transporters and Fe homeostasis mechanism can be found in Chaps. 6 and 7).

22.3.3 Selenium (Se)

Selenium was first discovered in 1817; however it was not until 1957 that Schwarz and Foltz confirmed Se as an essential nutrient necessary for normal growth and reproduction in animals and humans. There is less than 1 mg of Se in the human body, most of it in the liver, kidneys and pancreas (Haas 2001). At least 25 different selenoproteins and a variety of subsequent isoforms exist in the human body (Gromer et al. 2005) and they exert multiple actions on endocrine, immune, and inflammatory functions. Of particular importance to reproduction and pregnancy are the six antioxidant glutathione peroxidases (GPxs), which play a pivotal role in reducing hydrogen peroxide (H_2O_2) and lipid peroxides to harmless products thus reducing the propagation of damaging reactive oxygen species (ROS) (Gromer et al. 2005). As antioxidants, the GPxs help maintain membrane integrity, protect prostacylin production, and limit the propagation of oxidative damage to lipids, lipoproteins, and deoxyribonucleic acid (DNA). Selenium is essential for testosterone biosynthesis and the formation and normal development of spermatozoa. Evidence is accumulating that elevated dietary levels of Se may be protective against colorectal and prostate cancer (Clark et al. 1996) and against heart disease (Liu 2003). It also appears that Se may reverse the accelerated decline in immune response in the elderly (Rayman 2000). Besides antioxidant and anti-cancer action, Se also has a positive role in detoxification of heavy metals (Rayman 2000).

22.3.4 Copper (Cu)

Copper is a relatively scarce metal in humans with an average concentration of 1.4–2.1 mg/kg in healthy individuals (Milne 1998). It is normally bound to proteins or to organic compounds and is not found as free Cu ions. Nearly two thirds of the body's Cu is found in the skeleton and muscles (Turnlund et al. 1998).

Most enzymes employ Cu in electron transfer reactions; enzymes involved in energy production (cytochrome oxidase) and in the protection of cells from free radical

damage (superoxide dismutase). Copper in ceruloplasmin has a well documented role in oxidising Fe before it is transported in the plasma (Lonnerdal and Uauy 1998).

Copper is also involved with an enzyme that strengthens connective tissue (lysyl oxidase) and in brain neurotransmitters (dopamine hydroxylase and peptidyl alpha amidating monoxygenase).

The few Cu containing proteins found in the cytoplasm are involved in protection and detoxification of Cu and in the cellular response to Cu levels. Most features of severe Cu deficiency can be explained by a failure of one or more of these Cu dependent enzymes.

22.3.5 Iodine (I)

Iodine is an essential element found in trace amounts in the human body (15–20 mg). It is stored almost entirely within the thyroid gland (Zimmermann 2011). According to Australian Nutrition Foundation (2003), I is defined as a chemical element that is required for growth and survival.

Iodine functions as a component of thyroid hormones (triiodothyronine, T3 and thyroxine, T4) which play a vital role in the regulation of metabolic processes such as growth and energy expenditure (Zimmermann 2011). Iodine is also essential to the normal development of the foetal nervous system. It regulates the effect of oestrogen on breast tissue, and it may protect against the effects of radioactivity. Additionally, the extra-thyroidal I has other functions; it removes toxic chemicals and biological toxins, suppresses autoimmunity, strengthens the T-cell adaptive immune system and protects against abnormal growth of bacteria in the stomach, *Helicobacter pylori* in particular (Miller 2006).

22.3.6 Vitamin B12 – Cobalt (Co)

The importance of cobalt (in the form of vitamin B12) for humans was discovered in 1948 (Rickes et al. 1948), while the full chemical structure of B12 was not identified until the 1960s. Vitamin B12 is the only vitamin synthesized solely by certain microorganisms (algae, yeast), many of which are abundant in soils, and the only vitamin containing a metal ion, Co (Carmel 2007).

B12 owes its chemical name ‘cobalamin’ to the Co at the centre of its molecular structure. Vitamin B12 can be stored in the body for extended amounts of time, even for years. It is primarily stored in the liver; other places are the heart, spleen, brain, kidneys, bones, and muscles (Seatharam and Alpers 1982).

Two forms of vitamin B12 are used in the human body, methylcobalamin and 5-deoxyadenosyl cobalamin.

Cobalamin is a cofactor for only two enzymes, methionine synthase and L-methylmalonyl-CoA mutase (Carmel 2007). Proper functioning of methionine synthase is needed for methylation of DNA which is important in cancer prevention. Inadequate function of methionine synthase can lead to an accumulation of homocysteine, which has been associated with increased risk of cardiovascular diseases. 5-deoxyadenosylcobalamin is required by the enzyme that catalyzes the conversion of L-methylmalonyl-CoA to succinyl-CoA, a substance required for the synthesis of haemoglobin, which is perhaps the most well known function of vitamin B12. Additionally, vitamin B12 is involved in the development of nerve cells; it is required for growth and repair of cells and for the proper movement of carbohydrates and fats through the body (Seatharam and Alpers 1982).

22.3.7 Vitamin B9 (Folic Acid)

Folic acid is named after the Latin word, “folium” for leaf. Synonyms for folic acid are vitamin B9, folacin, pteroylglutamate, pteroylmonoglutamate, and folate. Folic acid is composed of three parts; pteridine connected to para-aminobenzoic acid (PABA), which forms pteroic acid. Finally, pteroic acid is connected to glutamic acid to form folic acid. These compounds must be present for activation of folic acid (Hathcock 1997). Folic acid helps in the metabolism of several amino acids such as histidine, glycine, serine, and methionine. This vitamin is essential for cell division, production of DNA and RNA, and assists in the prevention of changes in DNA. Folate is essential for the formation of red and white blood cells in the bone marrow and for their maturation and is a single carbon carrier in the formation of heme. Folic acid is also involved in the synthesis of neuromediators which play a key role in the metabolism of the brain and nerve tissues (De Walle and De Jong-Van Den Berg 2002).

22.3.8 Manganese (Mn)

Manganese is essential for life. The human body contains approximately 20 mg of Mn most of which is found in the liver, bones, and kidneys. Manganese serves two primary biochemical functions in the body, it activates specific enzymes, and it is a constituent of several metalloenzymes (Davis and Greger 1992). The enzymes Mn activates include hydrolases, decarboxylases, kinases, and transferases.

The manganese metalloenzymes include arginase, pyruvate carboxylase, glutamine synthetase, and Mn superoxide dismutase. Manganese functions with vitamin K in the formation of prothrombin. It participates in numerous

biochemical functions in the body including steroid and sulfomucopolysacchride biosynthesis, carbohydrate and lipid metabolism, and bone, blood clot, and protein formation (Friedman et al. 1987). Manganese is also essential for normal brain function, possibly through its role in biogenic amine metabolism (Bowman 2011).

22.4 Impact of Micronutrient Deficiencies on Human Health

22.4.1 Development of a Micronutrient Deficiency

Micronutrient deficiency occurs when needs and losses exceed dietary intake. A number of factors contribute to needs including normal metabolism, growth and development, pregnancy and lactation, and disease state (Gibson 2005). Loss of nutrients can also be part of normal physiological processes (e.g. menstruation, skin exfoliation) and it can be highly accelerated during disease and infections (Gibson 2005). Before we discuss the negative effects associated with deficits of individual micronutrients, it is important to identify various ways by which a mineral deficiency can arise.

Primarily, a micronutrient deficiency occurs as a consequence of an insufficient dietary intake of an essential nutrient. Diets of people living in developing countries are mostly based on plant foods with low concentrations of important micronutrients. Consumption of animal-protein foods such as meat, poultry and fish is often small, because of economic, cultural and/or religious constraints. As a result, the amount of micronutrients accessible for absorption from such diets is low, and is probably the main cause of micronutrient deficiencies.

A secondary, or conditioned, deficiency may occur even if the dietary content of the essential nutrient appears to be adequate (Keen et al. 2003). Conditioned deficiencies occur through several mechanisms (Fig. 22.2).

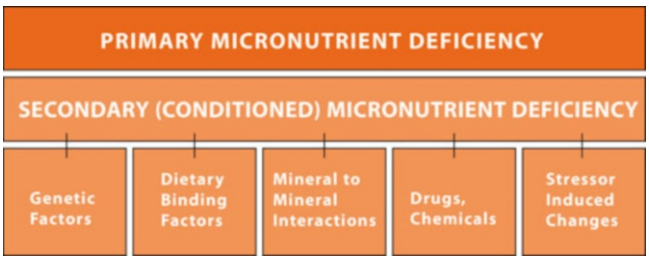


Fig. 22.2 Factors contributing to the development of a nutritional deficiency. Primary micronutrient deficiency is due to the inadequate dietary intake of a particular nutrient, while secondary micronutrient deficiency is caused by various factors

1. Genetic factors may be responsible for increased requirements of particular nutrients. For example, individuals with acrodermatitis enteropathica require a large amount of Zn in their diet because of a genetic defect in Zn absorption (Wang et al. 2002), whereas people with Menkes disease suffer from copper deficiency resulting from defects in the intracellular trafficking of this element (Keen 1998).
2. Interactions between the essential nutrients may also result in conditioned deficiencies. For example, dietary binding factors, including phytate and polyphenols, can create a complex with essential nutrients in the gut and consequently limit their absorption.

It is well known that Fe and Zn deficiencies occur in individuals who consume phytate-rich diets (Hambidge 2000).

In the research literature phytic acid and phytate are terms that are often used as synonyms, which can lead to misunderstandings. To explain, phytic acid is a free acid, a hexa phosphorous acid ester of the 6-hydroxyl group cyclic alcohol myo-inositol that has an essential characteristic for certain microorganisms. The correct chemical term for phytic acid is myo-inositol 1,2,3,4,5,6 hexakis dihydrogen phosphate. Phytates are the salts of phytic acid. The international abbreviation PA stands for both phytic acid and for phytate. The antinutritive effect of phytic acid is based on its molecular structure. At complete dissociation, the six phosphate groups of phytic acid carry twelve negative charges which, in weak acidic to neutral pH conditions, bind various di- and tri-valent cations (Ca, Mg, Fe, Zn, Cu, Mn) into a stable complex (Fig. 22.3).

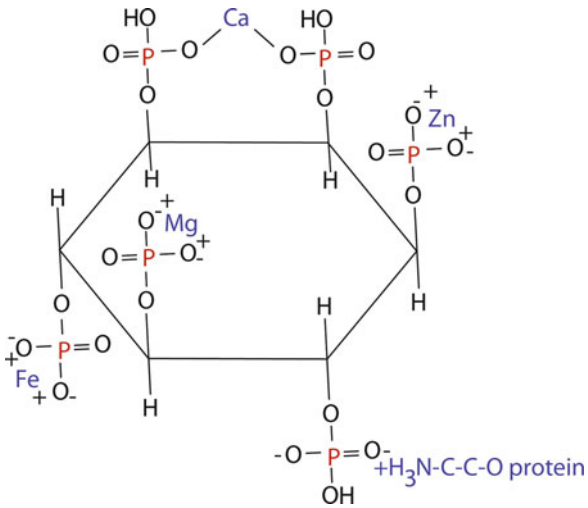


Fig. 22.3 Structure of phytic acid showing six phosphorous acid molecules bound to a phytic acid molecule, with negative charges binding divalent cations (Modified from Sohail and Roland 1999)

The formation of insoluble phytate mineral complexes in the intestinal tract prevents mineral absorption, and thus reduces the bioavailability of essential minerals (Elkhalil et al. 2011).

Zinc appears to be the trace element whose availability is the most influenced by phytic acid (Rimbach et al. 2008). Similarly, non-heme Fe is poorly absorbed in presence of phytates which irreversibly bind Fe in the gastrointestinal lumen, making it unavailable for absorption.

In addition, a number of studies over the years showed a 12–15 fold decrease in the absorption of Fe and Zn in the presence of phytate (Egli et al. 2004; Bohn et al. 2004).

A negative effect of phytate on absorption of other nutrients (manganese, calcium, copper) has also been observed (Lonnerdal et al. 1989).

Polyphenols, mostly present in the outer layer of the cereal grain, have also been shown to have inhibitory effect on absorption of certain micronutrients. Polyphenols are thought to act through the formation of complexes between the hydroxyl groups of the phenolic compounds and mineral molecules. Generally, the extent of inhibition varies inversely with the condensed polyphenol content (Elkhalil et al. 2011).

3. Mineral-mineral interaction can happen in a variety of ways. If the metabolism of one nutrient is dependent on the metabolism of another, a deficit of one may influence the availability of the other. Another interaction may occur when minerals share a common transport site or transport ligand (e.g. the absorption of Fe and metals close to Fe).

   The presence of competing metals in the diet can have a marked effect on their absorption.

   For example, zinc and cadmium are very strong antagonists of copper absorption (Danzeisen, et al. 2002). A number of studies over the years showed that absorption of metals close to Fe (e.g. cobalt, nickel, manganese and cadmium) is enhanced in Fe deficient subjects (Lynch 1997). Lead absorption is increased in Fe-deficient human subjects (Lynch 1997). The enhanced absorption of Mn and Co in Fe deficient rats can be inhibited competitively by Fe, and vice versa. Prolonged or excessive supplementation of the diet with zinc can lead to copper deficiency, very probably due to the Zn stimulated induction in the intestinal mucosa, mechanism not clear (Lynch 1997). Large doses of Fe may also reduce copper absorption, which has been demonstrated in children suffering from severe protein malnutrition (Kwashiokor) in Peru (Anderson 1981).
4. A fourth mechanism by which a conditioned deficiency can arise is through an effect of drugs or chemicals on the metabolism of the nutrient. Examples are diuretics that increase urinary loss of some nutrients. The increased turnover of folic and ascorbic acid, which occurs via oxidative damage secondary to smoking, is an example of how one toxin can induce essential nutrient deficiencies (Van Wersch et al. 2002).
5. A fifth mechanism by which a conditioned deficiency arises is through stressor-induced physiological changes in micronutrient metabolism. For instance, diabetes, renal diseases, hypertension modify the metabolism of some minerals, including Zn, Fe and copper (Keen 1996), possibly because of the induction by stress of an acute phase response. Inflammatory bowel disease often leads to deficiencies of vitamin A, K, D, Zn and Fe, while patients with asthma have lower concentrations of Zn and Se; higher Cu concentrations, and higher Cu/Zn and Cu/Se ratios (Goh and Morain 2003).

22.4.2 The Consequences of Deficiency of Individual Micronutrients on Human Health

22.4.2.1 Zinc Deficiency in Humans

The first cases of human Zn deficiency were described in early 1960s, in male adolescent dwarfs from the Middle East consuming plant-based diets (Prasad et al. 1963). Since that time, Zn deficiency has been identified in many other regions of the world. During the past 50 years, it has become evident that deficiency of Zn in humans is quite prevalent, and today it affects over two billion people in the developing world (WHO 2011). In 2002 Zn deficiency was incorporated as a major risk factor to the global and regional burden of disease, along with Fe, vitamin A and I deficiency (WHO 2002).

Zinc binds with 925 proteins in humans. Because Zn interacts with so many proteins Zn deficiency in humans is manifested in a wide range of symptoms, from acute, life-threatening problems to mild subclinical or marginal disorders which may only vaguely disturb well being (Walsh et al. 1994). The acute problems are often seen in profoundly ill patients treated in hospitals, whereas subclinical problems may be so indistinct that they often stay unrecognised (Prasad 2003).

The clinical manifestations of severely Zn deficient subjects include bullous pustular dermatitis, diarrhoea, alopecia, mental disturbances, and intercurrent infections due to cell-mediated immune disorders. Growth retardation, male hypo-gonadism, skin changes, poor appetite, anorexia, mental lethargy, abnormal dark adaptation, and delayed wound healing are usual manifestations of moderate deficiency of Zn (Barceloux, 1999). Systemic intestinal inflammation associated with Zn deficiency can lead to Fe deficiency anemia (Roy 2010).

Most recently, mild Zn deficiency has been shown to cause colitis in rats via impairment in the immune response (Iwaya et al. 2011), which shows a critical role of Zn in controlling inflammatory reactions in the intestine.

Additionally, a mild or marginal deficiency of Zn in humans is characterized by neurosensory changes, oligospermia in males, decreased serum testosterone in males, hyperammonemia, decreased serum thymulin activity, decreased IL-2 production, decreased natural killer cell activity, alterations in T cell subpopulations (Beck et al. 1997), impaired neuropsychological functions, and decreased ethanol clearance (Prasad 2002).

Zinc deficiency is associated with many diseases, for example, malabsorption syndrome, chronic liver disease, chronic renal disease, sickle cell disease, diabetes, malignancy, and other chronic illnesses (Prasad 2003). In these conditions, deficiencies of other micronutrients such as vitamins and other trace elements are often seen, which means that deficiency of one micronutrient rarely occurs in isolation.

22.4.2.2 Iron Deficiency in Humans

The human population is surprisingly Fe-deficient despite the earth and its rocks and soils being quite rich in Fe. The World Health Organisation (1995, 2005, 2011) on its website estimated in 2005 the global incidence of Fe deficiency to be between four and five billion people, and the current website identifies two billion severely deficient, that is, anemic. It appears that more than half the total problem is dietary in origin.

Total body Fe (measured by ferritin), transport Fe (measured by transferrin saturation), serum Fe, and other hematologic and biochemical markers are used to describe the degrees of Fe deficiency. Iron depletion refers to the earliest stage of diminishing Fe stores in the setting of insufficient Fe supply. Iron deficiency (without anemia) develops as these Fe stores are depleted further and begin to impair hemoglobin synthesis. Finally, Fe deficiency anemia results when the Fe supply is insufficient to maintain normal levels of hemoglobin.

Iron deficiency anemia is most severe and widespread among growing children and pre-menopausal women, because adult males even in old age resist anemia despite poor diets in resource-poor countries (Markle et al. 2007). Common symptoms of Fe deficiency are: tiredness, lethargy, shortness of breath (dyspnoea) and palpitations (irregular heartbeat).

Most Fe-deficient women and children are debilitated to some degree in both physical and mental work capacity.

Iron deficiency anemia results in increased child and maternal mortality, slower child development, delayed cognitive development in children and adolescents and decreased work productivity (Andrews 1999).

22.4.2.3 Selenium Deficiency in Humans

Selenium is an important trace element that has heightened interest because of its antioxidant and anticancer properties. Selenium deficiency has been mainly identified in people inhabiting geographical regions notable for low soil selenium content, such as volcanic regions (Poland, former Yugoslavia, China, and Russia). Deficiency symptoms for Se are difficult to determine and controversial in the research literature. Intake of Se that is borderline or only mildly deficient has not been connected with specific symptoms. With prolonged and severe deficiency, symptoms usually centre around two of the body areas, the heart and the joints (Diplock 1992). Keshan disease is a well known example of an endemic cardiomyopathy that has been observed in children, adolescents and pregnant women in the Keshan region of China, a place where Se levels in soil and food are extremely low (Lederer 1986). Similarly, Kachin-Beck disease is an osteoarthropathy reported to occur in Se-deficient populations (Navarro-Alarcon and Lopez-Martinez 2008). Deficient Se intake may also contribute to carcinogenesis, possibly due to the failure of GSH-Px to scavenge free radicals efficiently (Rayman 2000). Finally, numerous reports associate Se deficiency with several reproductive and obstetric complications including male and female infertility, miscarriage, preeclampsia, foetal growth restriction, preterm labor, gestational diabetes, and obstetric cholestasis (Sager 2006).

22.4.3 Copper (Cu) Deficiency in Humans

Copper has been recognized as an essential nutrient since the 1920s; however the first good evidence of a nutritional Cu deficiency was provided by studies on malnourished Peruvian children (6 months to 3 years) in the 1960s (Cartwright and Wintrobe 1964).

Disruptions in Cu homeostasis have implications in a number of diseases, and deficiencies in Cu absorption have a number of negative effects.

Unfortunately, most research into Cu deficiency has focused on acute, severe deficiency, which is relatively rare in humans on typical, varied diets. Marginal, chronic deficiency, however, is much more common. The determination of Cu needs and symptoms of marginal deficiency are complicated by the fact that while Cu deficiency doesn't necessarily lower the level of Cu-dependent enzymes, it does significantly lower their activity (Turnlund 1998).

Copper enzymes are widely distributed within the body; they perform several diverse functions (Cartwright and Wintrobe 1964) including transport of oxygen and electrons, catalysis in oxidation-reduction reactions and the protection of the cell against damaging oxygen radicals. At least ten enzymes are known to be dependent upon Cu for their

function. Inadequate Cu uptake during development interferes with the activity of cuproenzymes that leads to increased oxidative stress, decreased availability of nitric oxide, abnormal Fe metabolism, problems with cross linking in the extracellular matrix and altered cell signalling; in turn, these reactions affect a variety of tissues and organs including the brain, heart, lungs and skin (Uriu-Adams et al. 2010). Abnormalities associated with Cu deficiency include: anemia, neurological damage, hypercolesterolaemia, cardiomyopathy, osteoporosis and impaired immune function (Cordano 1998). Other changes may also develop, including subperiosteal haemorrhages, hair and skin depigmentation, and defective elastin formation (Lonnerdal 1998).

Despite the scarcity of Cu in the environment, overt Cu deficiency is rare in humans (Cordano 1998). Deficiency has been observed in individuals with restricted diets and in premature infants fed cow's milk formulas (Cordano 1998). Chronic ingestion of large quantities of Zn may reduce the efficiency of Cu absorption and has been reported to cause deficiency (Bertinato et al. 2003). Also, individuals suffering from malnutrition and severe malabsorption syndromes as well as patients undergoing chelation therapies are at increased risk of deficiency (Bleackley and MacGillivray 2011).

22.4.3.1 Iodine Deficiency in Humans

Current calculations by the International Council for Control of Iodine Deficiency Disorders have suggested that 29 % of the world's population, or approximately 1.6 billion people, live in areas of I deficiency and are at risk for IDD. Nearly 30 % of the world's 241 million school-age children remain I-deficient (ICCIDD 2011).

Those people at highest risk for IDD are located primarily in mountainous regions far from the sea such as the Himalayas, the European Alps, the Andes, and the mountains of China.

Population effects of severe I deficiency include endemic goiter (enlarged thyroid gland), decreased fertility rate, increased infant mortality, and mental retardation. If iodine deficiency is extremely severe, thyroid hormone production falls, and patients become hypothyroid.

Hypothyroidism in the fetus, neonates, and in young children prevents central nervous system maturation, especially neuronal myelination, leading to permanent mental retardation. In the most severe form, the constellation of mental retardation and growth abnormalities is called cretinism. Cretenism is rare, but is still seen in regions of severe I deficiency in Southern and Eastern Europe, Asia, Africa, and Latin America.

22.4.3.2 Vitamin B12 and Folate Deficiencies

B12 and folate are B complex vitamins that are necessary for normal red blood cell formation, tissue and cellular repair, and DNA synthesis. A vitamin B12 and/or folate deficiency reflects a chronic shortage of one or both of these vitamins. Since the body stores 3–6 years worth of vitamin B12 and about a 3 month's supply of folate in the liver, deficiencies and their associated symptoms can take months to years to manifest in adults (Seatharam and Alpers 1982). Infants and children will show signs of deficiency more rapidly because they have not yet established extensive reserves.

A cobalt deficiency has never been produced in humans. Signs and symptoms of one are actually those of a vitamin B12 deficiency. Untreated vitamin B12 deficiency can lead to serious complications, such as micrositic anemia, nerve damage, and growth abnormalities. Indications of a deficiency of vitamin B12, when they do reach a stage where they have shown up, can be quite severe: fatigue, paleness, anorexia, mental confusion, delusions, paranoia, weight loss, etc. Vitamin B12 deficiency is estimated to affect 10–15 % of individuals over the age of 60 (Baik and Russell 1999).

Folic acid deficiency causes cell multiplication to slow down, particularly regarding cells with a high renewal rate such as blood cells and cells in the intestine, liver and skin.

Signs of this hypovitaminosis are therefore anaemia, depression, dementia, increased rate of heart disease, digestive and neurological problems and mucous membrane problems (gums, for example) (Flood and Mitchell 2007). In pregnant women, the consequences of folic acid deficiency are even more dramatic: anomalies with the development of maternal tissue (placenta, blood circulation), anomalies with foetal development (spina bifida, anencephalia), delayed growth of the foetus, increase in the risk of premature birth and low folate reserves in the baby (Van Wersch et al. 2002).

22.4.3.3 Manganese Deficiency in Humans

The requirements of Mn for plants are generally higher than those for animals owing to its high requirement in the synthesis of cell wall materials which animal cells do not have (Graham and Stangoulis 2001). Consequently, plant-based diets should generally be adequate in Mn for humans, except in areas of severe Mn deficiency in the soil itself. While an outright Mn deficiency has not been observed in the general human population, suboptimal Mn status may be more of a concern. As reviewed by Freeland-Graves and Lianis (1994), several disease states have been associated with low levels of serum Mn. These include epilepsy, exocrine pancreatic insufficiency, multiple sclerosis, cataracts, and osteoporosis. In addition, several inborn errors of metabolism have been associated with poor Mn status (e.g. phenylketonuria, maple syrup urine disease). Impaired fertility, growth retardation, birth defects, bone malformations, seizures, and general weakness may also result from Mn deficiencies (Friedman et al. 1987).

22.5 Interaction Between Micronutrient Deficiencies

It is now well recognised that many population groups in developing countries do not suffer from single but rather from multiple nutrient deficiencies. Multiple deficiencies are a consequence of a range of common factors, including diets lacking in adequate nutritional quality mainly related to poverty, seasonal variation in food availability, and cultural food practices. Low bioavailability of nutrients, especially from plant sources, illness or infection, further intensify deficiency due to poor nutrient utilization. Existing public health strategies mostly focus on one micronutrient, for example, the National Nutritional Anaemia Control Programme and the National IDD Control Programme.

Major intervention programs address mostly Fe, I and vitamin A deficiencies, mostly as single nutrient interventions, with fewer programs operating for other limiting essential trace elements (Gibson 2003).

It cannot be said with certainty that there is a common underlying cause of multiple micronutrient deficiencies, or that one micronutrient deficiency specifically leads to another deficiency. However, it is obvious that micronutrient deficiencies are occurring simultaneously in many regions of the world (Fig. 22.4). A diet rich in phyate and low in animal proteins, common in many developing countries, can lead to insufficient intake and absorption of both Fe and Zn (Kennedy et al. 2003). Dijkhuizen et al. (2001) observed concurrent occurrence of vitamin A, Fe, and Zn deficiencies in lactating mothers and their infants in rural villages in West Java, Indonesia. In addition, Anderson et al. (2008) demonstrated a high prevalence of co-existing micronutrient deficiencies in Cambodian children, with Zn (73 %) and Fe (71 %) being the most prevalent deficiencies. Vitamin A, Zn, Fe, and I have been mentioned, but there are many more significant overlaps. For example, according to Singh (2004) selenium deficiency may impair utilization of I because it is a key component of the enzyme required to convert thyroxine to triidothyronine.

Vitamin A deficiency can aggravate both Fe and Zn deficiencies (Welch 1986). Similarly, vitamin B12 deficiency can cause anemia (Fe-resistant or pernicious anemia).

More importantly, newly published mechanisms of the regulation of Fe uptake by dietary Zn in humans (Graham et al. 2012) show that Zn deficiency could be the cause of up to half of the global burden of Fe deficiency anemia.

It is important to treat concurrent micronutrient deficiencies together, as the positive effects of this approach have been reported in number of studies. For example, Shoham and Youdim (2002) investigated the effect of 4-week Fe and/or Zn treatments on neurotransmission in the hippocampal region in rats. Iron or Zn alone was not effective whereas together they caused a significant increase in ferritin-containing mossy fiber cells (cells important for memory and learning). This is the classical response to the addition of two limiting essential nutrients acting together on a physiological or developmental pathway.

Ramakrishnan et al. (2004) undertook meta-analyses of such randomized controlled interventions to assess the effects of single vitamin A, Fe, and multi-micronutrient (Fe, Zn, vitamin A, vitamin B and folic acid) interventions on the growth of toddler children. In their summary of around 40 different studies, they clearly found greater

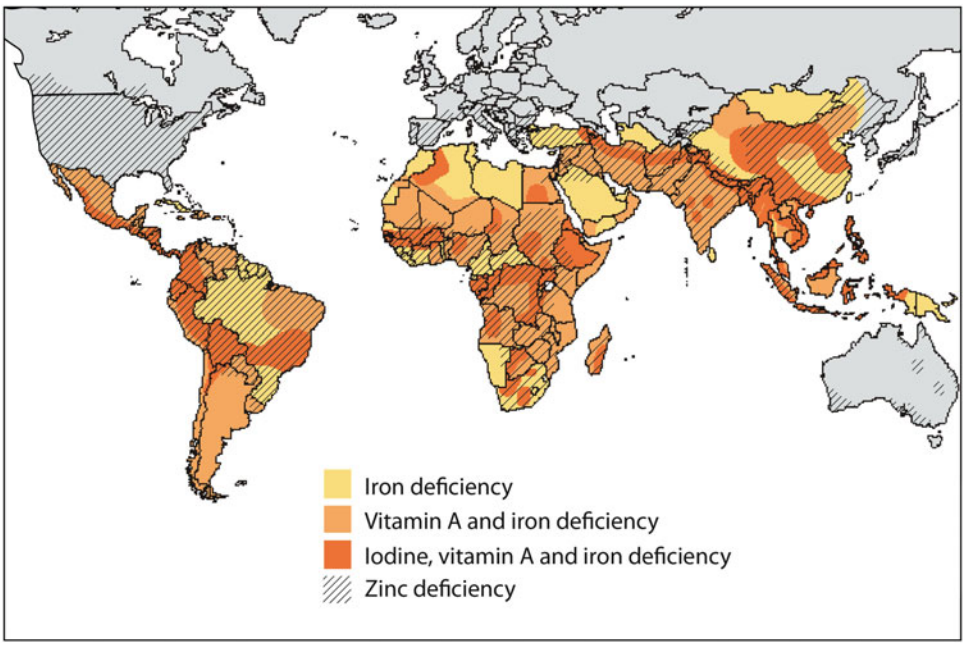


Fig. 22.4 Prevalence of common micronutrient deficiencies in developing countries. (Modified from Welch and Graham 1999)

benefits from multi-micronutrient interventions that they explained by the high prevalence of concurrent micronutrient deficiencies and the positive synergistic effects between these nutrients at the level of absorption and/or metabolism (for example, vitamin A and Fe, vitamin A and Zn, Fe and Zn, all three) (Ramakrishnan et al. 2004).

To sum up, micronutrient deficiencies often coexist in the same at-risk individuals owing to poor quality diets and recurring illnesses. Therefore, multiple, rather than single, micronutrient interventions are needed to positively affect various health and nutrition outcomes.

22.6 Agricultural Interventions for Reducing Micronutrient Deficiencies in Humans

Traditionally, agricultural programs have not been designed to promote human health, but are mainly focused on increasing yields, productivity and general food availability. For many years, the agriculture, nutrition and health sectors have operated as separate entities, and policies and government structures have been planned without looking at the interactions among these sectors (Bouis and Welch 2010).

Nowadays, more and more people realise that agricultural interventions may have a positive impact on nutritional status of humans, particularly if they are clearly implemented with that objective. Collaboration among health, agriculture and nutrition sectors is essential in the development of programs that can reduce the number of people suffering from micronutrient deficiencies in a most effective way.

Plant foods can serve as dietary sources of all essential minerals required by humans. Unfortunately, mineral concentrations of some plants are very low. This problem of low mineral density is particularly important in staple foods, such as cereal grains and tuber crops, which make up a large proportion of daily food intake in the developing world (Calloway 1995 in Grusak 2002). Thus, various agricultural interventions are used to increase the mineral content of these foods in order to ensure adequate intake of dietary minerals in all individuals. One well known agricultural strategy is crop mineral biofortification which is achieved by agronomic fertilisation with metal salts. Nutrient concentrations can be increased greatly, from less than double for Zn in rice to a 100 times in an example of selenium in wheat (Lyons et al. 2004). However, this approach involves some technology and costs. On the other hand, crop biofortification by breeding or by genetic modification offers a sustainable and low-cost way to provide essential micronutrients to people in both developing and developed countries (Graham et al. 2007).

Biofortification refers to the use of traditional crop breeding practices to increase the micronutrient concentrations in crops, in order to target the specific micronutrient deficiency of a target population (Nestel et al. 2006). It involves classical breeding approaches designed to characterize and exploit genetic variation for mineral content as well as new approaches involving gene discovery and directed genetic modification for creating new plant lines with improved mineral qualities (Graham and Welch 1996).

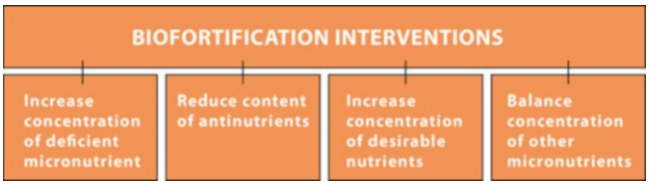


Fig. 22.5 The biofortification strategies for improving absorption of micronutrients in humans

The ultimate aim of each biofortification intervention is to increase the absorption of a certain micronutrient from a human diet. In this instance, four main biofortification strategies can be applied to crops (Fig. 22.5).

1. Increase the concentration of a deficient micronutrient

   Due to genotypic variation, the concentrations of micronutrients in grains of major crops can be quite diverse. For example, Fe concentration in the edible parts of differing varieties of maize (*Zea mays*), wheat (*Triticum* spp.), rice (*Oryza sativa*) and common beans (*Phaseolus vulgaris*) falls between 6 and 70 mg/kg dry weight. Similarly, higher concentrations of both Fe and Zn have been measured in emmer wheat (*Triticum turgidum* ssp. *dicoccoides*), a progenitor of domesticated wheat, than in a modern durum wheat cultivar (Grusak 2002). Plant mineral concentrations do also vary among plant tissues (e.g. leafy structures versus seeds) (Abbo et al. 2000).

   One way of increasing the concentrations of a deficient micronutrient through breeding processes is by identification and reintroduction of functional genes (responsible for higher concentrations of a particular micronutrient/group of nutrients) from wild relatives to domesticated varieties (Grusak 2002). In recent years, plant scientists have begun to identify genes which encode proteins of relevance to the membrane transport of various mineral nutrients. These protein products include membrane transporters, as well as proteins that facilitate the availability of minerals at the root-soil interface (Grusak 2002).

   The data are now available for genes encoding various divalent metal transporters (e.g. for Fe^{2+} Zn^2, Mn^2, Cu^2, Ni^2, and others) (Elde et al.1996; Grotz et al. 1998). The overexpression of a Ca^{2+}/H^+ antiporter in the model plant, *Arabidopsis thaliana*, has led to a moderate increase in leaf Ca concentration (Hirsch 1999) which demonstrates the potential usefulness of this approach. Similarly, the

genes for ferric reductases have been identified (Robinson et al. 1999).

Overexpression of the Fe-storage protein ferritin in rice grains has also been reported to result in a threefold (Lucca et al. 2001) increase in Fe concentrations. New discoveries clearly show that improvement of mineral content of plants by transgenic manipulation is a feasible option; however much more work is required before this technology can have a widespread use.

2. Reduce the content of substances ('antinutrients') that hinder the absorption of micronutrients (phytate and polyphenols)

   Phytic acid and polyphenols are considered as antinutrients because by mixing with minerals, they reduce the absorption of micronutrients in the human gut. The presence of high levels of phytic acid in cereals contributes to reduced absorption of micronutrients (particularly of Fe and Zn). One of the approaches for improved absorption of minerals in humans is the isolation of 'low phytate' cereals and legume genotypes and reduction of phytate in specific seed tissues (Joyce et al. 2005). Phytic acid fulfils essential biological functions in plants; it represents an important metal cation reserve; it also acts in the leaves in the signalling cascade triggered by drought/osmotic stress; maintain basal resistance against a wide range of pathogens, etc. (Guttierit et al. 2006). In addition, phytic acid and polyphenols possess diverse biologically beneficial properties for humans (antineoplastic, antioxidant and anti-inflammatory) (Trowel 1973; Welch 2008).

   Reduction of phytic acid and polyphenol concentrations should be done with caution as there is an obvious need for a careful balance between the beneficial and antinutritive properties of these nutrients. Therefore, the first priority of breeding programs should be to increase the content of a particular micronutrient rather than to decrease phytate concentrations of major crops (Bouis and Welch 2010).

3. Increase the concentration of substances favouring absorption of a certain desirable micronutrients (e.g. ascorbate, prebiotics)

   Prebiotics (fructans and arabinoxylans) are nutrients naturally present in cereal grains, resistant to digestion and absorption in the human upper gastrointestinal tract but selectively fermented in the colon by gut microflora (Roberfroid 2005). This favorable fermentation stimulates the production of short chain fatty acids (SCFAs) that confer several health benefits to humans, one of them being increased absorption of certain micronutrients.

   In recent years, there has been more and more efforts put towards increasing fructan and arabinoxylan levels in grains by genetic manipulation. Two quantitative trait loci (QTLs) for increased wheat fructan content have been identified (QGfc.awww-6D.2 on chromosome 6D and QCfc.aww-7A.1 on chromosome 7A (Huynh et al. 2008). Lines having alleles at these loci had 17–27 % higher fructan content.

   Furthermore, two major QTLs (QGax.aww-2A.1and QGax.aww-4D.1 on chromosome 6B) that influence grain arabinoxylan concentration has been identified (Nguyen et al. 2011). Wheat lines carrying the favourable alleles at these loci have higher arabinoxylan content of up to 19 %. These new findings are important as they can help in breeding grains with higher fructan and arabinoxylan content, which consequently may mean improved absorption of certain micronutrients. However, the entire efficacy of these interventions is yet to be confirmed.

   It is suggested that ascorbic acid promotes Fe absorption from the diet by reducing the negative effect on Fe absorption of certain ligands such as phytates and tannins present in the diet. In recent years, more and more research is devoted towards understanding the pathways leading to ascorbic acid biosynthesis in plants. The overexpression of GalUR gene enhanced vitamin C content of certain plants (Agius et al. 2003).

   Similarly, overexpression of the enzyme DHAR (dehydroascorbate reductase) responsible for recycling ascorbate has also been shown as an effective way in increasing the concentration of this vitamin in wheat plants (Khan et al. 2012).

4. Balance the concentrations of other micronutrients

   Increasing the concentrations of individual micronutrients through biofortification approaches is important and could significantly contribute to the nutritional enhancement of the species of interest; however, this strategy can be successful only if the concentrations of other essential nutrients are taken into account and balanced appropriately. The increase of a particular micronutrient by more than what brings the deficient nutrient up to a relative abundance roughly matching that of the other nutrients in the system should be avoided because replacing one imbalance (e.g. Zn too low) with another (e.g. Zn too high) will induce a deficiency of another micronutrient and so give little progress in health (Graham et al. 2012). Additionally, nutritional enhancement of crops should be achieved without a reduction in other major traits, such as agronomic performance, crop yield and protein concentration.

   Breeding of nutrient-rich staple food crops is the main goal of a number of international organisations. The Consultative Group on International Agricultural Research (CGIAR) Global Challenge Program,

HarvestPlus (http://www.harvestplus.org/), aims to reduce micronutrient malnutrition, mainly provitamin A, Zn, Fe, Se and I, through various biofortification programs. AgroSalud (www.agrosalud.org) is another international consortium supporting the production and dissemination of Fe and Zn-rich bean and rice varieties in Latin America.

Delivery of more micronutrients into staple food crops through plant breading is a huge task and the challenge is to minimize the number of genes involved to accomplish this end. It is feasible to help meet this challenge by use of supplemental fertilizers especially on soils inherently low in these nutrients. So far, little prospect of breeding for high Se or I content has been seen (Welch 1986) so fertilizer strategies are still the most appropriate for these (Welch 1996; Cao et al. 1994).

In the same way, the improvement of mineral content of plants by transgenic manipulation is also a viable option; however many of the molecular elements relevant to mineral transport still need to be identified before this technology can have a widespread utility. Our knowledge base is most lacking in the area of mineral loading into the phloem pathway, which is a critical process in the delivery of mineral nutrients to developing seeds (Grusak 2002). Due to simpler genetics, it may prove more effective to breed for increased bioavailability promoting substances (e.g. prebiotics) to enhance the absorbability in the human gut of minerals in staples than to increase mineral concentrations in staple food grains (Graham et al. 2007). However, further research is required to delineate positive and negative health effects of reduced phytic acid concentrations, important for the health of both plants and humans, in plant tissues.

Of other challenges more related to the human nutrition field, there is a need for the development of sensitive and specific biomarkers that can be used with confidence across a range of applications and conditions (Raiten et al. 2011). Similarly, large-scale trials of efficacy of biofortification programs in human populations still need to be performed.

Ideally, in the future, programs would be greatly improved with the inclusion of biomarkers that could simultaneously reveal the status of multiple micronutrients, going beyond the usual big four (i.e., vitamin A, Fe, Zn, I) and in line with assessment of status associated with optimal health (Wasantwisut and Neufeld 2012). Furthermore, extending our understanding of absorption mechanisms of individual/multiple micronutrients in humans will help to maximise the benefits of various biofortification approaches and to reduce any possible drawbacks.

In the end, the control of micronutrient deficiencies is a major goal of the whole scientific community. A successful strategy requires strong collaboration among scientists of several disciplines (plant physiology, biotechnology, human nutrition, epidemiology and medical care) and is therefore crucial for offering the most effective solutions. Improving the status in one micronutrient or even several micronutrients simultaneously in the case of multiple deficiencies can have wider benefits to health of people in developing countries. Multiple micronutrient interventions will help not only to accelerate progress in reducing micronutrient deficiencies but also to address several of the underlying drivers, poverty, food insecurity, gender inequity, and lack of access to basic services. The largest effect of well designed biofortification interventions will undoubtedly be observed in developing countries, but measurable and cost-effective outcomes should also occur in developed countries, even in populations characterised by appropriate diets, given that secondary micronutrient deficiencies can occur by multiple means.

See Also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements), Chapter 6 (Uptake of Elements from a Biological Point of View), Chapter 7 (Biological Functions of the Elements), Chapter 12 (Arsenic in Groundwater and the Environment), Chapter 15 (Bioavailability of elements in soil), Chapter 17 (Soils and Iodine Deficiency), Chapter 18 (Geophagy and the Involuntary Ingestion of Soil), Chapter 19 (Natural Aerosolic Mineral Dusts and Human Health), Chapter 20 (The Ecology of Soil-Borne Human Pathogens), Chapter 23 (Environmental Medicine), Chapter 24 (Environmental Pathology)

References

Abbo S, Grusak MA, Tzuk T, Reifen R (2000) Genetic control of seed weight and calcium concentration in chickpea seed. Plant Breed 119:427–431

Agius F et al (2003) Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. Nature 21:177–181

Anderson A (1981) Nat Food Inst 52 (Soberg, Denmark)

Anderson VP, Jack S, Monchy D, Hem N, Hok P, Bailey KB, Gibson RS (2008) Co-existing micronutrient deficiencies among stunted Cambodian infants and toddlers. Asia Pac J Clin Nutr 17(1):72–9

Andrews N (1999) Disorders of iron metabolism. N Engl J Med 341:1986–1995

Baik H, Russell RM (1999) Vitamin B12 deficiency and elderly. Annu Rev Nutr 19:357–377

Barceloux DG (1999) Zinc. Clin Toxicol 37:279–292

Beck FW, Kaplan J, Fitzgerald JT, Brewer GJ (1997) Changes in cytokine production and T cell subpopulations in experimentally induced zinc-deficient humans. Am J Physiol 272:E1002–E1007

Bertinato J, Iskandar M, L'Abbe RM (2003) Copper deficiency induces the upragulation of the copper chaperone for Cu/Zn superoxide dismutase in weaning male rats. J Nutr 133(1):28–31

Blair MW, Astudillo C, Grusak M, Graham R, Beebe S (2009) Inheritance of seed iron and zinc content in common bean (*Phaseolus vulgaris* L.). Mol Breed 23:197–207
Bleackley R, MacGillivray RT (2011) Transition metal homeostasis: from yeast to human disease. Biometals 24:785–809
Bohn T, Davidsson T, Walczyk T, Hurrell RF (2004) Phytic acid added to white-wheat bread inhibits fractional apparent magnesium absorption in humans. Am J Clin Nutr Rev 79:418–423
Bouis HE, Welch RM (2010) Biofortification—a sustainable agricultural strategy for reducing micronutrient malnutrition in the global South. Crop Sci 50:S20–S32
Bowman B, Kwakye GF, Hernandez HE, Aschner M (2011) Role of manganese in neurogenerative diseases. J Trace Elem Med Biol 25(4):191–203
Bryce J et al (2008) Maternal and child undernutrition 4: effective action at the national level. The Lancet 371:510–526
Cakmak I (2009) Enrichment of fertilizers with zinc: an excellent investment for humanity and crop production in India. J Trace Elem Med Biol 23(4):281–289
Calesnick B, Dinan AM (1988) Zinc deficiency and zinc toxicity. Am Fam Physician 37:267–270
Cao X-Y et al (1994) Iodination of irrigation water as a method of supplying iodine to a severely iodine-deficient population in XinJiang, China. Lancet 344(8915):107–110
Calloway DH (1995) Human nutrition: food and micronutrient relationships. International Food Policy Research Institute, Washington, DC
Carmel R (2007) Haptocorrin (transcobalamin I) and cobalamin deficiencies. Clin Chem 53(2):367–368
Cartwright G, Wintrobe MM (1964) Copper metabolism in normal subjects. Am J Clin Nutr Rev 14:224–232
Clark LC, Combs GF, Turnbull BW, Slate EH, Chalker DK, Chow J et al (1996) Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin. JAMA 276:1957–1963
Cordano A (1998) Clinical manifestation of nutritional copper deficiency in infants and children. Am J Clin Nutr 67:1012S–1016S
Danzeisen R, Fosset C, Chariana Z, Page K, David S, McArdle HJ (2002) Placental ceruloplasmin homolog is regulated by iron and copper and is implicated in iron metabolism. Am J Physiol Cell Physiol 282:C472–C478
Darshan D, Anderson GJ (2009) Interacting signals in the control of hepcidin expression. Biometals 22:77–87
Davis C, Greger JL (1992) Longitudinal changes of manganese-dependent superoxide dismutase and other indexes of manganese and iron status in women. Am J Clin Nutr 55:747–752
De Walle H, De Jong-Van Den Berg LT (2002) Insufficient folic acid intake in the Netherlands: what about the future? Teratology 66 (1):40–43
Dijkhuizen MA, Wieringa FT, West CE, Martuti S, Muhilal (2001) Effects of iron and zinc supplementation in Indonesian infants on micronutrient status and growth. J Nutr 131:2860–5
Diplock A (1992) Selenium, antioxidant nutritions, and human diseases. Biol Trace Elem Res 33:155–156
Duxbury MA, Bodruzzaman M, Lauren JM, Sadat RM, Welch N, E-Elahi, Meisner CA (2005) Increasing wheat and rice productivity in the sub-tropics using micronutrient enriched seed. In: Andersen P. et al (ed) Micronutrients in South and South East Asia. Proceedings of international workshop, Kathmandu, Nepal. 8-11 Sept 2004. International Centre for Integrated Mountain Development (ICIMOD), Kathmandu, pp 187–198
Edison S, Bajel A, Chandy M (2008) Iron homeostasis: new players, newer insights. Eur J Haemotol 81:411–424
Egli I, Davidsson L, Zeder C, Walczyk T, Hurrell R (2004) Dephytinization of a complementary food based on wheat and soy increases zinc, but not copper, apparent absorption in adults. J Nutr 134:1077–1080
Eide D, Broderius M, Fett J, Guerinot ML (1996) A novel iron-regulated metal transporter from plants identified by functional expression in yeast. Proc Natl Acad Sci USA 93:5624–5628
Elkhalil A, Osman HA, Osman O (2011) Phytic acid analysis by different bacterial phytases. Aust J Basic Appl Sci 5(11):2295–2302
Flood V, Mitchell P (2007) Folate and vitamin B12 in older Australians. Med J Aus 186:321–322
Freeland-Graves J, Lianis C (1994) Models to study manganese deficiency. In: Klimis-Tavantzis DJ (ed) Manganese in health and disease. CRC Press, Boca Raton, pp 59–86
Friedman BJ, Freeland-Graves J, Bales CA et al (1987) Manganese balance and clinical observations in young men fed a manganese-deficient die. J Nutr 117(1):133–143
Genc C, Verbyla AP, Torun AA, Cakmak I, Willsmore K, Wallwork H, McDonald GK (2009) Quantitative trait loci analysis of zinc efficiency and grain zinc concentration in wheat using whole genome average interval mapping. Plant Soil 314:49–66
Gibson RS (2003) Concurrent micronutrient deficiencies in developing countries: problems and solutions. Proc Nutr Soc 28:21–34
Gibson R (2005) Principles of nutritional assessment, 2nd edn. Oxford University Press, Oxford
Goh J, Morain CA (2003) Nutrition and adult inflammatory bowel disease. Aliment Pharmacol Ther 17(3):307–320
Graham RD (1984) Breeding for nutritional characteristics in cereals. Adv Plant Nutr 1:57–102
Graham RD (2008) Micronutrient deficiencies in crops and their global significance Chapter 2. In: Alloway BJ (ed) Micronutrient deficiencies in global crop production
Graham CG, Cardano A (1939) Copper depletion and deficiency in the malnourished infant. John Hopkins Med J 124:139–150
Graham RD, Stangoulis JCR (2001) Trace element uptake and distribution in plants. J Nutr 133(5):1502S–1505S
Graham RD, Welch RM (1996) Breeding for staple food crops with high micronutrient density. International Food Policy Research Institute, Washington, DC
Graham RD, Ascher JS, Ellis PAE, Shepherd KW (1987) Transfer to wheat of the copper efficiency factor carried on rye chromosome arm 5RL. Plant Soil 99:107–114
Graham RD, Welch RM, Saunders DA, Ortiz-Monasterio I, Bouis HE, Bonierbale M, de Haan S, Burgos G, Thiele G, Liria R, Meisner CA, Beebe SE, Potts MJ, Kadian M, Hobbs PR, Gupta RK, Twomlow S (2007) Nutritious subsistence food systems. Adv Agron 92:1–74
Graham RD (2008) Micronutrient deficiencies in crops and their global significance. Chapter 2 in Alloway BJ (ed) Micronutrient deficiencies in global crop production (pp. 41–61), Dordrecht, Springer
Graham RD, Knez M, Welch RM (2012) How much nutritional iron deficiency in humans globally is due to an underlying zinc deficiency? Adv Agron 115:1–40
Gromer S, Eubel JK, Lee BL, Jacob J (2005) Human selenoproteins at a glance. Cell Mol Life Sci 62(21):2414–2437
Grotz N, Fox T, Connolly E, Park W, Guerinot ML, Eide D (1998) Identification of a family of zinc transporter genes from Arabidopsis that respond to zinc deficiency. Proc Natl Acad Sci USA 95:7220–7224
Grusak M (2002) Enhancing mineral content in plant food products. J Am Coll Nutr 21(3):178S–183S
Guttieri MJ et al (2006) Agronomic performance of low phytic acid wheat. Crop Sci 46:2623–2629
Haas E (2001) Selenium. Health world on-line available at: http://www.healthy.net/scr/article.aspx?ID=2068
Hambidge M (2000) Human zinc deficiency. J Nutr 130:1344S–1349S

Hathcock J (1997) Vitamins and minerals: efficacy and safety. Am J Clin Nutr 66:427–437
Hentze M, Muckenthaler MU, Andrews NC (2004) Balancing acts: molecular control of mammalian iron metabolism. Cell 117 (3):285–297
Hirsch IK (1999) Expression of Arabidopsis CAX1 in tobacco: altered calcium homeostasis and increased stress sensitivity. Plant Cell 11:2113–2122
Hotz C, Brown KM (2004) Assessment of the risk of zinc deficiency in populations and options for its control. Food Nutr Bull 25: S99–S199
Huynh B, Mather DE, Wallwork H, Graham RD, Welch RM, Stangoulis JCR (2008) Genotypic variation in wheat grain fructan content revealed by a simplified HPLC method. J Cereal Sci 48:369–378
ICCIDD (2011) International council for the control of iodine deficiency disorders. http://www.iccidd.org/
Iwaya H, Kashiwaya M, Shinoki A, Lee JS, Hayashi K, Hara H, Ishizuka S (2011) Marginal zinc deficiency exacerbates experimental colitis induced by dextran sulfate sodium in rats. J Nutr 141 (6):1077–1082
Joyce C, Deneau A, Peterson K, Ockenden I, Raboy V, Lott JNA (2005) The concentrations and distributions of phytic acid phosphorus and other mineral nutrients in wild-type and low phytic acid Js-12-LPA wheat (Triticum aestivum) grain parts. Can J Bot 83(12):1599–1607
Kawashima Y et al (2001) Dietary zinc-deficiency and its recovery responses in rat liver cytosolic alcohol dehydrogenase activities. J Toxicol Sci 36:101–108
Keen CL (1996) Teratogenic effects of essential trace metals: Deficiencies and excesses. In: Chang L, Magos L, Suzuki T (eds) Toxicology of metals. CRC Press Inc, New York, pp 997–1001
Keen CL, Uriu-Hare JY, Hawk SN, Jankowski MA, Daston GP, Kwik-Uribe CL, Rucker RB (1998) Effect of copper deficiency on prenatal development and pregnancy outcome. Am J Clin Nutr 67:1003S–1011S
Keen C et al (2003) The plausability of micronutrient deficiencies being a significant contributing factor to the occurrence of pregnancy complications. J Nutr 133(5):1597S–1605S
Kennedy C, Nantel G, Shetty P (2003) The scourge of hidden hunger: global dimensions of micronutrient deficiencies. FAO Food, Nutr Agric 32:15–16
Khabaz-Saberi H, Graham RD, Pallotta MA, Rathjen AJ, Williams KJ (2002). Genetic markers for manganese efficiency in durum wheat (*Triticum turgidum* L. var. *durum*). Plant Breed 121:224–22
Khan A et al (2012) Ascorbic acid: an enigmatic molecule to development and enviromental stress in plants. Int J Appl Technol 2 (3):97–111
Lederer J (1986) Sélenium et Vitamine E, Jouvence, France
Liu T (2003) Selenium. Chem Eng News 81(36):94
Lonnerdal B, Uauy R (1998) Genetic and environmental determinants of copper metabolism. Am J Clin Nutr 67:951
Lonnerdal B, Sandberg AS, Sandstrom B, Kunz C (1989) Inhibitory effects of phytic acid and other inositol phosphates on zinc and calcium absorption in suckling rats. J Nutr 119:211–214
Lonergan PF, Pallotta MA, Lorimer M, Paull JG, Barker SJ, Graham RD (2009) Multiple genetic loci for zinc uptake and distribution in barley (*Hordeum vulgare*). New Phytol 184:168–179
Lucca P, Hurrell R, Potrykus I (2001) Genetic engineering approaches to improve the bioavailability and the level of iron in rice grains. Theor Appl Genet 102:392–397
Lynch S (1997) Interaction of iron with other nutrients. Nutr Rev 55(4):102–110
Lyons GL, Stangoulis JCR, Graham RD (2004) Exploiting micronutrient interaction to optimize biofortification programs: the case for inclusion of selenium and iodine in the harvest plus program. Nutr Rev 62:247–252
Markle W, Fisher MA, Smego RA (2007) Understanding global health. McGraw-Hill Medicall, New York
Miller W (2006) Extrathyroidal benefits of iodine. J Am Physician Surg 11(4):106–110
Milne D (1998) Copper intake and assessment of copper status. Am J Clin Nutr 67:1041S–1045S
Mocchegiani E, Giacconi R, Muzzioli M, Cipriano C (2000) Zinc, infections and immunosenescence. Mech Ageing Dev 121 (1–3):21–35
Navarro-Alarcon M, Cabrera-Vique C (2008) Selenium in food and the human body: a review. Sci Total Environ 400:115–141
Nestel P, Bouis HE, Meenakshi JV, Wolfgang P (2006) Biofortification of staple food crops. Am Soc Nutr 136(4):1064–1067
Nguyen Van Lam H, Wallwork H, Stangoulis CRJ (2011) Identification of quantitative trait loci for grain arabinoxylan concentration in bread wheat. Crop Sci 51(3):1143–1150
Prasad AS (1963) Zinc and iron deficiencies in male subjects with dwarfism and hypogonadism but without ancylostomiasia, schistosomiasis or severe anemia. Am J Clin Nutr 12:437–444
Prasad AS (1991) Discovery of human zinc deficiency and studies in an experimental human model. Am J Clin Nutr 53:403–412
Prasad AS (2002) Zinc deficiency in patients with sickle cell disease. Am J Clin Nutr 75:181–182
Prasad AS (2003) Zinc deficiency. BMJ 326:409–410
Prentice AM et al (2008) New challenges in studying nutrition disease interactions in the developing world. J Clin Invest 118:1322–1329
Raiten J et al (2011) Executive summary—biomarkers of nutrition for development: building a consensus. The Am Clin Nutr 94(2): 633S–650S
Ramakrishnan U, Cossio GT, Neufeld ML, Rivera J, Martorell R (2004) Multimicronutrient interventions but not vitamin A or iron interventions alone can improve child growth: results of 3 meta analyses. J Nutr 134:2592–2602
Rayman M (2000) The importance of selenium for human health. Lancet 356:233–241
Reuter DJ, Robinson JB (1997) Plant analysis: an interpretation manual 2nd edn CSIRO Publishing, Melbourne
Rickes EL, Brink NG, Koniuszy FR, Wood TR, Folkers K (1948) Crystalline vitamin B12. Sci Total Environ 107:396
Rimbach G, Pallauf J, Moehring J, Kraemer K, Minihane MA (2008) Effect of dietary phytate and microbial phytase on mineral and trace element bioavailability, a literature review. Curr Top Nutraceutical Res 6(3):131–144
Robertfoid M (2005) Introducing inulin-type fructans. Br J Nutr 9: S13–S25
Robinson N, Procter CM, Connolly EL, Guerinot ML (1999) A ferricchelate reductase for iron uptake from soils. Nature 397:694–697
Roy C (2010) Anemia of inflammation. Hematology 30:276–280
Sager M (2006) Selenium in agriculture, food and nutrition. Pure Appl Chem 78:111–133
Schwarz K, Foltz CM (1957) Selenium as an integral part of factor-3 against dietary necrotic liver degeneration. J Am Chem Soc 79: 3292–3293
Seatharam B, Alpers DH (1982) Absorption and transport of cobalamin (vitamin B12). Ann Rev Nutr 2:343–349
Shoham S, Youdim MB (2002) The effect of iron deficiency and iron and zinc supplementation on rat hippocampus ferritin. J Neural Transm 109:1241–1256
Singh M (2004) Role of micronutrients for physical growth and mental development. Indian J Pediatr 71(1):59–62
Sillanpaa (1990) Micronutient assessment at country level: an international study. FAO Soils Bulletin No.63 FAO, Rome

Sohail S, Roland AD (1999) Fabulous phytase: phytase enzyme proving helpful to poultry producers and environment. Highlight Agric Res 1(46)
Steele M, Frazer DM, Anderson AJ (2005) Systematic regulation of intestinal iron absorption. Life 57:499–503
Stewart CP et al (2010) The undernutrition epidemic: an urgent health priority. Lancet 375:282
Subramanian KB, Bharathi C, Jegan A (2008) Response of maize to mycorrhizal colonization at varying levels of zinc and phosphorus. Biol Fertil Soils 45(2):133–144
Trowel L (1973) Definition of dietary fiber and hypotheses that it is a protective factor in certain diseases. Am J Clin Nutr 26:417–427
Turnlund J (1998) Human whole-body copper metabolism. Am J Clin Nutr 67:960S–964S
Turnlund JR, Keyes WR, Peiffer GL, Scott KC (1998) Copper absorption, excretion and retention by young men consuming low dietary copper determined by using the stable isotope 65Cu. Am J Clin Nutr 67:1219–1225
United Nations (1998) WHO, UNICEF, UNU (1998) IDA: Prevention, assessment and control. Report of a joint WHO/UNICEF/UNU consultation, Geneva. World Health Organization, Geneva
Uriu-Adams JY, Scherr RE, Lanoue L, Keen CL (2010) Influence of copper on early development: prenatal and postnatal considerations. Biofactors 36(2):136–152
Vallee B, Falchuk KH (1993) The biochemical basis of zinc physiology. Physiol Rev 73:79–118
Van Wersch JW, Janssens Y, Zandvoort JA (2002) Folic acid, vitamin B(12), and homocysteine in smoking and non-smoking pregnant women. Eur J Obstet Gynecol Reprod Biol 103:18–21
Walsh T et al (1994) Zinc: health effects and research priorities for the 1990's. Environ Health Perspect 102:5–46
Wang K, Zhou B, Kuo YM, Zemansky J, Gitschier J (2002) A novel member of a zinc transporter family is defective in acrodermatitis enteropathica. Am J Hum Genet 71:66–73
Wasantwisut E, Neufeld (2012) Use of nutritional biomarkers in program evaluation in the context of developing countries. J Nutr 142(1):186S–90S
Welch RM (1986) Effects of nutrient deficiencies on seed production and quality. Adv Plant Nutr 2:205–247
Welch RM (1996) The optimal breeding strategy is to increase the density of promoter compounds and micronutrient minerals in seeds: caution should be used in reducing anti-nutrients in staple food crops. Micronutr Agric 1:20–22
Welch RM (2008) Linkages between trace elements in food crops and human health. Micronutr Defic Glob Crop Prod 287–309
Welch RM, Graham RD (1999) A new paradigm for world agriculture: meeting human needs: productive, sustainable, nutritious. Field Crop Res 60(1):1–10
Wissuwa M, Ismail AM, Graham RD (2008) Rice grain zinc concentrationsas affected by genotype, native soil-zinc availability, and zincfertilization. Plant Soil 306:37–48
World Health Organization (1995) http://www.who.int/nut/malnutrition_worldwide.htm Malnutrition Worldwide. World Health Organization, Geneva
The World Health Report 2002: reducing risks, promoting healthy life. Geneva
WHO (2009) Trace elements in human nutrition and health. World Health Organization, Geneva
WHO (2011) World Health Statistics. World Health Organization, Geneva
WHO (2012) The World Health Report 2012, World Health Organization, Geneva
Zimmermann B (2011) The role of iodine in human growth and development. Semin Cells Dev Biol 22:645–652

Part III

Environmental Toxicology, Pathology, and Medical Geology

José A. Centeno

Joint Pathology Center, Biophysical Toxicology,
Malcolm Grow Medical Clinic, Joint Base Andrews Air Naval Facility,
Washington, DC, USA
e-mail: jose.a.centeno@us.army.mil

Introduction

In recent decades, there has been an increasing awareness of the importance of the interaction of mammalian systems with their natural environment. The primary focus has been on understanding exposure to hazardous agents in the natural environment through air, water, and soil. Such appreciation has led to a myriad of investigations focused on identifying those natural (and sometimes anthropogenic) environmental risk factors that may be involved in the development of diseases in humans and other animals.

Environmental medicine may be defined as the study of how environmental risk factors affect human health, which includes the practice of how to minimize and/or prevent any adverse effects. Humans are continually exposed to hazardous agents in the natural environment through air, water, soil, rocks, and even the workplace. The aim of environmental medicine is to better characterize exposure to a particular risk factor, to identify the type of adverse effects in tissues, and to determine the relationship between environmental exposure and genetic susceptibility for disease. It has been long thought that the genetic makeup of an individual plays a key role in the cellular response to environmental toxic agents such as chromium, arsenic, and nickel. Understanding the effects on humans from exposure to many of these elements present in the natural environment is a complex undertaking and one that requires an understanding of pathologic, toxicologic, speciation analysis, and epidemiological techniques. In this section, these scientific fields and techniques are summarized, with a focus on their application to the complexities of disease processes.

As indicated by Chap. 22, toxicology and epidemiology have significant roles in the study and implementation of environmental medicine. Environmental epidemiology is the study of associations between environmental exposures and the occurrence of disease within a population. The chapter presents common problems in environmental epidemiology, with exposure scenarios and case studies illustrating the application of epidemiology in environmental medicine. The objective of this chapter is to provide nonepidemiologists with basic skills to critically read and understand most epidemiological studies and recognize strengths, weaknesses, and biases related to design and exposure assessment. The chapter uses three recently published studies and illustrates different study designs, different outcome measures, and different types of exposure. At the end of the chapter, the authors provide a point-to-point checklist to guide the reader through reading and understanding an epidemiological paper.

In Chap. 24, the authors provide the reader with a comprehensive description of the principles and practices of environmental medicine focusing specifically on those aspects of the environment directly related to geological materials and geological processes. This chapter

describes in detail several fundamental concepts in environmental medicine, including the characterization of the mechanism(s) of exposure and the study of the internal (cellular) response to hazardous substances. The importance of understanding individual variability in response to exposure to a particular environmental agent is emphasized and illustrated using clinical and epidemiological studies. This chapter describes a number of approaches to the study and implementation of environmental medicine practices, including toxicology, surveillance, and intervention. Finally, the case of chronic arsenic exposure from contaminated drinking water is illustrated as an example of environmental medicine and medical geology.

Several diseases are related to the accumulation of trace elements, metals, and metalloids within different organ tissues, either as the primary or as a secondary manifestation of diseases. The accumulation of metals is associated with varying degrees of organ injury. In the liver, for example, these disorders include Wilson's disease (hepatolenticular degeneration), primary biliary cirrhosis associated with copper accumulation, and hemochromatosis associated with excess iron accumulation. Chapter 25 provides a comprehensive overview of environmental pathology and exposure to toxic metals. The chapter is concentrated on major organ systems of the body and discusses pathologic states induced by various elements in the skin, brain, lungs, heart, and liver. This is by no means all inclusive but rather serves as an introduction to our current understanding of pathologic states in these organ systems. Most studies have concentrated on chronic high-level exposures, and much less work has been done on low-dose exposures to toxic elements. Much remains to be learned, and in the chapters that follow, technical analyses are discussed that assist in broadening our understanding.

In Chap. 26, toxicologic testing and the geosciences are reviewed. Toxicology is critical in medical geology, and toxicologic principles are applied through clinical toxicology, risk assessment and hazard control, monitoring, and surveillance. A general framework of these areas is provided by an examination of toxico kinetics and toxicodynamics.

Understanding the speciation of trace elements in environmental medicine, nutrition, and diseases may play a significant role in promoting legislative and regulatory actions in human health protection pursuant to the decrease of chronic diseases and/or amelioration of debilitating effects. Chapter 27 describes in detail the emerging scientific field of trace element speciation, including a discussion on a wide range of techniques for speciation analysis that can be used to establish the chemical form (oxidation state), morphology, and compounds formed by trace elements, metals, and metalloids.

Environmental Epidemiology

23

Jesper B. Nielsen and Tina Kold Jensen

Contents

23.1 Introduction

Epidemiology is the study of the occurrence of disease in populations. It originates from an observational discipline that describes changes in the prevalence or incidence of a specific disease – changes that may be observed over time, between geographical regions, or between populations. Thus, basic epidemiology delivers numbers with no explanation. Prevalences or incidences are, however, only really useful if associated with explanatory variables. These variables may relate to genetics, lifestyle, age, gender, occupation, environment, etc. Environmental epidemiology is therefore the study of associations between environmental exposures and the occurrence of disease within a population.

Few environmental diseases are pathognomonic in the sense that only one specified exposure may cause a certain disease. In most cases, several chemical exposures may cause the same disease, aggravate an existing disease, or in some situations even offer a certain degree of protection. Likewise, several sociodemographic factors and occupational exposures may affect exposure as well as disease. Proof of causation in epidemiological studies is therefore seldom, and associations between exposure and disease may often be biased.

This chapter does not replace epidemiological textbooks, but it is intended to introduce and discuss some more basic features related to study design and measures of exposure and outcome, as well as bias. The purpose is that non-epidemiologists should be able to critically read and understand most epidemiological studies, know strengths and weaknesses of different common study designs, and be able to recognize the more general types of bias occurring in health and exposure assessment.

The chapter will present some common problems related to environmental epidemiology and primarily use three exposure scenarios (Cases 23.1–23.3) based on recently published scientific articles to illustrate some of these problems. Most problems are general in nature and an inherent consequence of the chosen study design. The three

J.B. Nielsen (✉)
Institute of Public Health, University of Southern Denmark, Odense, Denmark
e-mail: jbnielsen@health.sdu.dk

T.K. Jensen
Research Unit of Environmental Medicine, Institute of Public Health, University of Southern Denmark, Odense, Denmark

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_23,

examples have been chosen to illustrate different study designs, different outcomes or health effects, and different types of exposures. The studies selected are a case-control study of residential radon exposure and lung cancer (Barros-Dios et al. 2002), a cohort study on malignant mesothelioma and environmental exposure to asbestos (Metintas et al. 2002), and an ecological study on adverse pregnancy outcomes and exposure to arsenic in drinking water (Yang et al. 2003). All three articles are available as full text articles free of charge on the Internet.

23.2 Study Design

Study designs can be broadly categorized according to whether they are describing distributions of a health outcome (descriptive studies) or elucidating its determinants (analytical studies). Descriptive studies describe general characteristics of the distribution of an outcome in relation to person, place, and time. Analytical studies are used to test specific hypotheses and infer that exposure precedes outcome. They can be categorized into case-control or cohort studies according to whether the study subjects are selected on the basis of outcome or exposure (Fig. 23.1). This section will briefly introduce the different study designs and discuss their strengths and weaknesses. For further reading and more specified details, the reader is referred to epidemiological textbooks (e.g., Rothman and Greenland 1998).

23.2.1 Descriptive Studies

Descriptive studies describe general characteristics of the distribution of an outcome in relation to person, place, and time. The identification of descriptive characteristics is an important first step in the search for determinants or risk factors for specific outcomes, and thereby for the formulation of hypotheses to be tested in analytical studies. Descriptive studies use information from diverse sources such as census data, disease registers, and vital and clinical records, as well as national figures on consumption of food, drinking water, etc. Because this information is already available, descriptive studies are generally far less expensive and time-consuming than analytical studies. Usually they preclude the ability to test epidemiological hypotheses. Descriptive studies can be categorized into ecological or cross-sectional studies.

23.2.1.1 Ecological Studies

In an ecological study data from entire populations are used to compare outcome frequencies between different groups during the same time period or in the same population at different points in time. It is not possible to link exposure information to the occurrence of outcome in a particular individual. Furthermore, the studies are unable to control for confounding. Therefore, they cannot be used to test hypotheses or infer causality. They are, however, quick and inexpensive and use already available information (Table 23.1).

Case 23.1 is an ecological study where birthweight distributions in two different regions of Taiwan with different levels of arsenic in drinking water are compared. No individual exposure information was collected and the place of birth determined the exposure status. The women may, however, have moved to that area just before the delivery and therefore not have been exposed to the drinking water of that region in pregnancy at all. Furthermore, no information about the actual intake of water in pregnancy was available and it was not known whether the women with high water intake were those who delivered prematurely. In addition, no information on how much water the women drank at home was available, and the study was unable to control for confounding as, for example, the women's smoking habits. The study is, however, good for generating hypotheses to be tested in analytical studies (see also Chaps. 12 and 24, this volume).

Case 23.1. Prevalence of adverse pregnancy outcome among 18,259 first-parity singleton live births in Taiwan was linked to place of birth. Two geographic regions with different median levels of arsenic in the drinking water were included. Children from the arsenic-endemic area had on average a 30-g lower birthweight (statistical significant) and the rate of preterm deliveries was increased by 9% (insignificant). No data on individual exposures were available and the exposure to arsenic in drinking water in the arsenic-endemic area varied between <0.15 ppb and 3.6 ppm (20,000-fold), whereas the exposure in the area of comparison was below 0.9 ppb. In the arsenic-endemic area, 83% of drinking water resources had arsenic concentrations above 0.9 ppb.

Design: Ecological study
Outcome: Preterm delivery and birthweight
Exposure: Arsenic in drinking water
Reference: Yang et al. (2003).

23.2.1.2 Cross-Sectional Studies

In cross-sectional studies the status of an individual with respect to the presence or absence of both exposure and outcome is assessed simultaneously (Fig. 23.1). Thus, a cross-sectional study provides information about the frequency and characteristics of an outcome by a "snapshot" of the population at a specific time. Such data are of great value to public health administrators when assessing the health status or health-care needs of a population. However, as exposure and outcome are assessed at the same point in

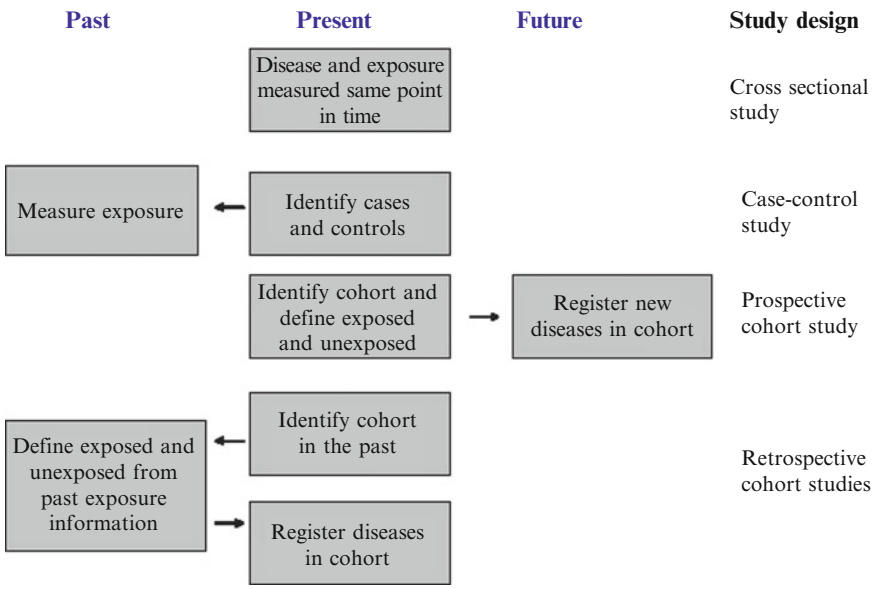


Fig. 23.1 A schematic view of different study designs

Table 23.1 Strengths and limitations in different types of epidemiological studies

Study type	Ecologicalstudy	Cross-sectional study	Case-control study	Cohort study
Strengths	Quick and inexpensive often using already available information	Quick and inexpensive	Optimal for evaluation of rare diseases	Valuable for rare exposures
		Provide information about health status of great public health relevance	Can examine multiple etiologic factors for a single disease	Can examine multiple effects of a single exposure
			Relatively quick and inexpensive compared to cohort studies	Can elucidate temporal relationship between exposure and disease
			Well suited for evaluation of diseases with long latency periods	Minimizes bias in exposure assessment
				Allows direct incidence rates to be calculated
Limitations	Unable to link exposure with disease in particular individuals	Cannot determine whether exposure preceded or resulted from the disease	Inefficient for evaluation of rare exposures	Inefficient for evaluation of rare diseases
	Unable to control for confounding	Considered prevalent and will reflect determinants of etiology as well as survival	Cannot compute incidence rates in exposed and unexposed individuals	Prospective: extremely expensive and time-consuming
	No individual exposure information		The temporal relationship between exposure and disease may be difficult to establish	Retrospective: requires the availability of adequate records
			Prone to bias, particularly recall and selection bias	Losses of followup can affect results

time, cross-sectional surveys cannot always distinguish whether the exposure preceded the outcome development or whether the presence of disease affected the individual's level of exposure. It is, in other words, not possible to determine whether the exposure preceded or was caused by the disease (Table 23.1). Thus, cross-sectional studies have found that infertile couples report more psychological distress symptoms, which implies that stress therefore causes infertility. It is, however, not known whether the couples became infertile because of the stress or whether the infertility and its consequences and treatment caused the stress. Cross-sectional studies are like ecological studies, which are valuable for raising a question of the presence of an association rather than for testing a hypothesis.

23.2.2 Analytical Studies

23.2.2.1 Case-Control Studies

In a case-control study, subjects are selected on the basis of whether they have (cases) or do not have (controls) a specific outcome. In its most basic form, cases with the outcome of

interest are selected from hospitals or the general population and compared with a group (controls) without the outcome. More refined study designs exist, and interested readers are referred to Rothman (Rothman and Greenland 1998, Rothman 2002). The proportions with the exposure of interest in each group are compared (Fig. 23.1). The case-control design is a good way to study diseases with long latency periods, because investigators can identify affected and unaffected individuals and assess antecedent exposures rather than waiting a number of years for the disease to develop. Therefore, case-control studies are time and cost efficient. In addition, by selecting the cases on the basis of outcome, the study can identify an adequate number of affected and unaffected individuals. Consequently, this strategy is particularly well suited for rare diseases, which in cohort studies would need inclusion of very large numbers of individuals in order to accumulate a sufficient number of cases with the outcome of interest. Finally, case-control studies allow evaluation of a range of potential etiologic exposures and their effect on the outcome. The case-control design can therefore be used to test specific a priori hypotheses or explore the effect of a range of different exposures.

The major drawback of case-control studies is that both the exposure and the outcome have already occurred at the time when the participants enter the study. This may affect the motivation to participate and the way that participants remember and report their exposures. This study design is therefore particularly vulnerable to selection and information bias, especially recall bias (see below). Furthermore, case-control studies are not efficient for rare exposures, as too few cases would then be exposed. In addition, only one outcome can be studied because the cases and controls are selected on the basis of that outcome. In case-control studies, no absolute measures of risk or incidence can be calculated. Instead the odds ratio estimates the relative risk or incidence rate ratio. This is, however, not a reason for not conducting case-control studies, as they offer advantages mentioned before and provide answers to hypotheses relatively fast (Table 23.1).

One of the first issues to be considered in the evaluation of a case-control study is the definition of disease or outcome of interest. It is important that the definition of disease (outcome) is as homogeneous as possible, because very similar manifestations of disease may have very different etiologies. For example, congenital malformations which encompass many different diseases such as congenital heart malformations, cleft-palate, or neural tube defects are often compiled into one outcome because of the rare nature of each of these disease categories. They do, however, have very different etiologies and combining them does not give clues to the risk factors of each particular outcome. It is therefore important to establish strict diagnostic criteria for the disease under study.

The selection of appropriate controls is perhaps the most difficult and critical issue in a case-control study. Controls are necessary to evaluate whether the exposure observed in the case group differs from what would have been expected in a comparable group of individuals without the disease. Controls must be selected, not to represent the entire non-diseased population, but the population of individuals who would have been identified and included as cases had they also developed the disease. They can be chosen from hospitals or the general population. Hospital controls are selected from people admitted to the same hospital as the cases but with a different disease. The advantage of this approach is that people admitted to hospitals are easy to identify, motivated, and more likely to be aware of antecedent exposures. The disadvantages are that there might be different selection factors leading to admission to that hospital for different diseases. Furthermore, they differ from healthy individuals and may therefore not represent the exposure distribution in the population from which the cases derived. Controls can also be chosen from the general population. This can be done in a number of ways including canvassing households in the targeted neighborhood, random digit telephone dialing, or identification from population registers or voting lists. This is, however, usually more costly and time-consuming. Furthermore, the quality of the information obtained and the participation rate from cases and controls may differ as healthy individuals from the general population do not recall exposures with the same level of accuracy and they are less motivated to participate.

It is often argued that cases should be representative of all persons with the disease. This is, however, not true and case-control studies can be restricted to a particular type of case from whom complete and reliable information on exposure and disease can be obtained (for example, in a limited age range). Then control subjects should be selected to be comparable to the cases. Such case-control studies will provide a valid estimate of the association between exposure and disease and a judgment of the generalizability of the findings can then be safely made.

Case 23.2. A total of 163 cases of primary lung cancer (response rate 70%) and 241 cancer-free controls (response rate 62%) were included. Cases were on average 8 years older than controls and had a 40% higher rate of cancer within the family. Close to 92% of cases were smokers as compared to 55% of the controls. Residential radon was measured in 98% of the homes for an average of 150 days. Residential radon exposure was close to 20% higher among cases than controls. This study concludes that residential radon exposure at levels below official guidelines of 148–200 Bq/m^3 may lead to a 2.5-fold increase in lung cancer risk. Further, synergism (an effect greater than that

expected by their separate actions) between residential radon exposure and smoking was demonstrated.

Design: Population-based case-control study

Outcome: Confirmed primary lung cancer

Exposure: Indoor radon concentration

Reference: Barros-Dios et al. (2002)

Case 23.2 is an example of a population-based, case-control study where lung cancer patients were compared with healthy controls from the same area in Spain. Controls were proportionally stratified randomly but excluded if they had respiratory tract disease, lived in the area less than 5 years, or were younger than 35 years of age. Exposure information was obtained from next of kin if the case or control had died. Information about radon exposure was measured, so no recall bias was present. The length that the participants lived at their current address was, however, not taken into account. In addition the participation rates were 10% higher among cases than controls and cases were approximately 8 years older, which may have introduced selection bias (see below). Moreover, more than 90% of the cases were smokers as compared to 55% of controls.

23.2.2.2 Cohort Studies

In a cohort study, a group of individuals are defined on the basis of presence or absence of exposure. At the time of exposure classification, subjects must be free from the outcome under investigation. Participants are then followed over a period of time to assess the occurrence of the specified outcome among those who are exposed and unexposed (Fig. 23.1). Most often, the followup period must be at least several years to allow an adequate number to develop the outcome so that meaningful comparisons of disease frequency between exposed and unexposed individuals can be made. As the participants by inclusion criteria are free of disease at the time when their exposure status is defined and the study initiated, the temporal sequence between exposure and disease can be more clearly defined.

For many exposures the proportion of exposed individuals with the outcome is too small to make meaningful comparisons between exposed and unexposed. Therefore, cohort studies are particularly well suited for assessing the effect of rare exposures. Thus, cohort studies can enroll participants on the basis of their exposures and thereby include a large number of exposed, for example, among a cohort of heavily exposed workers. Furthermore, cohort studies offer less potential for selection bias and direct measurement of association (incidence rates) can be calculated among the exposed and unexposed. Finally, cohort studies allow the examination of multiple outcomes of a single exposure.

As cohorts studies often involve a large number of individuals followed for many years, they are time-consuming and expensive. Furthermore, only a proportion of those eligible actually participate in the study, and they often differ from the non-participants in motivation and attitudes toward health. As outcome is compared among exposed and unexposed, this does not usually affect the relationship except when non-response is related to both exposure and outcome. A way to address this problem is by comparing participants with non-participants with respect to basic available information such as age and socioeconomic status. In addition, losses to followup may seriously affect the results, especially if it differs between exposed and unexposed individuals or is related to exposure or outcome or both. Losses to followup should therefore be minimized and for those lost to followup, attempts to gain information about outcome from independent sources should be made (for example, through death or disease registers).

Cohort studies are often categorized into prospective or retrospective studies according to whether the outcome of interest has occurred at the time the study is initiated (Fig. 23.1). In a prospective cohort study, the cohort is identified and categorized according to exposure. After a followup period the frequency of the outcome among exposed and unexposed is compared. In a retrospective cohort study both exposure and outcome have occurred at the time of the start of the investigation. A historical cohort is identified at the start of the study and past exposures in the cohort are identified from already existing information. Then the frequency of outcomes (which has occurred) is determined. Prospective or retrospective solely refers to whether the outcome has occurred at the start of the study. Case-control studies can also be both prospective and retrospective, but they are most often retrospective, i.e., the outcome defining the case has occurred when the study is initiated. Retrospective cohort studies can usually be conducted more quickly and cheaply than their prospective counterparts, because all relevant events have already occurred at the time the study is initiated. They do, however, depend on routine availability of relevant exposure data in adequate details from pre-existing records. Because these data were collected for other purposes, the quality is often not optimal. Moreover, information on potential confounding factors is often unavailable (see also Chap. 19, this volume).

Case 23.3. In a cohort of 1886 villagers in a rural area in Turkey, the incidence of malignant pleural mesothelioma (MPM) was studied. The villagers were environmentally exposed to asbestos dust due to the use of asbestos-contaminated white soil. The soil was used as a whitewash or plaster material for walls, as insulation, and also in pottery. Exposure was assessed on a subgroup level through measurement of airborne fiber concentrations both indoors and outdoors. During a 10-year observation period, 24 cases of MPM were diagnosed within the cohort corresponding to

an annual incidence rate close to 130/100,000. This incidence rate exceeds the expected in the general Turkish population by more than 100-fold, and is comparable to risks of MPM observed in occupational settings with much higher exposures.

Design: Cohort-study

Outcome: Malignant pleural mesothelioma

Exposure: Inhalation of dust from asbestos in soil

Reference: Metintas et al. (2002).

Case 23.3 is an example of a retrospective cohort study examining the incidence of malignant mesothelioma among people in villages exposed to white soil containing asbestos compared to the incidence among the background population in Turkey. It is an example on how rare exposures can be studied if highly exposed cohorts are chosen. The authors sampled 11 out of 403 villages, and it is difficult to rule out if a selection bias is present. Exposure levels were measured in white soil and information on potential confounding factors was obtained by interview with relatives. As it is often the problem with retrospective studies, the information about confounders is limited and no information on smoking habits was obtained.

23.3 Exposure Assessment

Outcome and exposure assessment are equal partners in a well-balanced epidemiological study. Accordingly, the very same questions on validity, bias, or confounding should be considered. Further, the representability of the exposure assessment with respect to individual, time, and place should be scrutinized. Thus, exposure assessments include qualitative as well as quantitative questions. This section will illustrate some common problems related to exposure assessment, but for a more in-depth discussion of this theme, readers are referred to more specific literature (a starting point is included in the Further Reading list).

23.3.1 What?

The qualitative questions relate to the validity of the analytical methods: (1) what is measured; (2) what is the specificity of the method, and (if the exposure is a mixture), (3) does the mixture change qualitatively over time or between areas included in the study. In relation to exposure to metals, the analysis of the total concentration of metal is often insufficient as different metal species will have specific toxicological profiles. Thus, for the assessment of intake of mercury or lead from soil by children, it would be relevant to know the species and salts occurring, as the intestinal absorption as well as toxicity of these metals depends on these features.

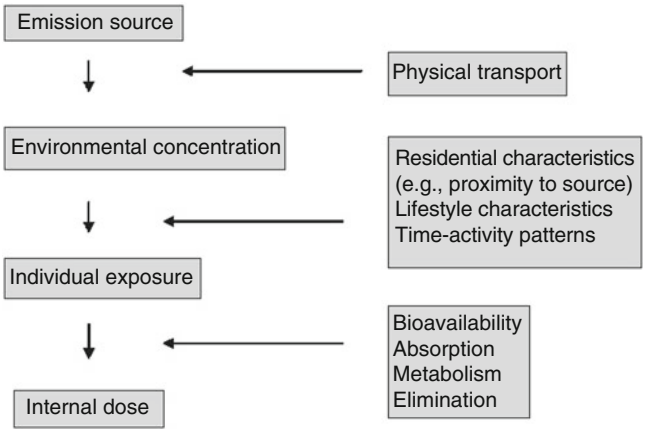


Fig. 23.2 Sources of variability in exposure assessment

Likewise, the authors in Case 23.3 use a well-validated method for fiber collection and only analyze and report the fraction of fibers (>5 um) relevant for the outcome (mesothelioma). This is a relevant approach, but it is not clear from the description in the article what proportion of the larger fibers analyzed were asbestos fibers. Neither is the exact type of asbestos described, which could be expected to influence the risk for the exposed individuals. These queries are especially important if the exposure is expected to vary qualitatively over time or between geographic regions included in a study.

23.3.2 Who?

Exposure may be assessed through measurements at different levels, beginning with very unspecific measurements at the emission source and ending with specific measurements of internal dose (Fig. 23.2). Ecological studies (like Case 23.1) will use exposure assessments at a very crude level, which causes severe uncertainty with respect to the individual exposure. Thus, in Case 23.1, 83% of the drinking water resources in the arsenic-endemic region had arsenic concentrations above 0.9 ppb. It is, however, not known whether the women with adverse pregnancy outcomes actually consumed water from these wells or from the 17% with uncontaminated drinking water nor how much water they drank. Advancing the exposure assessment will clarify different factors of variability and increase the knowledge of individual exposure (Fig. 23.2). An important point is, however, that the interindividual differences in exposure will not necessarily decrease with more specific exposure data. Instead, the uncertainty will be replaced by variability that may potentially be used in modeling or otherwise taken into account. In cohort or case-control studies, the possibility of getting more specific information on exposure is better as illustrated in Cases 23.2 and 23.3, where the exposure was

assessed in the homes of the participants. However, the inherent problem that the exposure assessment will also have to be done retrospectively needs to be addressed. Prospective studies will have the advantage that exposure assessment can be planned in advance. Individual exposure and information on internal dose requires personal monitoring equipment or the use of biomarkers of exposure. These approaches are, however, often time-consuming and expensive, and they may not be an option in larger cohort studies.

23.3.3 How Much?

Using exposure assessments without individual exposure data raises some general questions relating to the representativity of the concentrations measured in the environment. At their best, these concentrations will represent an average exposure, but may over- as well as underestimate individual exposures. In Case 23.3, the concentration of airborne fibers was measured in a few homes in each village, but it is not known whether these exposure levels represent the exposure in the remaining houses. In Case 23.2, radon was measured in 98% of the homes, but not related to the time spent indoors. Besides averaging exposures between persons and locations, a measurement of a contaminant in the environment will be a snapshot in time. A single measurement or even a few will not be very informative regarding variation in exposure levels over time. The ideal exposure assessment in an epidemiological study is seldom achieved, but the key information that allows for a useful exposure assessment is quantitative information on changes in exposure over time and between locations. This will help exchange uncertainty with variability.

23.3.4 When?

Knowledge of variations in exposure over time is critical for several outcome measures. This information is needed if the health outcome depends on specific windows of susceptibility within the exposed population. Thus, for adverse pregnancy outcomes, neither information on exposures dating 20 years back nor information on exposure after birth is relevant. Likewise, when the outcome appears after a delay in time, e.g., cancer, current exposure is not relevant. Thus, in Case 23.3, the most relevant information on exposure to asbestos fibers probably dates 10–20 years back.

Information on present exposure levels is mainly useful if the outcome appears without delay, in prospective studies, or when the exposure can be assumed to be unchanged over time. However, most exposures change with time. If risk is a function of time of exposure, exposure profiles including information on variation will be valuable (Fig. 23.3). In these profiles, the exposure concentration or dose is plotted as a function of time. Concentration versus time is used to describe the exposure, while amount versus time characterizes dose. If the elimination rate of the chemical is known, dose characterization may be used to estimate accumulation of contaminants. Further, exposure profiles may be used to identify more limited time periods with higher than average exposures, which may be relevant for some outcomes. Thus, a single short-term exposure to very high concentrations may induce adverse effects, even if the average exposure is much lower than an apparent no-effect-level. Such short-term peak exposures clearly remain unidentified if only average values are available. In other exposure scenarios, however, average values are sufficient to perform a valid exposure assessment. Therefore, an epidemiological study protocol should include careful considerations concerning the most relevant collection of data on exposure.

Exposure/dose profiles are important

- When risk is a function of time of exposure
  - Reprotoxicity
  - Neurotoxicity
- When risk may occur following short peak exposures
- When a chemical is accumulated
  - Lead, mercuty, etc.
- When outcomes appear after delay
  - Cancer

Fig. 23.3 Importance of exposure/dose profiles

23.3.5 Modeling Exposure

Several models, e.g., Monte Carlo simulations, have been developed through recent decades to estimate risk and exposure, to assess changes in exposure over time, or to identify worst case scenarios. These models are often very useful as they are able to accommodate and use vast amounts of information on parameters of importance for modeling individual exposures. If physiological and behavioral parameters are included, these models may even estimate target organ deposition. One of the major achievements of these models is that they have enabled risk and exposure assessors to replace the often very conservative estimates of worst case scenarios with more realistic scenarios based on probability functions. It is, however, important to remember, that a model never gets more valid than the validity of the exposure information obtained and entered into the model.

23.4 Bias

Two types of errors may occur in epidemiological studies: random and systematic errors. Random errors are, as the word implies, random and are minimized when the study size or the precision of information is increased whereas

systematic errors are unaffected by the size of the study. If a participant is weighed on an imprecise weighing scale, his or her weight may be over- or underestimated, but the error is random. Therefore, an increase in number of participants will reduce bias. If, however, the weighing scale is systematically overestimating the true weight of the participants, there is no effect of increasing the number of participants. It will still overestimate their weight, and a systematic error is introduced.

Systematic errors are often referred to as bias. Bias may be defined as any systematic error in an epidemiological study that causes an incorrect estimate of the association between exposure and outcome. Because epidemiological studies involve humans, even the most perfectly designed study will have the potential for one or more types of errors. Consequently, evaluating the role of bias as an alternative explanation for an observed association is a necessary step in the interpretation of any study. Therefore it is essential to discuss types of biases that might be present as well as the most likely direction and magnitude of their impact. A study can be biased because of the way in which the study subjects are selected (selection bias), the way the study variables are measured (information bias), or by the lack of measurement of other exposures related to the outcome (confounding).

23.4.1 Selection Bias

Selection bias is a systematic error in the study that occurs when the association between exposure and outcome differs for those who participate and those who do not participate in the study. The participation rate in a study is never 100%, therefore it is important to gain information about age and sociodemographic status from non-participants and compare these with the participants.

Selection bias may change the estimates both toward and away from the null hypothesis. Selection bias is of particular importance if the participation rate is low, varies between cases and controls or between exposed and unexposed. In Case 23.2, 70% of the lung cancer patients and 61% of the controls participated. Patients may be more interested in participating as they believe that the exposures studied may have caused their disease.

A special form of selection bias occurs when the prevalence of an outcome for a group of workers is compared with the prevalence for the general population. This comparison is biased because the general population includes many people who cannot work because they are too ill. Consequently, the outcome is more frequent in the general population. This bias is often referred to as healthy worker effect.

23.4.2 Information Bias

Information bias is caused by systematic differences in the way data on exposure or outcome are obtained from the various study groups. The participants are thereby misclassified with respect to either exposure or disease. This misclassification can be either differential or non-differential. Consider the smoking information in Case 23.2. If both cases and controls underestimate the number of cigarettes that they smoke on average, this would lead to a categorization of heavy smokers as light smokers. The classification of exposure (smoking) is unrelated to the outcome (lung cancer) as both cases and controls underreport to the same extent. The information bias is therefore a non-differential misclassification. A non-differential misclassification will produce estimates of the effect that are diluted and will tend to support the null hypothesis. Now imagine that lung cancer patients underreport their smoking to a greater extent than controls. The classification of exposure (smoking) is then related to the outcome (lung cancer), and the bias is differential and may under- as well as overestimate the effect.

A common type of information bias is recall bias, which may occur in case-control studies where a subject is interviewed to obtain exposure information after the outcome has occurred. Cases then tend to have a different recall than controls as a result of their disease. In Case 23.2 lung cancer patients and controls were interviewed. If lung cancer cases remember and report their smoking differently than controls, this causes a recall bias. Case 23.2 was further complicated by the fact that some cases had died by the time of the investigation and next of kin were interviewed instead. They may not remember exposures as precisely as the cases themselves.

23.5 Confounding

Confounding is a mixing of effects. A confounder is an exposure other than the one investigated, which is associated with outcome, but unequally distributed between the groups compared. Furthermore, it must not be an intermediate step in the causal pathway from exposure to disease. Confounding can cause bias in either direction.

In Case 23.2, smoking is a confounder. It is associated with the outcome (lung cancer) and unequally distributed among cases and controls (92% of cases and 55% of controls smoked). Furthermore, it is not an intermediate step in the causal pathway between radon exposure and lung cancer. Therefore, if the study does not take smoking into account, the effect of radon exposure would be overestimated as the estimate would really measure the aggregated effect of smoking and radon on lung cancer.

Confounding may be controlled by restricting the study population, thus all participants are equal with respect to a potential confounder (e.g., restricting the study to a specific age category). In Case 23.2, the study could have included only non-smokers. Another way to deal with confounding is by matching the study subjects with respect to the confounder. In Case 23.2, a smoking control could be included every time a smoking case was included. Matching poses special challenges and is not discussed further in this chapter, but readers are referred to epidemiological textbooks (see Further Reading section). Confounding control may also be addressed during analysis of the data by multiple regression analysis or by stratifying data, i.e., study lung cancer and radon exposure among smokers and non-smokers separately.

23.6 Statistics

The majority of statistical analyses involves comparisons between groups of subjects. Initially, a hypothesis, called the null hypothesis, states that there is no difference in the outcome of interest between the groups of subjects. Statistical analysis is then an evaluation whether to accept or reject this hypothesis. The selected study subjects are only subsamples of the entire population, and probabilities are used to describe the certainty by which the null hypothesis is rejected. This probability is given as a p value in most statistics. Thus, a p value of 0.05 means that there is 95% certainty that the null hypothesis is not true and should be rejected. There is, however, a remaining probability of 5% that the rejected null hypothesis was actually true. This is known as a false positive result and termed a type I error. Thus, the risk of a type I error is determined by the size of the p value that is used as the level of rejection of the null hypothesis. False negative results may also occur, i.e., the acceptance of a null hypothesis that is not true, and this is called a type II error. In Case 23.1, the authors use a probability of $p < 0.001$ to conclude that the mean birthweight is different in the two regions. There is only a 0.1% risk of a type I error, e.g., that the difference observed is actually only a chance finding and not true. Type I and type II errors are interdependent. Thus, whenever the risk of a type I error is reduced, i.e., by decreasing the p value used as level of rejection of the null hypothesis, the risk of a type II error is increased and vice versa. For a thorough statistical explanation, readers are referred to statistical textbooks.

In many epidemiological studies large numbers of comparisons are made between different subgroups within the observed group of subjects. If a probability of $p = 0.05$ is used as the level for rejection of the null hypothesis, this means that for every 20 comparisons one will, just by chance, be a false positive finding. Therefore, most statistical packages include methods to reduce this risk of type I errors when doing large numbers of comparisons. Failing to apply these methods in multicomparison scenarios may invalidate conclusions.

23.7 Check List for Evaluating an Epidemiological Paper

Reading and understanding an epidemiological study report may be time-consuming and it is often difficult to evaluate its validity. We have tried to develop a checklist to guide the reader through the most pertinent questions relating to the validity, strengths, and weaknesses of an epidemiological paper. Answering the questions on the checklist will make the reader recognize possible problems in a paper.

1. What type of study design was used?
   See strength and limitations for the different design in Table 23.1.
   - Cross-sectional study – The problem with cross-sectional studies is that information about exposure and outcome is collected at the same point in time. It is therefore impossible to draw conclusions about causation.
   - Ecological study – No individual exposure information is collected in an ecological study. It is therefore impossible to infer causation in these studies.
   - Case-control study – Exposure among a group of cases selected on the basis of an outcome, i.e., disease, is compared to exposure among a control group without the outcome.

   Study Design:
   - How were the cases defined?
   - How were the controls selected? Community or hospital controls?
   - Were cases and controls from the same source population?
   - Were cases and controls matched?
   - What was the number of controls per case?

   Study population:
   - What was the target population?
   - What recruitment procedures were used?
   - Did the participation rate among cases and controls differ?

   Validity:
   - Over what time period was the study population recruited?
   - Were the cases and controls comparable with respect to characteristics, response rates and time of recruitment?
   - Was any information about non-responders obtained?

- Cohort study – A cohort is defined and categorized into exposed and unexposed, who are then followed in order to determine new occurrences of the outcome.

Study Design:

- Was the cohort defined retrospectively or prospectively?
- How large was the cohort?
- How many were exposed?
- How many observed events were there?

Study population:

- What was the target population?
- What recruitment procedures were used?
- How many were lost to followup (percentages)?

Validity:

- Over what time period was the study population recruited?
- Was there any description of the losses to followup?
- Do the losses to followup introduce bias?

2. What were the hypotheses of the study?
   - Was the study originally designed to test these specific hypotheses?
3. How was the data quality?
   - Were the data collected for the purpose of the study or were they obtained from other sources, for example, registers or hospital files?
4. How was exposure assessed?
   - Who provided the information about exposure (subject, family or others)?
   - Was the quality of exposure information assessed?
   - Were the subjects and/or interviewer blinded to the hypothesis?
   - How was the exposure information linked to the cases?
   - Is risk a function of time, and were exposure profiles included?
5. How was information about outcome measured?
   - Self-reported, by health personnel or from registers
6. Was adequate statistical analysis used?
   - Did the analyses control for potential confounders?
   - How wide were the confidence intervals?
   - Is a type 1 or type 2 error possible?
7. Bias

   Selection and information bias:
   - It should be detected from the questions asked under case-control and cohort studies.

   Confounders:
   - Was adequate information about the confounders obtained?
   - How accurate and adequate was the information about confounders?
   - Did the study control for confounding by restriction, matching, stratification, or multiple regressions?
8. Have other studies reported similar findings (consistency)?
   - Have a number of studies conducted by different investigators, in different geographical areas, and among different cultures at various points in time using different methodology found similar results?
   - Lack of consistency should lead to a high degree of caution at any causal interpretation of the findings.
9. What is the strength of the association?
   - The magnitude of the observed association is useful to judge the likelihood that the exposure itself affects the risk of developing the disease, and therefore, the likelihood of a cause-effect relationship. Specifically, the stronger the association – that is the greater the magnitude of the increased (or decreased) risk observed – the less likely that it is merely due to the effect of unexpected and uncontrolled confounding. This does not imply that a weak association cannot be causal, merely that it is more likely to exclude alternative explanations.
10. Is there a plausible biological mechanism of action and do experimental studies show similar results?
    - Because what is considered biologically plausible or tested in animal studies at a given time depends on the current knowledge, the lack of these criteria do not necessary mean that a relationship is not causal.
11. Did the exposure precede outcome?
    - Many lifestyle factors are likely to be altered as the first symptoms of a disease appear.

See also the Following Chapters. Chapter 3 (Natural Distribution and Abundance of Elements), Chapter 12 (Arsenic in Groundwater and the Environment), Chapter 19 (Natural Aerosolic Mineral Dusts and Human Health), Chapter 24 (Environmental Medicine), Chapter 25 (Environmental Pathology)

Further Reading

Basic Epidemiology

Rothman KJ (2002) Epidemiology. An introduction. Oxford University Press, New York

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Exposure Assessment

Paustenbach DJ (2000) The practice of exposure assessment: a state-of-the-art review. J Toxicol Environ Health B 3:179–291

References for Case-Studies

Barros-Dios JM, Barreiro MA, Ruano-Ravina A, Figueiras A (2002) Exposure to residential radon and lung cancer in Spain: a population-based case-control study. Am J Epidemiol 156: 548–555

Metintas S, Metintas M, Ucgun I, Oner U (2002) Malignant mesothelioma due to environmental exposure to asbestos – follow-up of a Turkish cohort living in a rural area. Chest 122:2224–2229

Yang C-Y, Chang C-C, Tsai S-S, Chuang H-Y, Ho C-K, Wu T-N (2003) Arsenic in drinking water and adverse pregnancy outcome in an Arseniasis-endemic area in Northeastern Taiwan. Environ Res 91:29–34

Environmental Medicine

24

Jefferson Fowles, Philip Weinstein, and Chin-Hsiao Tseng

Contents

24.1 Introduction: What Is Environmental Medicine?

A broad definition of environmental medicine is the study of how the environment affects health, including the characterizing of exposures, and the practice of how to minimize any adverse effects. The term environmental medicine is effectively synonymous with the term environmental health, except that the latter term is often confused with "health of the environment," so we will stick to the former term. The word environment also requires interpretation, because there are many subsets of environments that have special branches of environmental medicine associated with them. For example, occupational medicine (the study of the effects of the work environment on health) and social medicine (the study of the effects of social structures and dynamics on health). Of particular interest in this book are those aspects of the environment that relate directly to substances and processes of geological origin. This chapter will attempt to provide a brief outline of the principles and practice of environmental medicine by drawing largely on examples with geological relevance.

24.2 External Processes: How Are We Exposed to Hazardous Substances?

24.2.1 Environmental Media

A medium (pl. media) is a vehicle by means of which exposure occurs. Thus, when humans are exposed to arsenic (for example), it is often by means of their drinking water, that is, water is the environmental medium of exposure. Other media by which substances of geological origin may come into contact with humans include air (e.g., inhalation of radon gas or airborne particulates), soil, and food. Although there is some direct ingestion of soil in food preparation or in conditions such as pica, contaminants in

J. Fowles (✉)
Tox-Logic Consulting, LLC Petaluma, CA, USA
e-mail: jefffowles1@gmail.com

P. Weinstein
University of South Australia, Adelaide, Australia

C.-H. Tseng
Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_24,

Table 24.1 A simple framework for classifying health effects of the environment

Environmental medium	Health effect	Type of example from medical geology
Air	Physical	Death by asphyxiation: absence of oxygen in a mine shaft; chronic obstructive pulmonary by disease caused inhalation of particles
	Chemical	Neurotoxicity by inhalation: geothermal hydrogen sulfide accumulated in basement
	Biological	Pneumonia: infection by airborne anthrax from dried out soils
Water	Physical	Death by drowning: inundation by tsunami or other flooding
	Chemical	Skin cancer and vascular pathology: chronic ingestion of arsenic in drinking water
	Biological	Meningitis: infection by amoebae in hot springs
Soil/food	Physical	Direct injury: from displacement of natural or artificial structures in earthquake
	Chemical	Goiter: disruption of metabolism in areas with iodine deficient soils
	Biological	Tetanus: infection of cuts by contact with soil containing Clostridium spores

soil often enter humans by way of the food chain. Soil and food can therefore be conveniently thought of as one medium. For example, in regions where iodine is deficient in the soil, it is also deficient in the food chain, and goiter (or cretinism in children) may result as a deficiency disease in people dependent on local produce. The soil/food medium is complicated by the factor of bioavailability, which determines the ability of a substance to migrate, in its current chemical state, from the soil fraction to living organisms. (see also Chap. 27, this volume). These three environmental media: air, water, and soil/food form the basis of a useful analytic framework in environmental medicine, as illustrated in Table 24.1.

24.2.2 Hazardous Substances

To carry the analytic framework further, the nature of the injurious agent conveyed by the above media can be examined. It is in this examination that environmental medicine most clearly shows itself as a broad science; the injurious effects of landslides are just as relevant in this context as are the toxins traditionally studied by toxicologists. The latter affect health because of their chemical nature, and are arguably most readily associated with the field of medical geology. A broader perspective was illustrated in Table 24.1 with physical as well as microbiological health effects that result, however indirectly, from geological processes.

24.2.3 Mechanisms of Exposure

Given a hazardous substance that has made its way to a human through an environmental medium, it remains for that substance to exert its influence by entering into the biochemical and metabolic functioning of the body at the cellular level. First and most obviously cellular function may be interrupted at the physical level by direct cellular injury in a rockfall, or by asphyxiation as the respiratory systems fails following a tsunami (Table 24.1). Chemical and biological substances, on the other hand, enter the body through existing organ systems, by means of which they reach the cells and organ systems that they affect. Inhalation, ingestion, and absorption are such modes of penetration to the cellular level, with the systems responsible to include the respiratory, gastrointestinal, and integumentary (skin) systems respectively. Thus carbon monoxide from incomplete coal combustion is inhaled, passes through the lungs to red blood cells in the alveolar capillaries, and irreversibly binds hemoglobin so that cells are starved of oxygen. Lead from old paint is ingested with house dust, passes through the intestinal tract where it is taken up by the bloodstream, and transported to the brain where it exerts its neurotoxic effect on brain cells. Methylmercury accidentally spilled on the skin is absorbed directly through the skin and into the bloodstream, and is likewise transported to the brain where it exerts its neurotoxic effect. An understanding of these particular mechanisms of introducing hazardous substances to the body at the cellular level are of critical importance to the practice of environmental medicine: only with such an understanding can appropriate barriers be devised to protect individual as well as public health (see also Chap. 6, this volume).

24.2.4 Case Study: Soil, Water, and Amoebae

New Zealand is located in the Pacific Ring of Fire, on the boundary of two major tectonic plates. As these rub together, a number of geothermal phenomena are produced, including volcanism (see Chap. 10, *Volcanic Emissions and Health*, this volume) and the geothermal heating of freshwater springs. Such hot springs are an integral part of New Zealand's cultural history; since the first humans colonized the islands some 1,000 years ago, hot springs have been used for recreation, cooking, healing, and heating.

About 100 natural thermal springs are dotted around the country, and they are the delight of New Zealand children and tourists alike. The diving, jumping, and water sliding

unfortunately took a serious blow in the late 1960s when isolated deaths began to occur from primary amoebic meningoencephalitis (PAM).

The amoeba responsible for PAM, *Naegleria fowleri*, invades the nasal mucosa and olfactory nerve endings, and then tracks up to the meninges and brain through a sieve-like weakness in the skull known as the cribriform plate. The infection is usually accompanied by a fever, sore throat, and headache and progresses in a few days to nausea, vomiting, neck stiffness, and sometimes olfactory hallucinations. Death is usual by day 5 or 6, and almost inevitable by day 10. Active and healthy young people are often the victims, which fuels public outrage disproportional to the public health impact of this disease (there have been less than a dozen deaths from PAM in New Zealand, and less than 200 reported cases globally).

An epidemiological association (see Sect. 24.4) was soon established between PAM and exposure to hot springs, and *N. fowleri* was isolated from water samples from suspect springs. Once the microbial hazard, environmental medium, and mechanism of introduction had all been established, appropriate barriers could be devised to protect the public from further infections. There was a publicity campaign requiring signs at hot pools to warn the public to keep their heads above water. This safety message persists in the New Zealand psyche, and many home spa users adhere to it even if they can no longer remember exactly why.

Interestingly, there is a geological link in this example because *N. fowleri* was only found in soil-contaminated waters. Free living in soil, the organism thrives in the warm waters of hot springs (and therefore not surprisingly also does well at body temperature). It is introduced into the water if the spring runs over bare soil before entering the pool, if the pool has unsealed edges or bottom allowing direct soil contact or runoff with rain, or if the pool surrounds are of exposed soil that enters the pool on bathers' feet. A second public health intervention was therefore to isolate the hydrological environment from the geological environment: standards were introduced for piping water into pools; improving pool construction; and installing wide, paved pool surrounds.

There have now been no cases in New Zealand for over 20 years. This provides an excellent example of the effectiveness of the sound practice of environmental medicine.

24.3 Internal Processes: How Does the Body Respond to Hazardous Substances?

24.3.1 Concepts of Absorption, Distribution, Metabolism, and Repair

The basic processes of absorption, distribution, and metabolism are critical in gauging the body's internal exposure and response to any toxic chemical. In particular, these processes are crucial in understanding the significance of laboratory animal results in relation to human health risks, when modeling or estimating exposures in epidemiology studies, and when discussing the potential for increased susceptibility of subgroups within the population. Figure 24.1 depicts this relationship (see also Chaps. 7 and 9, this volume).

24.3.1.1 Absorption

Chemicals are absorbed through oral ingestion, inhalation, or through the skin at a rate relating to their water solubility, size, ionization state, acid dissociation constant (pKa), and exposure concentration. In general, ionically neutral compounds are absorbed more readily than ionized compounds. Acids tend to be better absorbed in the acidic stomach, while basic compounds may be better absorbed under the more alkaline conditions in the small intestine. When considering metals, active uptake mechanisms play an important role, such as the use of calcium transport mechanisms serving to increase the uptake of lead. Some characteristics of compounds and their influence on absorption are shown in Table 24.2.

It is thought that water-insoluble metal forms are not bioavailable for absorption by the stomach and small intestine. However, some insoluble forms become bioavailable in the acidic stomach. For example, the gastrointestinal effects for copper sulfate and copper oxide in humans are similar (Pizarro et al. 2001).

24.3.1.2 Distribution

The distribution of a chemical in the body is a function of its water solubility, ionization state, molecular size, and affinity

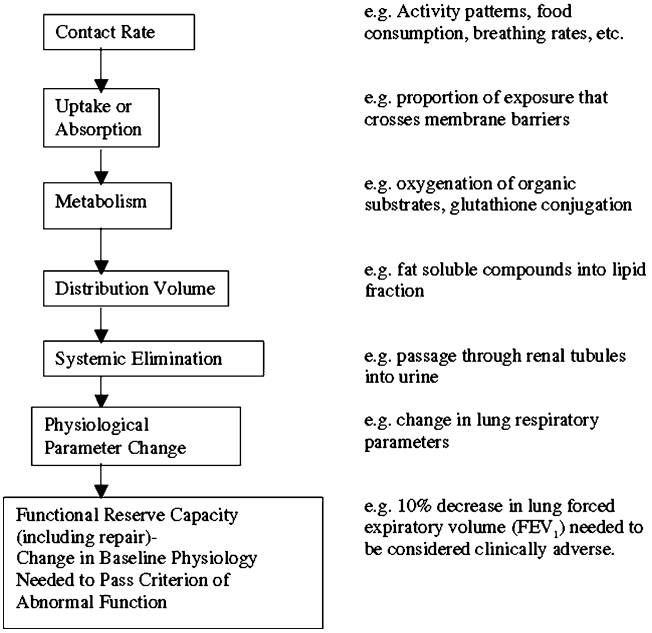


Fig. 24.1 The pathway from exposure to response

Table 24.2 Substance characteristics that influence absorption or bioavailability

Substance characteristic	Uptake mechanism	Example
pH of weak organic acids and bases	Ionically neutral states are more readily absorbed	Aniline absorption in the small intestine.
		Aspirin absorbed in the stomach
Valence states of metals	Specific active transport mechanisms	Chromium VI uptake into cells, while Chromium III is excluded
		Inorganic arsenic transported by phosphate transport mechanisms
Cationic metal form may have competitive antagonists	Specific active transport mechanisms	Lead competes with iron for uptake and reduces iron bioavailability; molybdenum competes with copper for absorption
		A child who gets enough iron and calcium will absorb less lead
Sorption status of metals in soils (bioavailability)	Ionic water soluble forms are more available	Cadmium is bound to organic counter ions in soil at high pH and therefore less bioavailable
Molecular weight	Smaller compounds tend to cross membranes more easily	Limited bioavailability of some large marine biotoxins (e.g., gymnodamine) through oral ingestion
Water solubility	Water-soluble liquids that are fat soluble cross membranes; water-soluble unionized compounds cross membranes more easily than ionized counterparts	

for specific receptor sites in tissues. Generally, the more water soluble the compound, the more easily it can be excreted, and most metabolic processes work toward this end. Extreme cases, such as the case with DDT, can be illustrative. DDT is extremely lipophilic and therefore is retained in adipose tissue throughout the body and the fat found in the bloodstream. In cases such as this, the distribution affects the compound's toxicity and biological half-life (in this case, approximately 8 years) through storage reservoirs that prevent access to metabolic systems.

The chlorinated dibenzo-*p*-dioxins are another example of lipophilic compounds distributing into adipose tissue, and because those isomers with chlorines at the 2,3,7,8-positions are highly resistant to metabolism, the result is a biological elimination half-life of 7–11 years (U.S. EPA 2000).

The tissues of highest concentration may or may not be the critical target organs. For example, copper is stored primarily in the liver, brain, heart, kidney, and muscles. About one-third of all the copper in the body is contained in the liver and brain, and the critical toxicological effect from chronic exposure is in the liver. Another one-third is contained in the muscles, where no toxic effect is known to occur. The remaining one-third is dispersed in other tissues.

Some compounds have deep storage depots, such as fluoride or lead in bone. When the chemical is in these depots, it is not bioavailable for activity at distal sites, and the elimination half-lives of such compounds are very long. The half-life for lead in bone is estimated to be about 25 years. Lead has a half-life of about 25 days in the blood. Methylmercury has a half-life for elimination of about 50 days in adult males and females. However, infants have poor elimination of organic and inorganic heavy metals in the first 6 months prior to development of metal transport systems, and it is possible that mercury is eliminated much more slowly in the child (Brown 2001).

24.3.1.3 Metabolism

Metabolism of a substance is subject to enormous interspecies and interindividual differences, through genetic polymorphisms, which in turn affects susceptibility to the chemical.

Metabolizing enzymes serve the functions of detoxification, intoxication, and facilitation of excretion of compounds. Key enzymes are the cytochrome P-450 family of heme-containing oxygenases, acetyltransfer-ases, glutathione-S-transferases, and glucuronidases. These enzymes are needed to increase the polarity and water solubility and thus the rate of excretion of organic environmental contaminants.

Through metabolic enzyme differences, chemicals that are benign to humans can be highly toxic to some species and vice versa. The toxicity of paracetamol (acetaminophen) to cats is a classic example of how species-specific metabolism can influence toxicity. Cats are particularly susceptible to paracetamol intoxication because of their impaired glucuronic acid conjugation mechanism and rapid saturation of their sulfate conjugation pathway, whereas humans rely heavily on a much more robust glucuronic acid conjugation system, which effectively detoxifies the critical metabolites.

For some metals, metabolism is also influential in determining bioavailability and toxicity. Arsenic methylation, the primary process by which the metal is metabolized in the body, has generally been considered a method of detoxification. Methylated arsenicals produced metabolically from inorganic arsenicals are excreted faster and have a lower affinity for tissue sulfhydryl groups than inorganic

arsenicals, especially arsenites (Chan and Huff 1997). It has also been shown that the incidence of skin cancer due to oral arsenic exposure is associated with individual methylating capacity (Hsueh et al. 1997). The epidemiological evidence indicates that the methylation detoxification mechanism does not become saturated even at high doses but that some inorganic arsenic always remains in human urine, regardless of the amount of arsenic exposure (Hopenhayn-Rich et al. 1993). Arsenic poisoning has been thought to occur only when the rate of exposure exceeds the rate of methylation (Le et al. 2000).

24.3.1.4 Repair

The body is constantly repairing itself from damage sustained from a myriad of environmental insults. Among the most significant of these mechanisms is with the repair of genetic lesions. Genetic lesions are central to the mechanisms of carcinogenesis and some developmental defects.

There are over 100 genes responsible for maintaining the integrity of our DNA. These include endonucleases, polymerases, and ligases. Each enzyme is important in one or more areas of DNA repair, which in turn affects individual susceptibility to particular agents. The clinical manifestations of defects in one or more of these enzymes can be seen in a number of genetic diseases. Xeroderma pigmentosa (XP) patients carry a 1,000-fold increase in skin cancer incidence, but no significant increase in internal cancers. UV-light-induced single DNA base point mutations are the most critical for these patients. XP patients have a defect in a key endonuclease required for carrying out excision-repair of point mutations (Hoffman 1994). Ataxia telangiectasia patients, on the other hand, can repair point mutations induced by UV light, but are very susceptible to x-rays because of a defect in repair of strand breaks. Similarly, Bloom's syndrome patients carry a defect in DNA ligase I. This defect effectively increases chromosome fragility and results in huge increases in cancer rates (e.g., 28 out of 103 patients died of cancers at a mean age of 20.7 years).

A number of metals are carcinogens by mechanisms that are not absolutely clear, but some metals do appear to inhibit one or more DNA repair enzymes. This could be a mechanism of indirect action for several metal carcinogens.

24.3.2 Dose–Response Relationships

A critical aspect of toxicology is the description of a dose–response relationship. Conceptually, a dose–response relationship requires either the severity of a particular end point, or the incidence of the adverse effect in the population, to increase with increasing dose. The increase does not need to be monotonic, but there should be a region of the dose range that is linear. Once much higher doses are reached, additional toxic effects of a different nature may manifest and cloud the estimation of the response at lower doses (see also Chap. 26, this volume).

For most acute toxicity end points, such as acute lethality, the dose–response relationship is described in a relatively straightforward manner, and linear or log-linear models describe these relationships quite adequately (Fowles et al. 1999).

For chronic toxicity end points, such as cancer, the behavior of the dose–response curve at the low dose end is critical to interpretation and assessment of health risks. Some compounds have associated biological mechanisms that support a "hockey-stick" shaped dose–response curve for cancer. This is particularly relevant to the case of overloading of functional reserve or repair capacity of the body at high doses.

The dose–response relationship is important for the following reasons:

- Validation of hypothetical causal relationships to chemical exposure
- Provision of a measure of toxicological potency which allows prioritization of hazards and risks by risk managers
- Description of the range of variability in responses in the test population

24.3.3 Varieties of Effects: "Non-Cancer" Toxicity and Carcinogenicity

Regulatory toxicology and toxicological risk assessment generally divides responses into those of a carcinogenic and non-carcinogenic nature.

24.3.3.1 Carcinogenicity

It has been estimated that over 80% of all cancers, population wide, are environmentally induced (Doll and Peto 1981). This estimate includes cancers from smoking, dietary carcinogens, and exposure to air and water pollution in addition to those from cosmic and solar radiation (see also Chap. 25, this volume).

One example of a type of cancer thought to have strong environmental links is prostate carcinoma (PC). In the United States, 10% of all 75-year-old men (black, white, and Japanese ethnicity) have latent PC. However, the active form of this carcinoma is 60:30:1 in these populations. This suggests that there are environmental, endocrine, or dietary factors that influence the progression of latent PC to active cancer.

Adult tissues, even those that are composed of rapidly replicating cells, maintain a constant size and cell number by regulating the rate of replication, by differentiation to

assume specialized functions, and by programmed cell death (apoptosis). Cancers are diseases in which there are somatic mutations of genes critical to maintenance of control over cell division that lead to loss of control over cell replication, differentiation, and death. The International Agency for Research on Cancer (IARC) has defined chemical carcinogenesis as: "the induction by chemicals of neoplasms that are not usually observed, the earlier induction by chemicals of neoplasms that are commonly observed, and/or the induction by chemicals of more neoplasms than are usually found."

Carcinogenicity is thought to occur in a multistep or "multistage" process, with several key events occurring in sequence for a given normal cell to convert into a malignant cell with unregulated growth. The number of genes altered in a cancer cell compared to a normal cell is not known; recent evidence suggests that 3–10 genetic events are involved in common adult malignancies in humans. Two distinct classes of genes, proto-oncogenes and tumor-suppressor genes, are involved in the cancer process (Barrett 1993).

In the current multistage model of carcinogenesis, development of a malignant tumor occurs in three stages: initiation, promotion, and progression. Initiation involves an irreversible change in a normal cell (usually an alteration of the genome) allowing for unrestricted growth. The initiated cell may remain latent for months or years. During this period of latency, the initiated cell is phenotypically indistinguishable from surrounding cells of the same type. Further development of the initiated cell into a neoplastic cell requires a period of promotion. Under the influence of a promoter, tumor formation is accelerated through clonal expansion of initiated cells. Promoters, which do not interact directly with DNA, are a diverse group of agents believed to act by a variety of mechanisms most often resulting in increased cell proliferation. The process of promotion is considered reversible and requires prolonged and repeated exposure to promoter agents. Progression is the final step in which pre-neoplastic foci develop into malignant cells. In this stage, tumor development is characterized by karyotypic changes, increased growth rate, and invasiveness. Progression may be spontaneous, influenced by environmental factors, or mediated by progressors. Resulting tumors may be either benign or malignant.

The mechanisms of metal-induced carcinogenesis are less clear than for genotoxic agents, and they probably involve a number of biochemical events that indirectly affect the integrity of the genome of particular cells. It has been found, for example, that nickel and arsenic exposure can induce DNA hypermethylation (Tang 2000), and that arsenic inhibits DNA ligase I and II, which plays a role in DNA repair. Such non-mutational epigenetic changes could also result in suppression of tumor-suppressor genes, such as the p53 gene, triggering tumorigenesis. These "indirect" mechanisms of carcinogenesis often translate into what are believed to be sublinear, or hockey-stick type dose–response curves. If this occurs, extrapolations from high-dose effects are likely to overpredict risks at low doses (Rudel et al. 1996).

Background tumor formation is a normal observation in control animals from rodent carcinogenicity studies. The incidence of spontaneous tumors varies between tissues, and the susceptibility of a given tissue or organ varies between species and strain, and can be influenced by other factors including diet. For example, it has been found, in lifetime studies, that the incidence of testicular interstitial cell adenoma is 49% in F-344 rats compared with 9% among FBNF1 rats (Haseman et al. 1998). A 40% food restriction lowered incidences in these strains to 19 and 4%, respectively.

24.3.3.2 Non-Cancer Effects

The non-cancer end points are subdivided into acute or chronic exposures and effects with different levels of further organization relating to the target tissue, organ, or system, depending on the risk management need.

The critical biological target is identified from an exhaustive search of published and unpublished literature until a reliable study shows an effect that occurs at doses below those causing any other measured effect. As these studies are typically done on laboratory animals, the most sensitive species and sex is used as the basis for identifying the critical dose.

One organizational scheme for target organ systems is shown in Table 24.3.

Generally speaking, the fewer the categories of classification, the more conservative or public health protective an assessment of the impacts of a mixture of chemicals.

24.3.4 Identifying Thresholds

The highest experimental "no observed adverse effect" level (NOAEL) is the basis for most practical thresholds of toxicological effects. The critical NOAEL is combined with uncertainty factors (UFs) to provide a margin of safety for the exposed population.

$$\begin{aligned}&\text{Acceptable Daily Intake/Air Quality Standard/}\\&\text{Water Quality Standard/ etc.}\\&\quad = \text{Experimental NOAEL} \div \text{Margin of Safety}\end{aligned}$$

or:

$$\text{NOAEL} \div (\text{UF}_\text{A} * \text{UF}_\text{H} * \text{UF}_\text{T} * \text{UF}_\text{D} * \text{UF}_\text{L})$$

Table 24.3 Target organs, tissues, or systems used by California Environmental Protection Agency for air toxicology non-cancer risk assessments

Eyes
Lung (upper and lower respiratory tracts)
Liver
Kidney
Immune system (e.g., reduced host-resistance)
Nervous system (peripheral and central)
Blood (i.e., anemia)
Endocrine
Reproductive and/or developmental
Gastrointestinal
Bones (i.e., fluorosis)

From OEHHA (2000)

A = animal to human; H = human variability; T = temporal factors; D = data gaps/quality; and L = LOAEL* to NOAEL

UFs = 1–10, depending on source of data, with a maximum cumulative UF of 3,000 (U. S. EPA 2002). The definition of NOAELs and LOAELs (lowest observed adverse effect levels) are shown in Table 24.4.

The NOAEL concept and its implications for human health risk assessment is shown in Fig. 24.2.

An alternative to the NOAEL is the benchmark dose (BMD) approach, which is favored by regulatory agencies when the data are of sufficient quality.

Some important characteristics of the BMD are

- Benchmark doses are used by the U.S. Environmental Protection Agency (U.S. EPA) and the World Health Organization (WHO) in many non-cancer risk assessments.
- BMDs use dose–response information.
- BMDs take into account statistical uncertainty and sample sizes.
- BMDs assume a distribution of responses rather than a point estimate.
- BMDs can assume a threshold and account for background responses as options.
- Dichotomous and continuous data can be used to calculate BMDs.

A number of mathematical models exist for calculating BMDs. Public domain software developed by the U.S. EPA for benchmark dose calculations can be found at the U.S. EPA Web site (www.epa.gov).

An example of a BMD relationship is shown in Fig. 24.3. A percentage response in the test population is taken to be the practical threshold for the adverse effect. Typically this is a 5–10% response incidence.

24.3.4.1 What Is an Adverse Effect?

Not all biological responses to a toxicant are considered adverse. Some effects are considered to be adaptive responses that have no short- or long-term consequences. This gray area in the definition of a threshold for adverse effects is one of considerable discussion and debate in regulatory toxicology. Table 24.4 illustrates the types of definitions that have been assigned as severity qualifiers to adverse effects by the U.S. EPA.

Adverse effects from human clinical and epidemiological studies have also been defined in the context of regulatory purposes (Table 24.5).

24.3.5 Variation in Effects: Genetic and Phenotypic Variability in Susceptibility

The variability in response to a chemical agent can create difficulties in establishing statistically significant associations in epidemiology. However, the variability in a given response often has biological roots that are increasingly important as regulatory and public health agencies try to determine ways to identify and protect the most sensitive individuals in society from adverse toxicological effects.

Toxicological risk assessment has traditionally relied on estimates of no-effect thresholds (i.e., NOAELs) combined with uncertainty factors, which are intended to account for the fact that individual variability in response exists. Default values between 1 and 10 are typically used in these calculations because the precise amount of individual variability is not known.

More recently, investigators have been determining the degree of variability in physiological parameters. For example, the U.S. National Research Council (NRC) reports a range of elimination half-lives for 13 different drugs that are 0.7- to 17-fold greater in newborn infants than in adults (NRC 1993). For a given rate of exposure, these drugs would remain for a longer time in an infant's body, thus likely increasing the infant's susceptibility.

The individual variability in some key toxicokinetic mechanisms have been described using clinical and epidemiological studies as shown in Table 24.6.

Table 24.6 shows that although toxicological risk assessment relies on the assumption that a ten-fold uncertainty factor for individual variability is health protective, in some cases clearly a factor of 10 comes nowhere near the amount of variability that actually exists. However, this type of research with application to risk assessment is relatively recent, and more data are needed in order to more precisely define what are appropriate default values for this parameter.

Although the long neglected field of human susceptibility to environmental toxicants is currently receiving renewed

Table 24.4 U.S. EPA—Interim guidelines for development of inhalation reference concentrations[a]

Effect or no effect level	Rank	General effect	Cal/EPA severity ranking
NOEL	0	No observed effects	<Mild
NOAEL	1	Enzyme induction or other biochemical changes, consistent with possible mechanism of action, with no pathologic changes and no change in organ weights	<Mild
NOAEL	2	Enzyme induction and subcellular proliferation or other changes in organelles, consistent with possible mechanism of action, but no other apparent effects	<Mild
NOAEL	3	Hyperplasia, hypertrophy, or atrophy, but no change in organ weights	≤Mild
NOAEL/ LOAEL	4	Hyperplasia, hypertrophy, or atrophy, with changes in organ weights	Mild
LOAEL	5	Reversible cellular changes including cloudy swelling, hydropic change, or fatty changes	Mild/severe
(LO)AEL[b]	6	Degenerative or necrotic tissue changes with no apparent decrement in organ function	Severe
(LO)AEL/FEL	7	Reversible slight changes in organ function	Severe
FEL	8	Pathological changes with definite organ dysfunction that are unlikely to be fully reversible	Severe
FEL	9	Pronounced pathologic changes with severe organ dysfunction with long-term sequelae	Severe
FEL	10	Death or pronounced life-shortening	Life-threatening

[a]Adapted from Hartung (1987)
[b]The parentheses around the LO in the acronym LOAEL refer to the fact that any study may have a series of doses that evoke toxic effects of rank 5–7. All such doses are referred to as adverse effect levels (AELs). The lowest AEL is the (LO)AEL

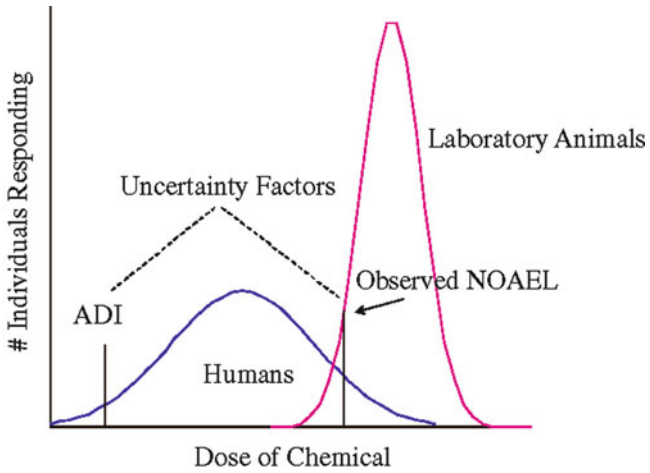


Fig. 24.2 Concept of extrapolation from animal experimental NOAEL to humans in deriving an acceptable daily intake (*ADI*). The laboratory animals are assumed to be, on average, less sensitive to the toxicant than humans, and are also assumed to have less individual variability (indicated by the relatively narrow distribution on the right)

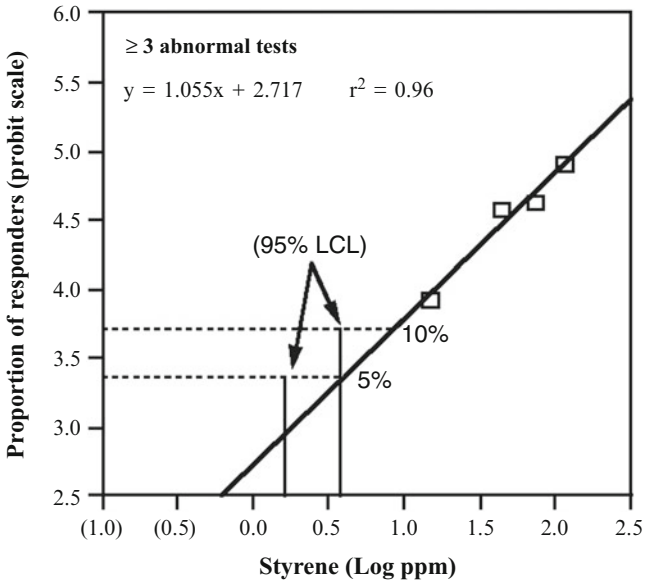


Fig. 24.3 Benchmark dose effect of styrene on neurobehavioral tests in plastics workers. (Data from Mutti et al. 1984; Rabovsky et al. 2001.) LCL is the statistical 95% lower confidence limit on the probit line

attention, there is only scant literature on factors influencing susceptibility to heavy metals (Gochfeld 1997).

24.4 Toxicology and Epidemiology: How Is Environmental Medicine Studied?

Toxicology, risk assessment, and epidemiology all have important roles in the study of environmental medicine. The scientific confidence in the public health actions that are taken in response to environmental contaminants is a function of how thoroughly each of these areas are addressed. History shows that heavy emphasis on one discipline alone can lead to actions that later are determined to be unfounded. The case of saccharin is one example. Saccharin was, for many years, considered to be a probable human carcinogen by the IARC and the U.S. National Toxicology Program due to its ability to induce bladder cancer in male rats. Many studies were done on rats to confirm this effect, and a dose–response relationship was developed and widely accepted. At one point, saccharin was one of the most studied compounds in terms of long-term cancer bioassays

Table 24.5 California environmental protection agency symptom and sign severity rating for human studies

Severity rating	Symptoms	Signs/laboratory findings
Mild adverse	Mild subjective complaints with few to no objective findings:	Statistically significant findings of pre-clinical significance:
	Mild mucous membrane (eye, nose, throat) irritation	Mild conjunctivitis
	Mild skin irritation	Mild lung function changes
	Mild headache, dizziness, nausea	Abnormal immunotoxicity test results
		Mild decreases in hemoglobin concentration
Severe adverse	Potentially disabling effects that affect one's judgement and ability to take protective actions; prolonged exposure may result in irreversible effects:	Clinically significant findings: Findings consistent with central or peripheral nervous system toxicity
		Loss of consciousness
		Hemolysis
	Severe mucous membrane irritation	Asthma exacerbation
	Blurry vision	"Mild" pulmonary edema
	Shortness of breath, wheezing	Clinically significant lung function changes
	Severe nausea	Cardiac ischemia
	Severe headache	Some cardiac arrhythmias (e.g., atrial fibrillation)
	Incoordination	Renal insufficiency
	Drowsiness	Hepatitis
	Panic, confusion	Reproductive/developmental end points (e.g., infertility, spontaneous abortion, congenital anomalies)
Life-threatening		Potentially lethal effects:
		Severe pulmonary edema
		Respiratory arrest
		Ventricular arrhythmias
		Cardiac arrest

Table 24.6 Reported ranges of variability in parameters related to susceptibility

Parameter	Width of 90% range for an average chemical or test
Systemic uptake pharmacokinetics	
Breathing rates	1.8- to 2.8-fold
Half-life or elimination	2.3- to 5.8-fold
Skin absorption	2.5-fold
Maximum blood concentration	2.3- to 11-fold
Area under concentration curve	3.0- to 8.1-fold
Blood concentration measurements	
Serum PCB concentrations	12-fold
Blood mercury levels	12-fold
Blood lead levels	13-fold
Pharmacodynamic or combined kinetic and dynamic parameters	
Cisplatin hearing loss	4.1-fold
Effects of methyl mercury in adults	12- to 78-fold
Fetal/developmental effects of methylmercury	460- to 10,000-fold
Hemodynamic responses to nitrendipine	8.3- to 17-fold
FEV1 response to cigarette smoke	8.3-fold
Salbutamol FEV1 increase (asthma)	128-fold
Acute toxicity	
Death from compounds metabolized by plasma cholinesterase	5.5-fold
Death from parathion	12-fold

Adapted from Hattis (1996)

and dollars spent. This carcinogen designation was then driven by toxicity testing because the mechanism for cancer formation from saccharin was not known, nor was it known why only male rats were susceptible, and not mice, any other species, or female rats (see also Chaps. 22 and 25, this volume).

Table 24.7 Basic framework for the identification, risk assessment, and epidemiological study of environmental contaminants

Discipline/Area of research	Goals and outcomes
Basic biochemical toxicological research	Identification of biochemical mechanisms
	Hypothesis of downstream physiological end points
	Development of biomarkers
Toxicity testing	Descriptive toxicology test battery (acute, chronic, mutagenicity, carcinogenicity, reproductive/developmental toxicity, sensitization)
	Identification of critical effects starting from high doses and reducing dose until no effect is seen
	Establishment of dose-response relationship
	Multiple species and both genders
Risk assessment and management	Identification of a critical dose (no observed effect level or benchmark dose)
	Application of margin of safety (uncertainty factors)
	Establishment of an acceptable chronic dose level
Epidemiological research	Probe established toxicological limits and biochemical mechanisms determined from animal studies to determine if they are applicable to humans
	Describe the probabilities of adverse effects in humans occurring following exposure
	Signal for further mechanistic research or the need to develop human biomarkers for exposure or effect

The epidemiology research showed that there was no evidence for elevated incidence of cancers in humans using saccharin, but there were concerns that the latency period required for cancer to develop in people had not been allowed to mature. The mechanism of cancer formation was later found to be due to induction of a male-rat-specific protein (alpha-2U globulin) that caused chronic irritation of the bladder, which led to bladder cancer at high doses (Turner et al. 2001). This led regulatory agencies to de-list saccharin as a human carcinogen in 1999 (IARC 1999). The relationship of these disciplines in environmental medicine is shown in Table 24.7.

24.4.1 Concepts of Dose and Duration

24.4.1.1 Cancer Potency

Chemicals that are carcinogenic to humans are identified through several authoritative bodies using established weight of evidence approaches. The IARC, the U.S. National Toxicology Program (NTP), and the U.S. EPA are the three most authoritative sources for identification of new and existing carcinogens.

There are 48 individual chemicals that are known human carcinogens under the IARC classification scheme.

The data from cancer bioassays is usually fit to a linearized multistage model, which is of the form originally described by Crump (1984).

$$\text{Linearised multistage model: } P(d) = 1 - e^{-(q_0 + q_1 d + q_2 d2 + \ldots + q_k dk)}$$

$$P(0) = 1 - e^{-q_0}$$

where P(d) is the probability of developing a tumor at a given dose rate and P(0) is the estimated background incidence. The q parameters are derived from the model.

Cancer potency is usually described by a dose–response slope factor (q1) and its respective 95% confidence limit (q1). The units of potency are usually in $(mg/kg/day)^{-1}$ which, when combined with a given dose level, gives a unitless risk factor (e.g., 10^{-6}, or 1 in a million). This concept applies to daily doses experienced over a chronic period, up to an entire lifetime. Much less is known about estimating risks from single acute exposures to carcinogens, and there are few animal studies through which to judge the difference in the dose–response relationship.

Often, a rodent cancer bioassay yields information only at a few doses that far exceed those found in the environment, and the dose–response is extended to very low doses.

Potency factors assume the absence of a threshold for cancer at low doses. Therefore, it may be inappropriate to apply potency estimates to carcinogens that are thought to have a threshold (e.g., non-genotoxic carcinogens, such as dioxin).

The utility of cancer potency factors lies in their use in cancer risk projections. When calculating cancer risks, if exposures to specific carcinogens can be quantified, then it is assumed that the risk of getting cancer from a long-term exposure is a function of exposure (i.e., mg/kg/day) multiplied by the respective cancer potency factor $(mg/kg/day)^{-1}$. The two most critical assumptions with this calculation are

1. The assumption that the basis for the cancer potency factor is a mechanism that applies to human physiology
2. The tumors seen at high doses in experimental studies are part of a linear or curvilinear function that extends to low doses that are more relevant to environmental exposures

Table 24.8 Expected variability in the abnormal responses to styrene

Fold differences between percentiles			
Number of abnormal neurological tests	Probit slope (1/log GSD)[a]	(5–95% range)[b]	(1–99% range)[b]
≥1	1.346	270	2,860
≥2	1.225	485	7,190
≥3	1.055	1,300	25,540

From Rabovsky et al. (2001)

[a]The probit slope was taken from the log-probit analysis using Tox-Risk V.3.5(30) described for Table 24.3

[b]The range of variability is obtained from the equation:

$$[\log(X) - \log(GM)] / \log(GSD) = t(\alpha)$$

where X is the distance on the dose axis from the mean to the 95th (or the 99th) percentile. To obtain the range of variability, the same distance from the mean to the 5th (or 1st) percentile is added to "X" and the bounds of the 5th–95th (or 1st–99th) percentile are the ranges of variability. Other abbreviations are GM, geometric mean; GSD, geometric standard deviation; and t (α), t distribution value at the desired test level (i.e., 0.05 or 0.01)

The California Environmental Protection Agency (Cal/EPA) has an active program that identifies carcinogens and lists potency values for each on its Web site (www.oehha.ca.gov).

24.4.1.2 Dose–Response Slope

The slope of the response curve from a toxicology study imparts significant meaning to the causal relationship of the chemical and effect being studied. The dose–response relationship also helps in understanding the risks from exposure to low doses.

Some additional uses of the dose–response relationship include the characterization of individual variability in the measured response to the chemical. Table 24.8 shows the relationship of probit slope term to the individual variability in a battery of neurological tests in response to styrene exposure in the workplace (Rabovsky et al. 2001). In general, the more shallow the dose–response slope, the greater the variability in the test population. When using a log-normal model to describe a dose–response relationship, slope terms that approach 1.0 show very large individual variability (Table 24.8).

24.4.2 Estimating Exposure (Analytical Chemistry, Biomarkers, and Modeling)

Exposures to metals can be measured or estimated in various ways. For chronic dietary exposures, blood samples for the metal may be the most direct and simple measure of exposure. For historical exposures, blood samples may not be appropriate if the body has had time to depurate the metal from the bloodstream. Lead has a half-life of about 25 days in the blood. Methylmercury has a half-life for elimination from the body of about 50 days in adult males and females. However, infants have poor elimination of organic and inorganic heavy metals in the first 6 months prior to development of metal transport systems. It is possible that mercury is eliminated much more slowly in the child. Cadmium is removed from the human body much more slowly, so that the elimination half-life is on the order of 20 years (Gochfeld and Syers 2001).

Metallothioneins (MTs) are metal-binding proteins that are considered central in the intracellular regulation of metals such as copper, zinc, and cadmium. Variability in tissue MT levels influence susceptibility of tissues and species to the toxic effects of some metals, such as cadmium and mercury.

Metallothionein is the major protein thiol induced in cells exposed to cytokines and bacterial products (Schwarz et al. 1995). This protein is inducible by exposure to some metals, and it appears to impart protection from some adverse effects. For example, zinc exposure during fetal life results in MT induction (Mengheri et al. 1993). Zinc pretreatment lowers cadmium carcinogenicity in laboratory animals presumably through induction of MT (Coogan et al. 1992). Induction of MT protects against cadmium toxicity in rats (Singh and Rana 2002). Similarly, MT is thought to contribute to the placental barrier to the transfer of mercury to the fetus (Yamamoto et al. 2001).

24.4.3 Is There a Health Effect? Animal Models

The widespread use of animal models in toxicity testing continually raises the possibility that a given adverse effect may not be relevant to humans, or that the animal model studied may substantially underpredict human risk. The case of saccharin already described, is an example of the former. The converse of the saccharin case, however, can also be true. Arsenic, for example, is a human carcinogen and a human neurotoxin, but it does not appear to cause cancer in laboratory animals at doses that are considered carcinogenic to humans. Thalidomide is a prominent case in which a relatively minor heart valve developmental defect in rodents translated into major deformities in human babies. Benzene is another example where humans appear to be the most sensitive species tested for its critical effect (leukemia).

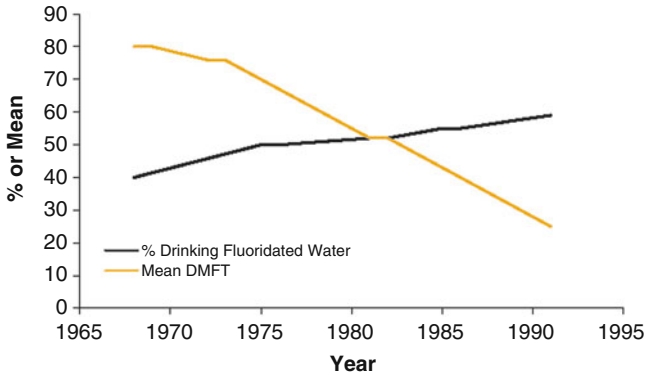


Fig. 24.4 Percentage of population residing in areas with fluoridated community water systems and mean number of decayed, missing (because of caries), or filled permanent teeth (DMFT) among children aged 12 years, United States 1967–1992. (From National Center for Health Statistics 1974, 1981; National Institute of Dental Research 1989; CDC unpublished data, 1988–1944)

Thus, animal models show a number of examples of failure in predicting effects on humans. However, by and large, animal models, particularly when multiple species are tested, are thought to provide adequate evidence for an initial risk assessment of a substance to proceed, provided that adequate margins of safety are used. Epidemiological studies will always be needed to ascertain if the safe limits proposed in risk assessment are adequately protective.

In the case of arsenic, there is significant interspecies variability in metabolism. However, in most mammals, arsenic is metabolized and detoxified by the addition of methyl groups, and is eventually excreted, primarily in urine, in a mono- or dimethylated form. The methylation pathway is dependent on folate metabolism, which provides methyl groups through conversion of S-adenosylmethionine to S-adenosylhomocysteine. The availability of methyl groups from this pathway is thus essential to the metabolic detoxification of arsenic and may at least partially account for both inter- and intraspecies differences in sensitivity to this toxicant. Human serum, for example, is folate-deficient compared to rodent serum (unpublished observation), and this observation may explain experimental findings that arsenic, a known carcinogen in humans, does not induce cancer in rodents.

Given the critical role of methylation in the disposition of arsenic, further characterization of the enzymatic basis of arsenic methylation is required. To date, human arsenic methyltransferase has not been isolated, but transferases are generally polymorphic. Understanding the factors affecting human sensitivity would improve the arsenic risk assessment. The objective of this section is to evaluate variations in arsenic metabolism as reflected in variations in urinary metabolites or other biomarkers of exposure as associated with the exposure level, nutritional status, genetic factors, and other variables. Included in this area are studies to improve mass balance data on typical human metabolism of arsenic at various doses and chemical forms. There is a need for the development and refinement of assay procedures to characterize arsenic methyltransferases in human tissues. In addition, these studies would compare biomarkers of arsenic metabolism in individuals exposed to varying levels of arsenic with differences that include, but are not limited to, nutritional status, age, sex, and genetic variations.

24.4.4 Evidence at the Population Level

Further evidence (follows from animal studies section) of the relationship between an exposure and a health effect can be sought through the epidemiological study of human populations. Such epidemiological studies take three main forms: cross-sectional, case–control, and cohort. Each study design has its own strengths and weaknesses, and *all answer different questions.*

The cross-sectional study is in essence a survey to determine how common something is, or at what level it occurs. In Fig. 24.4, for example, the prevalence of caries in the teeth of children was determined by cross-sectional survey. By carrying out such surveys in populations using different drinking water supplies, it was possible to answer the question: Are caries more common in populations with low fluoride exposure? An advantage of the cross-sectional study is that it is relatively quick, easy, and cheap to administer. Disadvantages include that it cannot take into account individual exposures and is subject to confounding. For example, if the people drinking low-fluoride water are also poorer, they may have more caries for dietary reasons.

The case–control study starts with identified cases of a disease and compares their exposure to controls who do not have the disease. In the case study on amoebic meningitis in Sect. 22.2, for example, children with the disease would be identified, and the frequency of their hot spring usage would be compared to that of children without the disease. With such a study it is possible to answer the question: Did children with amoebic meningitis swim in hot springs more frequently than children without amoebic meningitis? If yes, then hot springs are implicated by association, but note that this does not demonstrate a causal relationship.

An advantage of the case–control study is that risk factors for very rare events can be identified. A disadvantage is that they contain no denominator information: note that we have no idea how many children swam in the hot springs to give rise to the few rare cases that we are studying, and we therefore cannot say how dangerous that activity is.

The cohort study, seen by many as the epidemiological gold standard, starts with a group of people with a common exposure.

The risk factors of cases with the disease are then compared to the risk factors of controls without the disease, all still within this group or cohort of people. For example, a cohort might be defined as all people living in a village with high levels of arsenic in the water supply. Some of these people would develop skin cancer and become "cases," whereas others would not develop skin cancer and would instead become "controls." The epidemiologist would record the water consumption habits of all members of the cohort. With such a study it is possible to ask the question: Do people who drink more water develop skin cancer more frequently? If yes, the arsenic-rich water would be implicated by association, but again, a causal relationship has not been established. The advantage of this study design is that it does provide denominator information, and it is therefore possible to directly calculate the risk of developing cancer from drinking the water. The disadvantage is that this approach is both very time-consuming and potentially very expensive as very large cohorts are often required to give the study enough power to achieve statistical significance.

As mentioned, all of these studies are epidemiological, and can at best establish association between exposure and effect. To establish a causal relationship between an exposure and a health effect, several other criteria are generally considered, including the following (Hill 1965): a temporal relationship between exposure and effect; a biologically plausible relationship, including a dose–response effect; associations by epidemiological studies that are both strong and consistent; and reversibility of the association if the exposure is removed.

A number of approaches to the study of environmental medicine have been outlined in this chapter. Let's not forget that the aim of such a study is to devise interventions to reduce morbidity and mortality, which is the subject of the following section.

24.5 Health Protection: How Can Adverse Effects Be Minimized?

In Sect. 22.2 (and Table 24.1) we explored a useful framework for considering health effects from environmental exposures. Other than simply to satisfy the human compulsion to classify things, such a framework is a prerequisite to health protection: in order to minimize the adverse health effects of environmental exposure, we must first be able to consider all possible exposures. We can then prioritize surveillance and intervention so as to maximize health gain from the use of our (invariably limited) resources.

24.5.1 Risk Assessment

One common approach to risk assessment follows the four simple steps of hazard description, dose–response estimation, exposure assessment, and risk characterization. These can be simply illustrated here using an example such as exposure to volcanic gases during an eruption (see Chap. 10, *Volcanic Emissions and Health*, this volume). The hazard description would list the gases concerned, e.g., CO_2, H_2S, and HF, and detail the volume, concentration, and duration of emission. The dose–response estimation (effect) would consider, usually graphically, the relationship between the amount of gas and the relevant health effect. Thus respiratory distress may increase in a linear fashion with the dose of gas inhaled, and a threshold may exist at which consciousness is lost.

The exposure assessment would take into account how many people are likely to suffer from such health effects, for how long, and if any particular groups are at greater risk. For example, the number of people living in a village downwind from the eruption will be relevant, as will be the number of children in cots (arguably more susceptible to suffocation from gas that is heavier than air).

The risk characterization summarizes all this information into a prediction about the likely outcome if the hazard goes unchecked. For example, 2 deaths in children and 100 adults with respiratory distress may be predicted for such a village if it is not evacuated. Medical and civil defense authorities will then make a decision about the appropriate deployment of resources based on this risk assessment. Should the two buses available be deployed to evacuate this village, or would they be better deployed to evacuate the village on the other side of the mountain that is potentially in the path of a lava flow?

The problem is that even with the best risk assessment, there is always a degree of uncertainty, and decisions about intervention can therefore be difficult to make. It seemed sensible at one time to recommend that well water be used in Bangladesh to avoid the risk of gastrointestinal disease from pathogens in surface waters. The well water, however, turned out to contain arsenic at levels that were not anticipated (see Chap. 12, *Arsenic in Groundwater and the Environment*, this volume). It is sometimes unclear if a known exposure constitutes a health risk or not. For example, low-level chronic exposure to geothermal hydrogen sulfide has not been demonstrated to cause clear-cut pathology, but one might expect it to on the basis of respiratory and neurological toxicity at higher levels of exposure. In such cases, practitioners of environmental medicine often apply the "precautionary principle," which states that any substance suspected of adversely affecting health should be avoided (as far as possible) until proven otherwise.

24.5.2 Surveillance

Surveillance is the term used, in environmental medicine, to refer to the ongoing monitoring processes that inform public health intervention. Surveillance is the use of monitoring data to attempt to reduce morbid-ity and mortality: without the completion of this criterion, surveillance would be no more than data collection.

A good example of a surveillance system is the "notifiable disease" list, which compels medical practitioners in most countries to notify some authority responsible for disease control on each occasion that he makes a diagnosis of one of the diseases on the list. Lead poisoning causes learning difficulties and neurological complications, and is on the notifiable list in most countries. The relevant authority collects and analyzes the notified case data to determine the source of the exposure; for example, toddlers are commonly poisoned by ingesting flakes of old lead-based paint in poorly maintained houses, and adults may have occupational exposure in industries such as battery recycling.

The "surveillance loop" is completed when this information is used to impose recommendations for home improvement or factory practices. Note that the frequency of notified cases provides feedback as to the effectiveness of the interventions (at least on a regional or national scale), thereby forming a genuine "intelligence loop."

In the example above, health surveillance was carried out for the health effects of exposure to the element lead. It is also possible to carry out *hazard* surveillance, where the environmental levels of, e.g., lead are monitored directly, rather than (or as well as) monitoring the health effects caused by lead. The same requirements for surveillance hold, and the data on lead concentration and distribution are used to inform public health intervention: the lead level in the factory can now be kept below safety limits without the need for workers to develop symptomatic disease (which is obviously preferable). The implementation of one or both types of surveillance, health and hazard, are integral to the practice of environmental medicine. In conjunction with the environmental medicine framework discussed in Sect. 22.2, it should now be possible for the reader to devise, at least at the theoretical level, surveillance systems to deal with most situations that might be encountered in medical geology.

24.5.3 Intervention

24.5.3.1 Intervention Success Story: The Fluoridation of Public Water Supplies

In the 1930s American dental epidemiologists noted a considerable regional variation in the rate of dental caries. By carrying out a cross-sectional study (as described in Sect. 24.4.4), Dean (1936) established an association between differing levels of fluoride (F^-) in regional water supplies and caries rates in the populations drinking those waters. Although fluoride is a toxin that in high concentration can kill, it appeared in this case that a small amount in the drinking water was beneficial to dental health. In populations supplied with drinking water containing about 1 mg/L F^-, dental caries rates were reduced by about 50%. The variation in fluoride levels in drinking waters in those days was entirely natural, and resulted from the dissolution of fluorine by surface waters as they coursed over fluorine-rich substrates such as geological deposits of marine origin. Population health was therefore directly affected by living in an area with particular geological characteristics, a situation that is in essence the theme of this book.

Public health authorities now began to ask the obvious question: For populations not living in fluoride-rich areas, could dental health be improved by artificially supplementing F^- levels in drinking water? The suggestion was examined from several perspectives including cost-benefit considerations and human health risk assessment (Sect. 24.5.1). The risk was (and remains) one of balancing the fluoride delivery carefully so as to achieve the reduction in caries rate without also causing fluorosis—the condition of excessive chronic fluoride ingestion. Many middle-aged people in the western world have dental fluorosis (stained patches of brittle enamel) from receiving both fluoridated water and fluoride tablets as children. Industrial fluorosis is also known, largely historically, as an occupational hazard in the aluminum and fertilizer industries. In these cases, fluoride deposition into bone increases bone and cartilage density which can result in restricted flexibility and movement, especially around the lumbar spine (Derryberry et al. 1963). In the developing world, many people in specific geographic areas have a far more serious form of fluorosis, which is known as endemic fluorosis. This is a potentially crippling disease, the major manifestation of which is the overgrowth and distortion of bone, with tendinous, articular, and neurological involvement. Such severe disease only results after many decades of ingestion of drinking water with 10 mg/L F^- or more, a concentration one order of magnitude higher than the 1 mg/L observed to be beneficial by Dean (1936). Since the 1950s therefore, public drinking water supplies in most developed countries have been topped up to about 1 mg/L of F^-, with minor adjustments to account for differing levels of climate-dependent water consumption. The improvement in the dental health of children has been remarkable since that time, with large reductions in caries rates.

The debate continues to this day about the relative contribution of water fluoridation to this reduction in dental caries rates, because there have been concurrent improvements in nutrition, dental hygiene, and dental services that also contribute to a reduction in the incidence of caries. The issue has become further blurred by the advent of fluoridated

toothpaste, sports drinks, and other sources of fluoride that dilute the improvement attributed to fluoridated water alone. It is known, however, that populations receiving fluoridated drinking water show better dental health, on average, than do control populations without fluoridated water (WHO 1994). Health surveillance data (caries rates) collected by school dental services show that this continues to be the case, and hazard surveillance data (water F^- levels) collected by water treatment plants continue to ensure that populations are not at risk of fluorosis. The fluoridation of public water supplies is therefore a good example of a successful public health intervention informed by ongoing surveillance.

24.6 Case Study: Arsenic

24.6.1 Exposure to Arsenic

Arsenic is a metalloid element found ubiquitously in nature. It is present in the Earth's crust with an average concentration of 2 mg/kg. Arsenic can be found in soil, air, water, food, and some manufactured chemicals. Humans can be exposed to arsenic from either natural sources or anthropogenic sources. Natural sources of arsenic include rocks (soil), volcanic emissions, undersea smokers, and extraterrestrial material. Volcanic emission is the most important natural source of arsenic. Arsenic can be found in more than 200 mineral species, of which the most common is arsenopyrite. Anthropogenically, arsenic can be found in products of herbicides, fertilizers, pesticides, leather treatment, cotton desiccants, wood preservation, animal feeds as food additives, and pharmaceuticals (see also Chaps. 12 and 25, this volume).

Humans can be exposed to arsenic through ingestion of arsenic-containing water, food, and drugs (such as Fowler's solution containing 1% of potassium arsenite used to treat psoriasis and arsenic trioxide used to treat leukemia). Airborne arsenic can be absorbed into the bloodstream in workers involved in the processing of copper, gold, and lead ores; in the production and use of arsenic-containing pesticides; in the manufacturing of glass, semi-conductors, and pharmaceutical substances; in using arsenic as pigments and dyes; in burning coal containing high levels of arsenic (Guizhou Province, China); in smoking high-arsenic-contaminated tobacco; and in chimney sweeping.

Water contamination is the most common source of arsenic exposure. Currently, Bangladesh and West Bengal, India, have the most serious problem of groundwater contamination with arsenic in the world. Tracing back the history of these areas, surface water was replaced by tubewell water in 30 years ago to fight against infectious diarrheal diseases. These programs to provide "safe" drinking water from underground unexpectedly brought up another health problem of arsenic hazards. It is estimated that more than 95% of the 120 million people in Bangladesh drink tubewell water and more than one-third of the tubewell water contains arsenic above 0.05 mg/L (the maximum allowable level recommended by the WHO). In 2001 the U. S. EPA lowered the maximum allowable level of arsenic in drinking water from 0.05 mg/L to 0.01 mg/L. High arsenic level in drinking water is also reported in countries such as Argentina, Australia, Chile, China, Hungary, Mexico, Peru, Taiwan, Thailand, and the United States. See *Arsenic in Groundwater and the Environment* (Chap. 12), this volume.

24.6.2 Effects of Arsenic

Arsenic exists in four valence states: $^{-}3$, 0, +3, and +5. Elemental arsenic and arsine ($^{-}3$) exist in strongly reducing environments; arsenite (+3) is the dominant form in moderately reducing conditions; and arsenate (+5) is stable in oxygenated environments. Inorganic forms of arsenic are much more toxic than organic forms found abundant in seafoods, and, in general, inorganic arsenic of trivalent forms are more toxic than pentavalent forms. Immediate symptoms of an acute poisoning typically include vomiting, esophageal and abdominal pain, and bloody "rice water" diarrhea. However, a variety of symptoms and signs involving the gastrointestinal, dermal, nervous, renal, hepatic, hematopoietic, cardiovascular, respiratory, and ophthalmic systems can be observed (Table 24.9) (Chen et al. 1999; Tseng 1999). Treatment with chelating agents such as dimercaprol or dimercaptosuccinic acid during acute intoxication is classical but may have varying effects. Chelating agents may not be effective in chronic poisoning. See Biological Responses of Elements, this volume.

Long-term exposure to arsenic can cause a variety of cancers involving the skin (squamous cell and basal cell carcinoma), lung, bladder, kidney, and liver. Although arsenic does not induce point mutations, it can cause chromosomal aberrations, affect methylation and repair of DNA, induce cell proliferation, transform cells, and promote tumors.

A wide spectrum of non-cancerous diseases and clinical problems are also reported in long-term arsenic exposure (Table 24.9). Arsenic skin lesions are characterized by the coexistence of hyper- and hypopigmentation giving rise to a raindrop pattern (Fig. 24.5) and hyperkeratosis of the palms and soles (Fig. 24.6). In recent years, long-term exposure to arsenic from drinking water has also been found to be highly associated with hypertension and diabetes mellitus (Tseng et al. 2000, 2002). Preclinical microcirculatory defects (Tseng et al. 1995) and arterial insufficiency (Tseng et al. 1994) can also been demonstrated in subjects exposed to arsenic. Arsenic could also cause lower IQs in children

Table 24.9 Non-cancerous effects of arsenic on humans

Organ system	Diseases or symptoms/signs
Dermal	Hypo- and hyperpigmentation (raindrop pattern), hyperkeratosis of palms and soles, exfoliative dermatitis, Bowen's disease (pre-cancerous lesions), facial edema, non-pitting pedal edema
Cardiovascular	Arrhythmia, pericarditis, ischemic heart disease, peripheral vascular disease, cerebral infarction, hypertension, microcirculatory defects
Gastrointestinal	Abdominal discomfort, anorexia, malabsorption, body weight loss
Nervous	Peripheral neuropathy involving sensory and motor systems, cranial nerve involvement, hearing loss, mental retardation, encephalopathy
Renal	Nephritis and proteinuria
Hepatic	Fatty degeneration, non-cirrhotic portal fibrosis, cirrhosis, hepatomegaly
Hematopoietic	Bone marrow hypoplasia, aplastic anemia , leukopenia, thrombocytopenia impaired folate metabolism, karyorrhexis
Respiratory	Rhino-pharyngo-laryngitis, tracheobronchitis, pulmonary insufficiency (emphysematous lesions)
Ophthalmic	Conjunctivitis
Reproductive	High perinatal mortality, low birth weight, spontaneous abortions, stillbirths, pre-eclampsia, congenital malformation
Metabolic	Diabetes mellitus, goiter

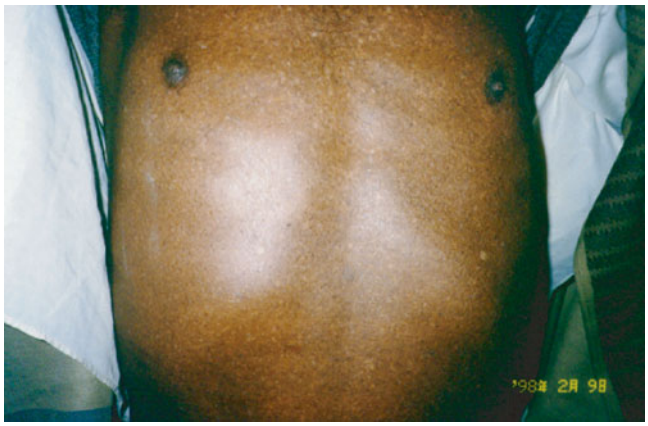


Fig. 24.5 Raindrop pattern of hyper- and hypopigmented skin lesions in a patient with long-term arsenic intoxication

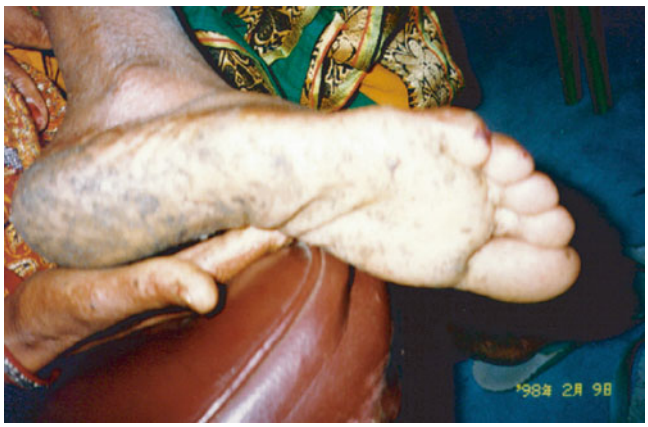


Fig. 24.6 Hyperkeratosis of the sole in a patient with long-term arsenic intoxication

exposed to arsenic in Thailand. The symptoms and signs that arsenic causes appear to differ between individuals, population groups, and geographic areas.

24.6.3 The Study of Arsenic Intoxication: The Example of Blackfoot Disease

Exposure to arsenic from drinking water in Taiwan has been shown to cause a severe peripheral vascular disease, which might progress from intermittent claudication, ulceration, gangrene, and spontaneous or surgical amputation (a case with spontaneous amputation is shown in Fig. 24.7). The disease has been named blackfoot disease after its clinical appearance (Tseng 1999, 2002). This disease was first reported in the early twentieth century and was confined to the southwestern coast of Taiwan where people used artesian well water from as deep as 100–300 m underground (Tseng et al. 1996). The prevalence ranged from 6.51 to 18.85 per 1,000 people in different villages. A series of epidemiologic studies and surveillance of the arsenic concentrations of the artesian wells carried out during the mid-twentieth century revealed the association between blackfoot disease and the consumption of high-arsenic-containing artesian well water. Besides arsenic intake from well water, residents in the endemic area could also be exposed to arsenic from a variety of other sources, because the artesian well water was extensively used for agricultural and piscicultural purposes. The amount of arsenic ingested by the residents of the endemic area was estimated to be as high as 1 mg per day. The lethal dose in humans is estimated to be 1 mg/kg/day.

Although studies in several other countries have demonstrated that arsenic exposure can be associated with some forms of peripheral vascular disease, similar endemic occurrence of severe blackfoot disease has not been observed. It is possible that nutritional status, coexistence of other factors, and interaction with other trace elements determine the development of the various clinical manifestations.

There are two main pathways of arsenic metabolism: the reduction reactions and the oxidative methylation reactions.

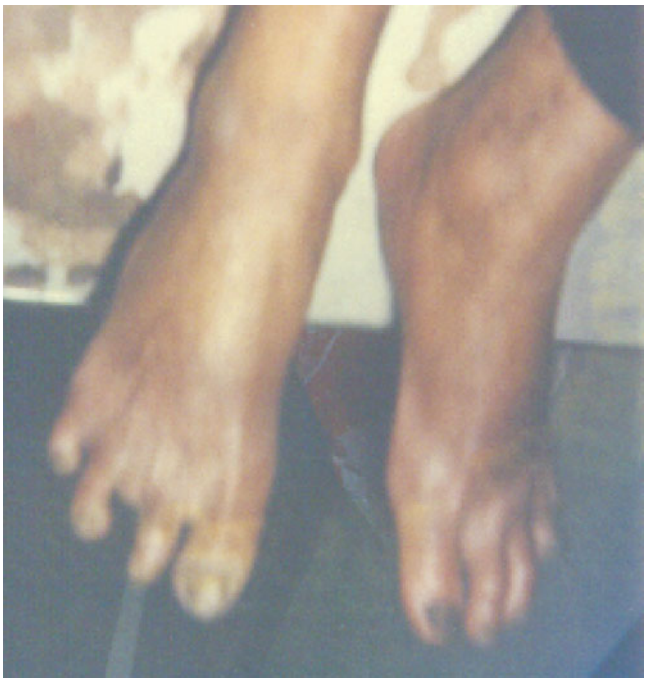


Fig. 24.7 Spontaneous amputation of the third toe in the right foot of a subject living in the blackfoot disease endemic area in Taiwan

Pentavalent arsenic is reduced to trivalent arsenic, followed by the methylation reactions to form mono-, di-, and trimethylated products. S-adenosyl methionine is the methyl donor and glutathione is an essential cofactor. Low amounts of methionine or protein in the diet decrease the methylation of inorganic arsenic in animals and similar nutritional deficiency was observed in the residents of blackfoot disease areas. Vitamin B_{12} is needed in the methylation process and insufficient intake of this vitamin in poor people and/or increased requirement during the reproduction ages in the women may put these people at higher risk for the development of arsenic-related health problems. Zinc and selenium may provide protective effects against the toxic effects of arsenic, and residents in the blackfoot disease areas were found to have deficiency of these elements in their diet. Lower levels of beta-carotene have also been shown to carry a higher risk of developing vascular disease and skin cancers in residents of the blackfoot disease endemic areas (Hsueh et al. 1997, 1998).

The absorption, distribution, and metabolism of arsenic differ significantly across species. Animals are less sensitive to the toxic effect of arsenic and most of the effects of long-term arsenic exposure on humans are not observed in animals. Genetic factors may play important roles on these metabolic cascades of arsenic; and thus, may also be involved in the development of the clinical effects of arsenic.

24.6.4 Public Health Intervention

The arsenic-related health problems in emerging endemic areas can be critical issues in public health. Drinking water poses the greatest threat to public health from arsenic. However, exposure from coal-burning, working environment, mining, and industrial emissions may also be significant in some areas. There is no universal definition of the disease caused by arsenic and there is no way to differentiate pathologically those vascular or cancerous lesions caused by arsenic from other etiologies. All of these complicate the assessment of the burden of arsenic on health. However, the use of interventional measures to terminate the hazards associated with arsenic should not wait until all these ambiguities are clarified. Up to now, there is no magic bullet for the treatment of the diseases associated with arsenic intoxication. The best strategy is prevention and avoidance of exposure. New sources of water and coal with low arsenic contents, techniques for arsenic removal from drinking water, decreasing industrial arsenic emissions, improving working environments, and promoting health education among the affected people are necessary.

As for the conditions in Bangladesh, a collaborative approach is required for scientists, health workers, policy makers, and members of the community to work together to plan and implement a sustainable and environmentally friendly water supply system. The government and professional people should have reliable, timely, and easily available information on the status of knowledge about the problems and what can be done to tackle the problems. Large-scale and concerted action is required from all sectors to take effective and practical remedial measures at affordable cost.

People need to be educated with correct knowledge about arsenic. Arsenic-related health problems are not infectious and they are manageable with a change of water-consumption pattern and adequate intake of food nutrients. Absorption of arsenic through the skin is minimal and thus hand-washing, bathing, laundry, etc., with water containing arsenic do not pose significant risk. However, arsenic-containing water is readily absorbed into the human body by the gastrointestinal tracts.

People at risk of arsenic exposure should be warned about the health hazards associated with arsenic. Tubewell water can be replaced by surface water. However, people should be educated to boil surface water before drinking to avoid the affliction of infectious diseases. Utilization of domestic arsenic removal devices are encouraged to obtain clean water. Adding alum or ferrous salts to arsenic-contaminated water to convert arsenic into insoluble substances is one of the methods. Rainwater harvesting can be helpful at a low cost during the monsoon season. Handy and low-cost technology

to detect arsenic components in water can be applied to identify safe water sources.

The successful eradication of blackfoot disease in Taiwan set an exemplar in the public health approach for the prevention of arsenic-related health hazards. Because of the link between the etiology of these endemic diseases with the artesian well water, the Provincial Government of Taiwan began to implement tap water supply systems to replace the artesian well water in the endemic areas. Programs that moved villagers to other residential areas had even been carried out in some seriously affected villages. Since the 1970s, the incidence rates of blackfoot disease decreased dramatically after the implementation of these public health measures. The eradication of blackfoot disease by changing the water supply system also demonstrated an excellent example that many environmental diseases can be successfully eradicated by removal of their vectors, even when the real etiology remains controversial.

See Also the Following Chapters. Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 9 (Biological Responses of Elements) • Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 23 (Environmental Epidemiology) • Chapter 25 (Environmental Pathology) • Chapter 26 (Toxicology) • Chapter 27 (Speciation of Trace Elements)

Further Reading

Barrett JC (1993) Mechanisms of multistep carcinogenesis and carcinogen risk assessment. Environ Health Perspect 100:9–120

Brown D (2001) Theoretical estimates of blood mercury levels from thiomersal injections using a one-compartment biokinetic model. www.iom.edu/iom/iomhome.nsf/WFiles/Brown/%24file

CDC (1988) Unpublished data, Third National Health and Nutrition Examination Survey, 1988–1994

CDC (1993) Fluoridation Census 1992, U. S. Department of Health and Human Services, Public Health Service, CDC, National Center for Prevention Services, Division of Oral Health, Atlanta, GA

Chan PC, Huff J (1997) Arsenic carcinogenesis in animals and in humans: mechanistic, experimental and epidemiological evidence. J Environ Sci Health Environ Carcinog Ecotoxicol Rev C15 (2):83–122

Chen CJ, Hsu LI, Tseng CH, Hsueh YM, Chiou HY (1999) Emerging epidemics of arseniasis in Asia. In: Chappell WR, Abernathy CO, Calderon RL (eds) Arsenic exposure and health effects. Elsevier Science Ltd, Oxford, pp 113–121

Coogan TP, Bare RM, Waalkes MP (1992) Cadmium-induced DNA strand damage in cultured liver cells: reduction in cadmium genotoxicity following zinc pretreatment. Toxicol Appl Pharmacol 113(2):227–233

Crump KS (1984) A new method for establishing allowable daily intakes. Fundam Appl Toxicol 4:860–866

Dean HT (1936) Chronic endemic dental fluorosis (Mottled Enamel). JAMA 107:1269–1272

Derryberry OM, Bartholomew MD, Fleming RBL (1963) Fluoride exposure and worker health – the health status of workers in a fertilizer manufacturing plant in relation to fluoride exposure. Arch Environ Health 6:503–514

Doll R, Peto R (1981) The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst 66(6):1192–1308

Fowles JR, Dodge D, Alexeeff GV (1999) The use of benchmark dose methodology with acute inhalation lethality data. Reg Toxicol Pharmacol 29:262–278

Gochfeld M (1997) Factors influencing susceptibility to metals. Environ Health Perspect 105(suppl 4):817–822

Gochfeld M, Syers K (2001) Scientific committee on problems of the environment. http://www.icsu-scope.org/index.html

Hartung R (1987) Dose–response relationships. In: Tardiff RG, Rodricks JU (eds) Toxic substances and human risk. Plenum Press, New York, pp 29–46

Haseman JK, Hailey JR, Morris RW (1998) Spontaneous neoplasm incidences in Fischer 344 rats and B6C3F1 mice in two-Year carcinogenicity studies: a national toxicology program update. Toxicol Pathol 26(3):428–441

Hattis D (1996) Variability in susceptibility–How big, how often, for what responses to what agents? Environ Toxicol Pharmacol 2:135–145

Hill AB (1965) The environment and disease: association or causation? Proc R Soc Med 58:295–300

Hoffman G (1994) Genetic toxicology. In: Klaassen C (ed) Cassarett and Doull's toxicology: the basic science of poisons, 5th edn. McGraw-Hill, New York

Hopenhayn-Rich C, Smith AH, Goeden HM (1993) Human studies do not support the methylation threshold hypothesis for the toxicity of inorganic arsenic. Environ Res 60(2):161–177

Hsueh YM, Chiou HY, Huang YL, Wu WL, Huang CC, Yang MH, Lue LC, Chen GS, Chen CJ (1997) Serum beta-carotene level, arsenic methylation capability, and incidence of skin cancer. Cancer Epidemiol Biomarkers Prev 6(8):589–596

Hsueh YM, Wu WL, Huang YL, Chiou HY, Tseng CH, Chen CJ (1998) Low serum carotene level and increased risk of ischemic heart disease related to long-term arsenic exposure. Atherosclerosis 141:249–257

International Agency for Research on Cancer (IARC) (1999) Sacchurin and its salts. IARC Monograph 73. WHO, Geneva

Le C, Lu X, Ma M, Cullen W, Aposhian V (2000) Speciation of key arsenic metabolic intermediates in human urine. Anal Chem 72:5172–5177

Mengheri E, Murgia C, Vignolini F, Nobili F, Gaetani S (1993) Metallothionein gene is expressed in developing rat intestine and is induced by zinc but not by corticosteroids. J Nutr 123 (5):817–822

Moolgavkar SH, Luebeck EG (1995) Incorporating cell proliferation kinetics into models for cancer risk assessment. Toxicology 102 (1–2):141–147

Mutti A, Mazzucchi A, Rusticelli P, Frigeri G, Arfini G, Franchini I (1984) Exposure-effect and exposure-response relationships between occupational exposure to styrene and neuropsychological functions. Am J Ind Med 5:275–286

National Center for Health Statistics (1974) Decayed, missing, and filled teeth among youth 12–17 years. U. S. Department of Health, Education, and Welfare, Public Health Service, Health Resources Administration, Vital and Health Statistics, vol 11, no 144, DHEW publication no. (HRA) 75–1626, Rockville, MD

National Center for Health Statistics (1981) Decayed, missing, and filled teeth among persons 1–74 years. U. S. Department of Health and Human Services, Public Health Service, Office of Health Research, Statistics, and Technology, Vital and Health Statistics, vol 11, no 223, DHHS publication no. (PHS) 81–1673, Hyattsville, MD

National Institute of Dental Research (1989) Oral health of United States children: the national survey of dental caries in U. S. school children, 1986–1987. U. S. Department of Health and Human Services, Public Health Service, National Institutes of Health, NIH publication no. 89–2247, Bethesda, MD

National Research Council (NRC (1993) Guidelines for developing community emergency exposure levels for hazardous substances. National Academy Press, Washington, DC

Office of Environmental Health Hazard Assessment (OEHHA) (2000) Air toxics hot spots program risk assessment guidelines; Part III: Non-cancer chronic reference exposure levels. California Environmental Protection Agency. www.OEHHA.ca.gov. February 2000

Pizarro F, Olivares M, Araya M, Gidi V, Uauy R (2001) Gastrointestinal effects associated with soluble and insoluble copper in drinking water. Environ Health Perspect 109(9):949–952

Rabovsky J, Fowles J, Hill M, Lewis D (2001) A health risk benchmark for the neurological effects of styrene: comparison with the NOAEL/LOAEL approach. Risk Anal 21(1):117–126

Rudel R, Slayton TM, Beck BD (1996) Implications of arsenic genotoxicity for dose response of carcinogenic effects. Reg Toxicol Pharmacol 23(2):87–105

Schwarz MA, Lazo JS, Yalowich JC, Allen WP, Whitmore M, Bergonia HA, Tzeng E, Billier TR, Robbins PD, Lancaster JR Jr (1995) Metallothiouein protects against the cytotoxic and DNA-damaging effects of nitric oxide. Proc Natl Acad Sci 92 (10):4452–4456

Singh, Rana (2002) Influence of antioxidants on metallothionein-mediated protection in cadmium-fed rats. Biol Trace Elem Res 88 (1):71–77

Tang EM (2000) Epigenetic effects on susceptibility to heavy metal and PAH induced DNA damage. Crisp Data Base National Institutes of Health, Bethesda, MD

Tseng CH (1999) Chronic arsenic intoxication in Asia: current perspectives. J Intern Med Taiwan 10:224–229

Tseng CH (2002) An overview on peripheral vascular disease in blackfoot disease-hyperendemic villages in Taiwan. Angiology 53 (5):529–537

Tseng CH, Chen CJ, Lin BJ, Tai TY (1994) Abnormal response of ankle pressure after exercise in seemingly normal subjects living in blackfoot disease-Hyperendemic villages in Taiwan. Vasc Surg 28:607–617

Tseng CH, Tai TY, Lin BJ, Chen CJ (1995) Abnormal peripheral microcirculation in seemingly normal subjects living in blackfoot disease-hyperendemic villages in Taiwan. Int J Microcirc 15:21–27

Tseng CH, Chong CK, Chen CJ, Tai TY (1996) Dose–response relationship between peripheral vascular disease and ingested inorganic arsenic among residents in blackfoot disease endemic villages in Taiwan. Atherosclerosis 120:125–133

Tseng CH, Tai TY, Chong CK, Tseng CP, Lai MS, Lin BJ, Chiou HY, Hsueh YM, Hsu KH, Chen CJ (2000) Long-term arsenic exposure and incidence of non-insulin-dependent diabetes mellitus: a Cohort study in arseniasis-hyperendemic villages in Taiwan. Environ Health Perspect 108:847–851

Tseng CH, Tseng CP, Chiou HY, Hsueh YM, Chong CK, Chen CJ (2002) Epidemiologic evidence of diabetogenic effect of arsenic. Toxicol Lett 133:69–76

Turner SD, Tinwell H, Piegorsch W, Schmezer P, Ashby J (2001) The male rat carcinogens limonene and sodium saccharin are not mutagenic to male Big Blue rats. Mutagenesis 16(4):329–332

United States Environmental Protection Agency (U. S. EPA) (2000) Exposure and human health reassessment of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and related compounds, EPA/600/P-00/001Ag.

United States Environmental Protection Agency (U.S. EPA) (2002) A review of the reference dose and reference concentration processes. EPA/630/P-02/002F. www.epa.gov/iris/Rfd.final.pdf

WHO (1994) Fluorides and oral health. WHO technical report series 846. WHO, Geneva

Yamamoto E, Shimada A, Morita T, Yasutake A, Yoshide M, Nishimura N, Suzuki JS, Satok M, Tokyama C (2001) Localization and role of placental metallothioueins as a barrier against maternal to fetal transfer of mercury. Toxicologist 60(1):357

Environmental Pathology

25

José A. Centeno, Florabel G. Mullick, Kamal G. Ishak†, Teri J. Franks, Allen P. Burke, Michael N. Koss, Daniel P. Perl, Paul B. Tchounwou, and Joseph P. Pestaner

Contents

25.1 Introduction

Humans are constantly exposed to hazardous pollutants in the environment—for example, in the air, water, soil, rocks, diet, or workplace. Trace metals are important in environmental pathology because of the wide range of toxic reactions and their potential adverse effects on the physiological function of organ systems. Exposures to toxic trace metals have been the subject of numerous environmental and geochemical investigations, and many studies have been published on the acute and/or chronic effects of high-level exposures to these types of agents; however, much fewer data are available concerning the health effects of low-dose chronic exposure to many trace metals. Chronic low-dose exposures to toxic elements such as cadmium and arsenic have been shown to cause these metals to accumulate in tissues over time, leading to multiple adverse effects in exposed individuals.

Exposure to toxic trace metals occurs via three principal routes: percutaneous absorption, ingestion, or inhalation. The toxic effects may affect specific target organ components, resulting in immunological-induced injury or specific functional changes. The diseases caused by metals can be genetic or acquired, and the effects can be acute or chronic. This chapter provides a review of some of these pathologies and discusses the critical organ systems that are affected. Examining such toxicities is a medical challenge in that a number of metallic elements, such as iron, copper, and manganese, are essential to life. Distinguishing normal and pathologic states is critical to our understanding of the pathogenesis of metal-induced diseases. The toxic properties of

†Deceased

J.A. Centeno (✉)
Joint Pathology Center Division of Biophysical Toxicology, Malcolm Grow Medical Clinic Joint Base Andrews Air Naval Facility, Washington, DC, USA
e-mail: jose.a.centeno@us.army.mil

F.G. Mullick • T.J. Franks • A.P. Burke
The Joint Pathology Center Silver Spring, MD, USA

K.G. Ishak†

M.N. Koss
Medical Laboratories, University of Southern California Medical Center, Los Angeles, CA, USA

D.P. Perl
Department of Neurology, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

P.B. Tchounwou
College of Science, Engineering and Technology, Jackson State University, Jackson, MS, USA

J.P. Pestaner
Medical Examiner Office, Washington, DC, USA

This chapter is dedicated to the memory of Kamal G. Ishak, M.D. Ph.D. (1928–2004).

The opinions and assertions expressed herein are those of the authors and are not to be construed as official or representing the views of The Joint Pathology Center, the US Department of the Army or the Department of Defense

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_25,

certain metals, such as lead and mercury, have been acknowledged since ancient times, but enhanced pathologic analyses have allowed us to learn much about how metals can affect specific organ systems. Also reviewed in this chapter are the pathologic states caused by metals in the skin, brain, lung, heart, and liver.

25.2 The Skin

The list of metals exhibiting dermal toxicity has been well catalogued. Such metals include compounds used in medicinal products, industrial processes, pesticides, cosmetics, dyes, and jewelry (Lansdown 1995). Of major concern is exposure to metals and metalloids through contaminated water and other environmental and geological media. Dermal toxicity is a result of local tissue responses to direct contact of a metal with skin or, alternatively, it may represent a manifestation of systemic toxicity following ingestion or inhalation. Allergic contact dermatitis induced by nickel is one such example of a local tissue response. The adverse cutaneous reactions resulting from chronic ingestion or inhalation of arsenical compounds exemplify systemic toxicity.

A variety of pathologic responses in the skin are associated with both acute and chronic exposures to metals. Categorization of these responses presents a challenge to the environmental pathologist, as the histologic features associated with metal-induced skin lesions may mimic virtually any known morphologic skin disease. The more frequently encountered morphologic changes include spongiotic dermatitis (allergic contact dermatitis and primary irritant dermatitis), granulomatous inflammation, pigmentation disorders, and cancer. A pertinent exposure history correlated with pathologic findings should be done to establish a precise diagnosis.

The occurrence of hyper- and hypopigmentation of the skin has been reported worldwide in populations chronically exposed to arsenic from contaminated drinking water (see also Chap. 12, this volume). Because of its widespread presence in the environment, arsenic has become one of the most studied elements in environmental toxicology and public health. The ensuing discussion centers on arsenic and arsenic-related skin diseases, which are considered by many to be the prototype of the development of disease following exposure to a metal.

25.2.1 Arsenic and Metal-Induced Cancer of the Skin

The most widely recognized toxic element affecting the skin is arsenic. Arsenic is the twentieth most abundant element in the Earth's crust. It is odorless and tasteless and exhibits both acute and chronic health effects in humans. In nature, arsenic can occur as metalloid alloys or in a variety of chemical compounds. In geological media such as rocks, arsenic is commonly found as a sulfide such as orpiment (As_2S_3) or realgar (As_2S_2) in the form of arseno–pyrite or mixed sulfides (AsFeS). Significant amounts of arsenic may also be found bound to gold, silver, copper, lead, zinc, and cobalt ores. Mining of these minerals may result in the mobilization and/or transport of arsenic into drinking water. Arsenic has also been used in a variety of agricultural applications (*e.g.*, pesticides, insecticides), industrial applications (*e.g.*, manufacturing of solid-state detectors), and medical applications (*e.g.*, drugs and medical treatments).

As with all toxic metals, the toxic effects of arsenic are related to the chemical and physical forms in which it appears: metallic, As(0); inorganic, As(III) and As(V); and organic, As(III) and As(V). Although arsenic exhibits both organic and inorganic forms, the inorganic trivalent arsenic compounds are considered to demonstrate the greatest toxicity. The molecular basis by which arsenic compounds may induce their toxicity in humans has been described. Impairment of cellular respiration through inhibition of various mitochondrial enzymes and uncoupling of oxidative phosphorylation is one of the major mechanisms by which arsenic exerts its toxic effects. At the molecular level, the toxicity of arsenic results from its ability to interact with sulfhydryl groups of proteins and enzymes and to substitute phosphorus in a variety of biochemical reactions (Li and Rossman 1989). *In vitro* experiments have demonstrated that arsenic reacts with protein sulfhydryl groups to inactivate enzymes such as dihydrolipoyl dehydrogenase and thiolase, thereby producing inhibited oxidation of pyruvate and beta-oxidation of fatty acids (Belton et al. 1985).

Humans are exposed to inorganic arsenic mainly through the oral and inhalation routes. Direct dermal exposure also occurs, but to a lesser extent. The oral route includes contaminated drinking water, food, drugs (including Chinese herbal medications), and tobacco. Inhalation occurs primarily in occupational settings; workers may be exposed to arsenic in the air as a by-product of copper and lead smelting, pesticide production, manufacturing of glass, and production of semiconductors (Chan and Huff 1997). Arsenic tends to concentrate in ectodermal tissues, including the skin, hair, and nails. Biomethylation is considered the major metabolic pathway for inorganic arsenic in humans. Historically, the enzymatic conversion of inorganic arsenic to mono- and dimethylated species has been considered a primary detoxification mechanism of inorganic arsenic; however, compelling experimental evidence obtained from several laboratories suggests that biomethylation, particularly the production of methylated metabolites that contain trivalent arsenic, is a process that can activate arsenic as a toxin and a carcinogen (Styblo et al. 2002; Wei et al. 2002).

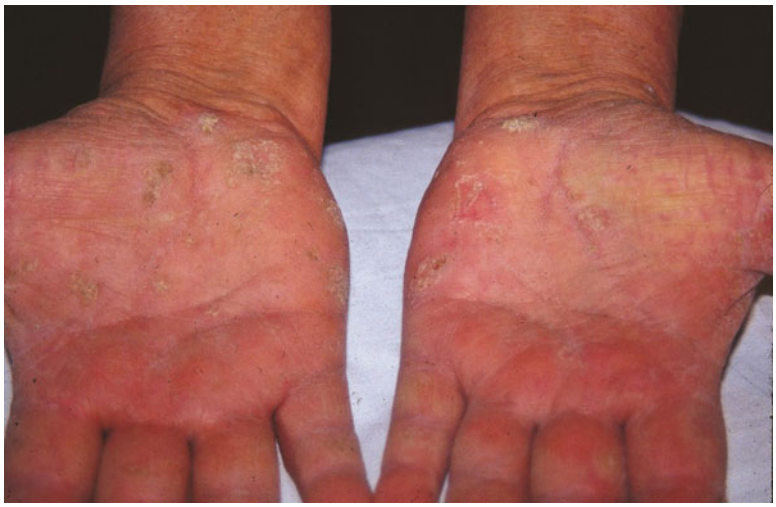


Fig. 25.1 Arsenic-induced hyperkeratosis of the hands

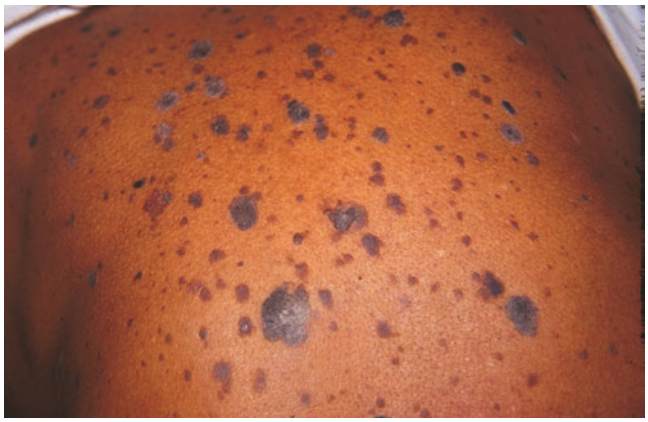


Fig. 25.2 Hyperpigmentation and hyperkeratosis lesions of the *back* induced by chronic exposure to arsenic from consumption of contaminated drinking water

Epidemiological studies have confirmed the role of arsenic in the induction of cancers of the skin. Of the metals known to exhibit dermal toxicity, only arsenic has been shown conclusively to be carcinogenic (Chen et al. 1992). Squamous cell carcinomas *in situ* (Bowen's disease) and basal cell carcinomas of the skin have been associated with chronic inorganic arsenic ingestion with a latency period of 2–20 years after exposure (Maloney 1996; Tsai et al. 1999). In addition, epidemiological studies have provided suggestive evidence linking arsenic exposure to various internal cancers, including angiosarcoma of the liver, lung cancer, and bladder cancer. In the majority of these cases, in which the internal cancer is ascribed to arsenic exposure, some dermatologic hallmark of arsenic poisoning (such as hyper- or hypopigmentation) is identified.

Arsenical keratosis is a well-established clinical entity resulting from chronic exposure. The lesions are usually most pronounced on the palms and soles, although they can occur on the trunk and other areas of the extremities (see Fig. 25.1). Arsenical keratoses are characterized by several specific pathologic features, including hyperkeratosis, parakeratosis, and acanthosis (see Fig. 25.2). Nuclear atypia is sometimes present as well. In severe atypia, squamous cells exhibit hyperchromatic nuclei and a disorderly arrangement within the epithelium. Within the spectrum of keratotic lesions, arsenical keratosis may be differentiated from the more commonly diagnosed actinic keratosis by the absence of epidermal atrophy and basophilic degeneration of the upper dermis. All arsenical skin changes, including keratoses, tend to occur in non-exposed sites with an absence of dermal solar elastosis noted histologically.

As a human carcinogen, inorganic arsenic remains an enigma because arsenic-related tumors occur in humans but not in laboratory animals. The biological mechanisms by which arsenic induces chronic effects including cancer are not well understood and are the subject of considerable research efforts. Arsenic does not directly damage DNA but rather causes chromosome aberrations, aneuploidy, cell transformation, and gene amplification in many cell types. Recently, arsenic-induced carcinogenesis has been studied using molecular biological methods. Among the oncogenes evaluated, the tumor-suppressor gene *p53* appears to play a role in arsenic-induced carcinogenesis. In one study from an endemic area, 48 cases representing a variety of arsenic-induced skin cancers (including Bowen's disease, squamous cell carcinoma, and basal cell carcinoma) were examined (Chung et al. 1998). All of the specimens demonstrated positive *p53* immunostaining. Positive *p53* staining was identified in all perilesional normal skin as well, which suggested that the *p53* mutation may be an early event in arsenic-related carcinogenesis. Another study demonstrated that *p53* mutation rates, sites, and types in arsenic-related skin cancer are significantly different from those in ultraviolet-induced cancer, which implicates arsenic as the etiologic agent and suggests a possible mechanism of action (Hsu et al. 1999).

In the absence of animal models, *in vitro* studies become particularly important in providing information on the carcinogenic mechanisms of arsenic toxicity. Arsenic and arsenical compounds have been reported to induce morphological changes in cultured cells. Experimental studies have indicated that arsenic is cytotoxic and able to transcriptionally activate a significant number of stress genes in transformed human liver cells (Tchounwou et al. 2000, 2001). Arsenic and arsenic-containing compounds have also been shown to be potent clastogens both *in vivo* and *in vitro*. Arsenical compounds have been reported to induce sister chromatid exchanges and chromosome aberrations in both human and rodent cells in culture (Nakamuro and Sayato 1981; Barrett et al. 1989). Arsenical compounds have also been shown to induce gene amplification, arrest cells in mitosis, inhibit DNA repair, and induce expression of the *c-fos* gene and the oxidative stress protein heme oxygenase in mammalian cells (Ramirez et al. 1997; Jingbo et al. 2002), and they have been implicated as promoters and comutagens for a variety of agents (Cavigelli et al. 1996).

Research has also indicated that inorganic arsenic does not act through classic genotoxic and mutagenic mechanisms but may be a tumor promoter that modi-fies signal transduction pathways involved in cell growth and proliferation (Simeonova and Luster 2000; Kitchin 2001). Modulation of gene expression and/or DNA-binding activities of several key transcription factors, including nuclear factor kappa B (NFκB), tumor-suppressor protein (p53), and activating protein-1 (AP-1) has been associated with arsenic exposure (Barchowsky et al. 1996). Mechanisms of AP-1 activation by trivalent arsenic include stimulation of the mitogen-activated protein kinase (MAPK) cascade with a consequent increase in the expression and/or phosphorylation of the two major AP-1 constituents, c-jun and c-fos (Simeonova and Luster 2000).

Much remains to be learned about arsenic and its toxicity to skin, and similar challenges remain when examining arsenic and its toxicity to other important organ systems in the body. As is discussed in the next section, arsenic damage to the nervous system is relatively acute, and the extent of damage is proportional to the exposure dose (see also Chaps. 12 and 22, this volume).

25.3 The Brain

25.3.1 Introduction

It should be recognized that a number of metallic elements, such as iron, copper, and manganese, are essential to life and play an important role in the functioning of the central nervous system (CNS); nevertheless, that certain metals, such as lead and mercury, have neurotoxic properties has been acknowledged since ancient times. The brain must be viewed in a somewhat different light than other organ systems because of its position behind the protective shield of the blood–brain barrier. Because of this critical protective barrier, one must recognize that, for a metal to induce neurologic damage, it must be able to cross the blood–brain barrier, enter the CNS, and gain access to target cells in sufficient quantity to produce pathologic damage. Compounds in which the metal is linked to a lipophilic organic compound tend to be particularly neurotoxic as they can readily cross the lipid membranes that comprise the blood–brain barrier. One example is mercury, which, in its inorganic form, is relatively nontoxic to the CNS; however, when mercury is methylated to form methylmercury, the compound rapidly crosses the blood–brain barrier, is readily taken up by neurons, and produces massive cellular destruction. This results in the severe parenchymal damage that occurred in the outbreak of severe CNS damage in Minamata, Japan (see discussion below). Finally, it should be recognized that the effects of toxins that damage neurons are particularly serious because of the inability of these cells to regenerate.

Certain metals function as classic toxins to the nervous system, in which damage is relatively acute following exposure, and the extent of damage is proportional to the exposure dose. One such example is arsenic. As noted earlier, arsenic is generally found as an impurity of ores containing copper, lead, gold, and zinc. Exposure to arsenic is a relatively rare event but does occur due to accidental ingestion, suicide, or murder and most commonly is related to occupational activities and consumption of contaminated drinking water (see Chap. 12, this volume). Arsenic is an active ingredient of herbicides, insecticides, and other pesticides. In cases of acute arsenic poisoning, sudden fatal circulatory collapse may follow ingestion of a single large dose. When smaller doses are involved, gastrointestinal symptoms will predominate initially to be followed after a period of 2–3 weeks by the development of a rapidly progressive peripheral neuropathy. The extent and severity of the neurologic symptomatology will depend on the arsenic dosage. In cases of chronic exposure to lower doses of arsenic, a sensory neuropathy is generally observed. Relatively little is known of the mechanism by which arsenic produces its toxic damage, but recent studies suggest that the process involves primary axonal damage.

25.3.2 Lead Poisoning: Acute and Chronic Effects

Lead is perhaps the most important metallic neurotoxin. Adults are relatively resistant to its effects, and only in high doses is a peripheral neuropathy encountered. In children, however, the effects of relatively low doses can be much

more devastating. High doses in children can cause acute lead encephalopathy, a life-threatening condition characterized by generalized cerebral edema with increased intracranial pressure, which leads to transtentorial and cerebellar tonsillar herniation. Clinical features of acute lead encephalopathy may include ataxia, seizures, stupor, coma, and often death. Almost always other associated systemic signs of lead exposure, such as anemia and the presence of lead lines on x-rays of the long bones, are noted in affected children. At autopsy, the brain is markedly swollen with compressed gyri, obliterated sulci, and collapsed lateral ventricles. Uncal and cerebellar tonsillar herniation are commonly encountered. Microscopically, there is a breakdown of the blood–brain barrier with transudation of fluid into the pericapillary space and ischemic damage to cerebral cortical neurons. Once again, the adult is apparently resistant to such toxic changes, and it is extremely rare to see acute encephalopathy in an adult even following extremely high-exposure doses. More typically, adults exposed to lead develop a peripheral neuropathy with wrist drop and/or ankle drop. Exposure in children leading to encephalopathy is primarily through the eating of peeling paint chips in houses dating from an era when lead-based paints were employed. In general, exposure to lead can come from the air, in the form of lead fumes (*e.g.*, firing range operators, stained glass workers, solderers); from water contaminated by lead plumbing components (particularly in acid conditions); the use of lead-containing vessels (lead crystal, lead glazes of pottery in which acid liquids have been stored); and airborne particulates, such as is found in restorers of old homes and bridge workers. Other workers who are particularly prone to high lead exposures are battery production workers and bronze workers.

Largely through the work of Needleman and colleagues (1979; 1988; 1990), the long-term effects of lower doses of lead exposure on children have been increasingly recognized. This relates to the adverse effects of lead on intellectual functioning as well as its association with behavioral problems in children exposed to what had been previously considered to be relatively low lead burdens. Documenting low-level, chronic lead exposure has not been easy to accomplish. It is clear that determining a single blood lead level in a child can be completely misleading. In the initial studies of Needleman et al., lead exposure was estimated through the calculation of lead levels in the dentine of deciduous teeth. These studies suggested that exposure to relatively low levels of lead in childhood is associated with a slight, but significant, drop in IQ. Furthermore, such low-level exposures have been associated with a tendency toward disturbed classroom behavior. Follow-up studies on children with elevated lead levels showed a markedly high dropout level from high school, increased absenteeism, and impaired eye–hand coordination. The neuropathologic substrate of all such effects of low-level, chronic lead exposure remains unclear, but these data have resulted in a progressive lowering of what is considered an "acceptable" blood lead level.

25.3.3 Mercury Poisoning: Inorganic Versus Organic Forms

The hazards of inorganic mercury poisoning have been known since Roman days, and the dangers of working in the mercury mines of Almedon, Spain, have been widely acknowledged for centuries. Central nervous system symptoms related to such exposure are generally related to the presence of high concentrations of inorganic mercury fumes. These produce tremor and irritability which typically are reversible upon removal from the exposure source. The behavioral disturbances that result gave rise to the notion of being "mad as a hatter," a saying that refers to the practice of employing mercury rather indiscriminately in the felt tanning industry of Victorian England and the subsequent frequency of cases of inorganic mercury poisoning among hatters. Nevertheless, oral mercury-containing calomel medicinals have been used relatively safely over the centuries, and inorganic mercury has a relatively low level of toxicity. On the other hand, exposure to organic mercury, such as methylmercury, causes dramatic nervous system destruction (Hunter and Russell 1954).

The importance of this distinction was dramatically demonstrated in the tragic outbreak of methylmercury poisoning among Japanese villagers living along Minamata Bay (Marsh 1979). In the small town of Minamata, Japan, on the southern portion of the island of Kyushu, a large chemical factory employed a significant amount of inorganic mercury as a catalyst for the production of the raw ingredients of plastic. Some of this mercury was discharged into Minamata Bay. The mercury deposited in the sediment of the bay was subsequently taken up by certain bacteria, which are capable of methylating it to form methylmercury (Jensen and Jernelov 1969). These bacteria entered the food chain to eventually deposit methylmercury in the fish, the major protein source of the local villagers.

The methylated form of mercury readily crosses the blood–brain barrier, inducing severe neuronal destruction in extremely small doses. Indeed, the solubility and neurotoxicity of methylmercury are so great that fatal poisoning has been reported in a research chemist who was exposed to minute amounts of dimethylmercury that gained access to his skin directly through his latex gloves. Clinically, exposed patients initially show a constricted visual field, paresthesias, gait ataxia, and impairment of vibration sense (stereognosis) and two-point discrimination. Subsequently, more profound visual loss (frequently progressing to total blindness), ataxia, and sensory–motor signs are noted. The neuropathologic

lesions of methylmercury poisoning are quite distinct, with acute necrosis of the calcarine and pre-central cortex and the cerebellum (Shiraki 1979). Nerve cell loss is severe, with a profound glial response. In the cerebellum, the neuronal loss predominates in the internal granular layer, with sparing of the Purkinje cells. Although blindness is a common complication of the disease, the retina appears to be intact, and the loss of vision is thought to be directly related to destruction of the primary visual (calcarine) cortex. In patients with prolonged survival after exposure to methylmercury, dramatic atrophy and glial scarring of the abovementioned regions are observed. Among the most severely affected victims of the Minamata outbreak were newborns who were exposed *in utero*. The methylmercury readily crossed the placenta and produced dramatic widespread nervous system destruction.

25.3.4 Tin Poisoning: An Example of Selective Toxicity of Organic Metal Complexes

A phenomenon similar to that of enhanced neurotoxicity of methylmercury also occurs following exposure to tin-containing compounds. Metallic tin and its inorganic salts have been included in medicinal preparations since the sixteenth century. These were widely administered in relatively high doses without any apparent adverse health effects; however, organic tin compounds, similar to organic mercury compounds, are highly lipid soluble and are also highly toxic to the nervous system (Cavanagh and Nolan 1994). Unfortunately, these characteristics were learned following a serious outbreak of organic tin poisoning in 1954 in France due to the introduction of the drug Stalinon for the treatment of staphylococcal infections. Each Stalinon capsule contained 15 mg of diiodide–diethyltin. Of those exposed to this drug, approximately half died, one-third recovered, and the remainder suffered from a wide variety of chronic neurologic deficits. Autopsies on the acute fatal cases showed evidence of an unusual form of diffuse edema that was confined to the white matter. Subsequent experimental work has shown that triethyltin produces a characteristic and selective white matter edema, presumably through its toxic effects on the functioning of oligodendroglial membranes. Interestingly, animals exposed to trimethyl tin fail to show evidence of white matter edema but instead develop a selective necrosis of neurons of the hippocampus. The mechanism behind this dramatic difference in the selective toxicity of these two extremely similar compounds remains unclear; nevertheless, the organic moiety on each form allows for its rapid uptake into the CNS, thus providing access to the different cellular targets for its toxicity.

25.3.5 Manganese-Induced Parkinsonism

Manganese is a neurotoxic metal that is capable of readily entering the CNS when it reaches the general circulation. Upon reaching the brain, manganese accumulates selectively in the globus pallidus, where it can destroy local nerve cells. Manganese is the twelfth most abundant element in the Earth's crust and the fourth most widely used metal in the world. Eight million tons of manganese metal are extracted annually, of which 94% is employed in the manufacture of steel. Manganese is also used in the manufacture of batteries. Potassium permanganate is a widely employed bactericidal and fungicidal agent in water purification processes. Additionally, methylcyclopentadienyl manganese tricarbonyl (MMT) is an organic manganese compound that has been used as an anti-knock additive to gasoline. Medically significant manganese toxicity is almost exclusively encountered in an industrial setting, either in association with manganese mining or in smelting operations. The classic papers of Mena and coworkers (Mena et al. 1967; Mena 1979) documented the development of psychiatric and parkinsonian features in a group of manganese miners in Chile, and more recent papers from Taiwan have documented cases seen in association with smelting (Olanow et al. 1994).

The clinical features of manganese neurotoxicity closely resemble Parkinson's disease but have more prominentdystonic features. Initial stages typically include psychiatric disturbances, such as behavioral abnormalities, hallucinations, and, at times, frank psychosis. This syndrome is referred to as "manganese madness" or *locura manganica*. Not all cases pass through the psychiatric phase of the disease and present with the extrapyramidal manifestations. The extrapyramidal features consist of bradykinesia, gait disturbance, postural instability, rigidity, micrographia, masked facies, and speech disturbances. Although tremors may be seen, they usually are not the "pill-rolling" tremors seen at rest that are typical of Parkinson's disease but are described as more of an intension tremor. Dystonic features can be rather prominent with facial grimacing and plantar flexion of the foot, which results in a very characteristic gait disturbance. The gait of manganese poisoning consists of the patient keeping the foot dorsoflexed and walking with elbows flexed and spine erect. This rather unique and characteristic gait is commonly referred to as *coq au pied* or "cock walk," as it resembles the strutting of a rooster.

The manganese that produces this form of neurotoxicity is generally thought to enter the body through the lungs. Experimental studies have shown that airborne manganese is readily taken up into the systemic circulation and is selectively deposited in the globus pallidus of the brain, where the major site of damage occurs (Newland et al. 1989; Olanow et al. 1996). Pathological studies of both

experimental models and human cases demonstrate evidence of selective neuronal loss and gliosis in globus pallidus (primarily the medial segment) and striatum. Accompanying the neuronal loss is a dramatic increase in the amount of stainable iron within the damaged regions (Olanow et al. 1996). Whether the increased iron participates in the neuronal damage through oxidative damage or is secondary to the parenchymal destruction remains unclear. The substantia nigra pars compacta is the location of the dopaminergic cells that project to the striatum and are the main targets of damage in Parkinson's disease, but they remain intact in manganese poisoning. Lewy bodies, the characteristic microscopic feature of Parkinson's disease, are not encountered in cases of manganese-induced parkinsonism. Presumably dysfunction of the extrapyramidal system occurs secondary to damage to the globus pallidus and striatum, the major targets of the striatonigral projections. Because the postsynaptic targets of dopaminergic transmission are largely lost, treatment by levodopa is generally considered to be ineffective in manganese-induced parkinsonism. This represents a major defining characteristic that distinguishes manganese-induced parkinsonism from Parkinson's disease.

25.3.6 Metals and Age-Related Neurodegenerative Diseases

The accumulation of metals in the brain has also been associated with several of the age-related neurodegenerative disorders. The two metals most often cited are aluminum and iron. Excess amounts of both metals have been identified in microprobe studies within the characteristic neurofibrillary tangles of cases of Alzheimer's disease (Perl and Brody 1980; Good et al. 1992a). Similarly, increased concentrations of iron and aluminum have been noted in the neurons of the substantia nigra in cases of Parkinson's disease (Good et al. 1992b). It is not considered likely that such elemental accumulations play an etiologic role in these diseases, although they may certainly have the capacity to contribute to their pathogenesis (Perl and Good 1992). Iron, through its ability to donate electrons in energy transfer reactions, may actively support oxidative damage. Evidence for oxidative damage in the brains of Alzheimer's disease and Parkinson's disease victims is well established (Markesbery 1997). Aluminum is a highly charged element that binds strongly to proteins and acts as a cross-linking stabilizer. Many of the proteins comprising the intraneuronal inclusions that characterize the neurodegenerative diseases (the neurofibrillary tangle being the most prominent example) are highly cross-linked and thus resistant to degradation. Whether aluminum serves this role in the natural history of the disorder remains unclear, although recent evidence appears to support this concept. These findings have raised concerns regarding the source of such elemental accumulations and whether aspects of environmental and geological exposures might play a role in the further onset of these age-related disorders. These questions remain largely unresolved.

One question that is resolved is the critical role oxygen plays in maintaining the tissues in the brain. Effects of metals on the lungs can secondarily affect respiratory effectiveness, which will have secondary effects on the brain. Toxicity of metals and the lung are reviewed in the next section.

25.4 Inhalation Injury

25.4.1 Introduction

Zenker (1867) first proposed the term "pneumonokoniosis" as a name for lung disease resulting from the inhalation of dust. While many forms of pneumoconioses are known today, each with its own etiologic agent, the definition of pneumoconiosis varies. Some authors restrict the term to non-neoplastic reactions of the lung to inhaled minerals or organic dusts, excluding asthma, bronchitis, and emphysema. Others use the term more broadly to define the accumulation of abnormal amounts of dust and the resulting pathologic reactions (Gibbs 1996). These reactions range from minimal responses to inert dust particles, such as interstitial dust macules, to lethal scarring associated with fibrogenic dusts (see also Chaps. 10 and 19, this volume). Damage caused by inhaled particles depends on a variety of factors, including the number, size, and physiochemical properties of the particles; the deposition, clearance, and retention of particles in the respiratory tract; the host's inflammatory response to the inhaled particles; and the duration of exposure and interval since initial exposure, as well as interactions with other inhaled particles, particularly cigarette smoke (Roggli and Shelburne 1994).

25.4.2 Deposition, Clearance, and Retention

The respiratory system acts like a filter by removing particles and bacteria from inhaled air and leaving exhaled air essentially free of contaminants (Brain and Valberg 1979). With each breath, air is drawn through the nares into the nasopharynx and trachea and on into the conducting system of the lungs. Air then passes by way of the bronchi, terminal bronchioles, respiratory bronchioles, and alveolar ducts to the alveoli. In healthy individuals, the tracheobronchial tree is covered by a thin, watery layer of mucus, which is continuously moved up and out of the lungs by ciliated bronchial epithelium. The velocity of inspired air decreases

over the course of the airways. Flow rates decrease slightly in the trachea and are markedly reduced in the third through fifth generations of bronchi. At the level of the terminal bronchiole, flow rates are not more than 2–3 cm per second (Morgan and Seaton 1975).

Particles that are nearly spherical in shape are called *compact particles*, whereas those with a length-to-width ratio of 3:1 or greater are called *fibers*. Compact particles with an aerodynamic equivalent diameter of 1–5 μm are the most likely to deposit in the lung parenchyma (Roggli and Shelburne 1994; Gibbs 1996), whereas particles with extreme shapes, such as fibers and plates, tend to deposit in the airway walls (Gibbs 1996). Fiber deposition is primarily a function of diameter and less a function of length (Gibbs 1996). Fibers with diameters below 3 μm, even when several hundred micrometers in length, can align axially with the airstream (Timbrell et al. 1970) and can deposit in the alveoli; however, the longer the fiber, the less its statistical chance of reaching the alveolus (Churg 1998a).

Deposition refers to the fraction of particles in inspired air that are trapped in the lung and fail to exit with expired air. Deposition of particles occurs in the respiratory tract as a consequence of the physical processes of inertial impaction, sedimentation, and diffusion. Inertial impaction occurs when the air current carrying a particle changes direction and the momentum of the particle carries it along its original path. This frequently results in particle deposition at bifurcation points in the respiratory tract (Brain and Valberg 1979). Sedimentation of particles occurs secondary to the force of gravity, whereas very small particles are deposited by diffusion and Brownian motion. In general, a high percentage of particles greater than 1 μm in diameter is filtered in the nasopharynx by inertial impaction and sedimentation. A smaller percentage of particles is deposited by the same mechanisms in the trachea and bronchi. When particles are under 0.5 μm in diameter, they deposit in the alveoli by sedimentation and diffusion (Morgan and Seaton 1975). Breathing rate, tidal volume, and nose versus mouth breathing influence the pattern of deposition (Roggli and Shelburne 1994; Gibbs 1996).

Clearance is the output of particles previously deposited in the respiratory tract. Mechanisms such as dissolution and absorption, coughing, sneezing, and tracheobronchial and alveolar transport systems keep the lungs relatively free of foreign material (Brain and Valberg 1979). The tracheobronchial system, also known as the mucociliary escalator, extends from the terminal bronchiole to the glottis. Particles deposited in the conducting system adhere to a thin layer of mucus secreted by the mucous glands and goblet cells of the bronchial tree; subsequently, particles are swept toward the glottis by ciliary action and are swallowed or expectorated. The rate of transport is approximately 3 mm per minute, with 80–90% of particles removed within 2 h. The bronchial mucus is arranged in two layers: an outer viscous gel, which does not absorb water, and an inner liquid solid phase in which the cilia beat. The function and integrity of the inner layer are affected by dehydration. Ciliary function is influenced by ionic charges, the presence of oxidants, high concentrations of oxygen, low concentrations of adenosine triphosphate, and the presence of oxidants and cigarette smoke. In the alveolar transport system, particles are dissolved, removed to lymphatics, engulfed by macrophages, or transported by surface fluid movement to the mucociliary escalator (Morgan and Seaton 1975). Retention is defined as the amount of particulate matter found in the lungs at any time and is dependent on the rates of deposition and clearance.

The diseased and blackened lungs of miners who die of coal worker's pneumoconiosis contain less than 2% of the dust originally deposited in the lungs (Brain and Valberg 1979). This illustrates the efficiency of the respiratory tract as a filter; however, the normal defense mechanisms of the lung can be overwhelmed by high exposure levels to inhaled foreign material, which increases the probability of a pathologic response.

25.4.3 Pathologic Responses to Inhaled Particles

Inhaled particles can produce a variety of injury patterns in the lung, which include diffuse alveolar damage (fume exposure), dust macules with or without small airway fibrosis, diffuse interstitial fibrosis, alveolar proteinosis, granulomatous interstitial pneumonitis, giant-cell and desquamative interstitial pneumonitis, and lung cancer. Generally, these injury patterns lack specific features that implicate a causative agent.

Diffuse alveolar damage (DAD), the histologic correlate to adult respiratory distress syndrome (ARDS), is the most common serious reaction to inhaled gases and fumes (Wright and Churg 1998). Microscopically, DAD is characterized by hyaline membranes lining edematous alveolar septa. Virtually any noxious gas or fume (*e.g.*, beryllium and cobalt metal) inhaled in sufficient concentration can potentially cause DAD (Wright and Churg 1998); however, numerous other agents, including infectious organisms, drugs, ingestants, shock, and sepsis, cause DAD that is microscopically indistinguishable from that produced by gases and fumes.

Dust macules are nonpalpable, peribronchiolar interstitial aggregates of pigmented dust and dust-laden macrophages. Initially, macules may have little associated fibrosis; however, with sufficient exposure, peribronchiolar fibrosis may occur. This pattern of injury generally has little functional deficit, but due to the radiodensity of the dust it is usually

associated with an abnormal chest radiograph (Churg and Colby 1998). Agents that produce dust macules include antimony, barium, chromium ore, iron, rare earths, tin, titanium, and tungsten.

A diffuse interstitial fibrosis pattern of injury, an uncommon complication of metal exposure, can be seen with iron (mild fibrosis), aluminum, hard metal (containing cobalt), copper, rare earths, and silicon carbide exposure. Asbestos and mixed dusts containing silicates can also produce interstitial fibrosis.

Alveolar proteinosis pattern is characterized by a relatively uniform filling of alveoli with granular, eosinophilic exudate containing dense bodies and acicular clefts accompanied by a variable amount of chronic interstitial inflammation and fibrosis. This pattern of injury is usually associated with acute exposure to high levels of silica dust and rarely to aluminum dusts.

Giant-cell (GIP) and desquamative interstitial pneumonitis (DIP) patterns of injury can result from exposure to "hard metals." Patchy interstitial fibrosis with mild chronic inflammatory cell infiltrates, accompanied by striking intraalveolar accumulations of macrophages, characterize these two injury patterns. The presence of enlarged, multinucleated alveolar macrophages, which may contain engulfed inflammatory cells, distinguishes GIP from DIP. GIP is characterized microscopically by interstitial fibrosis accompanied by noncaseating granulomas. This pattern can be seen in chronic beryllium disease.

Asbestos (particularly when associated with asbestosis), arsenic, beryllium, cadmium, chloromethyl ether, hexavalent chromium compounds, nickel, and radon have been linked to lung cancer. Lung cancer occurring in occupationally exposed individuals is histologically indistinguishable from cancer in unexposed individuals (see also Chap. 19, this volume).

Although many forms of pneumoconioses are known, only a handful are seen by the surgical pathologist with any regularity (Table 25.1). The ensuing discussion centers on asbestos and asbestos-related disease, which is considered by many to be the prototype inhalation injury.

Table 25.1

Asbestos-related disease
Benign diseases
Pleural plaques
Diffuse pleural fibrosis
Benign asbestos effusion
Rounded atelectasis
Asbestosis
Malignancies
Lung cancer
Malignant mesothelioma

25.4.4 Asbestos

Asbestos is a group of naturally occurring mineral fibers with the economically useful properties of flexibility, high tensile strength, acoustic insulation, and corrosion, thermal, and electrical resistance. These properties led to the widespread incorporation of asbestos into many products including fireproofing and insulating materials, cloth, cement, plastics, floor tiles, paper, paints, and brake, clutch, furnace, and kiln linings. Virtually everyone in the general population is exposed to a low level of asbestos fibers, and normal lungs can contain small numbers of them (Gibbs 1996).

Asbestos fibers occur as hydrated fibrous silicates that are mined directly from the Earth. The world's principal, commercially exploited mines are found in Canada, South Africa, Western Australia, and Russia (Roggli 1994). Based on physical and chemical features, asbestos is classified into two major mineralogic groups: serpentine fibers (containing only chrysotile) and amphiboles, comprised of crocidolite, amosite, tremolite, anthophyllite, and actinolite. Serpentine fibers are curly and pliable, whereas the amphiboles are straight, rigid, and brittle (see also Chap. 19, this volume).

Until the late 1970s, chrysotile (white asbestos) accounted for 95% of the asbestos used commercially (Becklake 1983), while crocidolite (blue asbestos) and amosite (brown asbestos) made up the remaining 5%. Current usage is almost 100% chrysotile asbestos. Tremolite, anthophyllite, and actinolite are encountered most commonly as contaminants of other minerals, including chrysotile ore and probably most forms of processed chrysotile products (Dupres et al. 1984). These forms of amphibole asbestos have seen little commercial use, in part because of their physical and chemical properties but primarily due to the lack of commercially useful deposits (Churg 1998a).

It is generally accepted that exposure to the amphiboles is far more pathogenic than exposure to chrysotile, due to the differential rate of fiber clearance between these groups (Gibbs 1996). Evidence indicates that amphibole fibers longer than 20 μm cannot be cleared from the peripheral lung (Morgan 1980). The straight, broad amphibole forms do not readily fragment, whereas the long fibers of chrysotile fragment into short, straight fibers that are cleared (Churg 1998a). Moreover, chrysotile is thought to be chemically unstable and likely to dissolve, whereas amphibole fibers are stable in the environment of the lung (Hume and Roe 1992). Zielhuis (1977) broadly classified exposure to asbestos into *direct*, *indirect* (bystander), *paraoccupational* (women washing contaminated work clothes), *neighborhood*

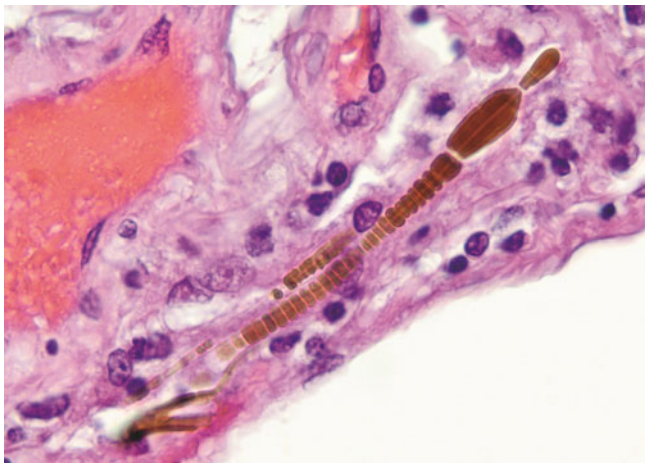


Fig. 25.3 Asbestos body (oil-immersion photomicrograph, ×400) with the characteristic clear core that is surrounded by an iron and mucopolysaccharide coating with terminal bulbs

(living in the vicinity of asbestos mines or factories), and *ambient* (atmospheric contamination).

Asbestos bodies are the most characteristic feature of asbestos exposure (see Fig. 25.3). These structures are composed of a clear asbestos core surrounded by a golden yellow coating of iron and mucopolysaccharides. The clear core distinguishes asbestos bodies from other ferruginous bodies. The coat may be continuous or bead-like with terminal bulbs. Fiber dimensions are important factors in determining whether a fiber becomes coated (Roggli 1992). Longer thicker fibers are more likely to become coated than shorter thinner ones (Morgan and Holmes 1985), and fibers less than 20 μm in length rarely become coated (Morgan and Holmes 1980). Because alveolar macrophages are unable to phagocytize long fibers completely within their cell cytoplasm, they coat the asbestos fibers with an iron–protein matrix (Roggli 1994). For many years, asbestos bodies were thought to be markers of asbestos exposure only in primary asbestos workers. Subsequent studies indicate that asbestos bodies are present in virtually everyone in the general population. Of the asbestos bodies identified in these studies, 98% contained amphibole cores, while the cores in the remaining 2% were chrysotile asbestos (Churg and Warnock 1981); however, asbestos bodies are present in such small numbers from background exposure that an observer should not see an asbestos body in more than 1 in 100 iron-stained, routine-sized paraffin sections (Roggli and Pratt 1983).

25.4.5 Asbestos-Related Diseases

A variety of benign and malignant diseases of the pleura and lung are associated with asbestos exposure. The benign diseases consist of pleural plaques, diffuse pleural fibrosis, pleural effusion, rounded atelectasis, and asbestosis. Carcinoma of the lung in the presence of asbestosis and malignant mesothelioma of the pleura comprise the malignant diseases. Although asbestos fibers are present in virtually everyone in the general population, disease always occurs with fiber burdens greater than those seen in the general population (Wagner et al. 1988). The various asbestos-related diseases occur at different fiber burdens; in general, the fiber burden required to induce disease is greater for chrysotile, with its tremolite contaminant, than for amosite and crocidolite (Churg 1998b; Churg et al. 1993).

25.4.5.1 Benign Pleuropulmonary Disease

Pleural plaques, the most common form of benign asbestos-related pleuropulmonary disease, are discrete, raised, irregular foci of dense fibrous tissue. Thin plaques are grayish-white and smooth, whereas thicker ones are pearly-white with either a smooth or bosselated surface. They vary in size from a few millimeters to 10 cm across, and they vary in consistency from leathery to heavily calcified and brittle (Roberts 1971). Pleural plaques occur most frequently on the posterolateral parietal pleura and on the domes of the hemidiaphragms. Generally, the apices, anterior chest wall, and costophrenic angles are not involved. Pleural plaques are seen predominantly in persons exposed to asbestos, and bilateral plaques are almost pathognomonic of asbestos exposure (Roggli 1994). There is evidence of a dose–response relationship between the presence of pleural plaques and the number of asbestos bodies in the lung (Roberts 1971). Microscopically, pleural plaques consist of dense, paucicellular, hyalinized collagen arranged in a basket-weave pattern in which asbestos bodies are not seen.

In contrast to pleural plaques, diffuse pleural fibrosis typically involves the visceral pleura. The lung may be encased in a thick rind of fibrotic pleura that bridges the major fissures and distorts the edges of the lung (Churg 1998b). In its most severe form, fusion of the visceral and parietal pleura obliterates the pleural cavity, which results in a condition known as fibrothorax. Microscopically, the thickened fibrous pleura consists of dense collagenous tissue admixed with varying numbers of chronic inflammatory cells.

Diagnosis of benign asbestos effusion is based on: (1) a history of exposure to asbestos, (2) confirmation of the effusion by radiographs or thoracentesis, (3) absence of another disease that could produce an effusion, and (4) no malignant tumor developing within 3 years after the effusion (Epler et al. 1982). Asbestos-induced pleural effusion is characteristically a serous or serosanguineous exudate with increased numbers of eosinophils. Most effusions are small (usually less than 500 mL) and asymptomatic and may persist from 2 weeks to 6 months (Fraser et al. 1999a). In 1982, Epler et al. reported that the prevalence of

asbestos-induced pleural effusion was dose related and, with a shorter latency period than other asbestos-related disease, it is the most common disorder in the first 20 years after asbestos exposure.

Rounded atelectasis is most often an incidental radiographic finding and is resected for suspicion of neoplasm (Churg 1998b). Radiographically, rounded atelectasis is a unilateral, rounded, pleural-based mass in the lower lobe of the lung, with one or more curvilinear densities radiating from the mass toward the hilum of the lung (Comet's tail sign). Grossly, the visceral pleura is irregularly thickened and invaginated into the underlying lung. Pleural fibrosis and folding, accompanied by a variable degree of parenchymal atelectasis and fibrosis, are seen microscopically. The majority of cases have a history of asbestos exposure; however, other causes of chronic pleuritis, such as congestive heart failure, pulmonary infarct, tuberculosis, and histoplasmosis, have also been associated with rounded atelectasis.

Roggli and Pratt (1992) describe asbestosis as the prototype of disease caused by the inhalation of mineral fibers. Asbestosis is defined as interstitial fibrosis of the lung parenchyma in which asbestos bodies or fibers may be demonstrated, and it is the only asbestos-related disease to which this term should be applied (American Thoracic Society 1986). Asbestosis shows a dose–response relationship, and there appears to be a threshold below which asbestosis is not seen (Browne 1994). The time from initial exposure to the appearance of asbestosis—the latency period—is inversely proportional to the exposure level and is generally several decades (Roggli and Pratt 1992). Studies suggest that cigarette smoke acts synergistically with asbestos by increasing the incidence of asbestosis (Barnhart et al. 1990; Roggli and Pratt 1992).

The clinical, physiologic, radiologic, and pathologic findings in asbestosis vary with the severity of the disease. The clinical and physiologic findings are not pathognomonic of asbestosis, and they can be seen in diffuse interstitial fibrosis of any cause. Patients with well-established disease usually present with shortness of breath, basilar end-inspiratory crepitations, and a restrictive defect with decreased diffusing capacity on pulmonary function testing. Small, irregular linear opacities, most prominent in the lower lobes, are seen on plain films (Fraser et al. 1999a).

The earliest microscopic manifestation of asbestosis is fibrosis of the walls of respiratory bronchioles. As the disease progresses, fibrosis involves the walls of terminal bronchioles and alveolar ducts, and ultimately it extends into the adjacent alveolar septae. The minimal histologic criteria for the diagnosis of asbestosis can be defined as the presence of peribronchiolar fibrosis and asbestos bodies, with or without accompanying alveolar fibrosis (Craighead et al. 1982). As peribronchiolar fibrosis occurs with inhalants other than asbestos, it is suggested that, in the absence of alveolar septal fibrosis, the histologic diagnosis of asbestosis be restricted to cases in which the majority of bronchioles are involved (Roggli 1989). The report from the College of American Pathologists suggests that there must be a minimum of two asbestos bodies in areas of fibrosis for the histologic diagnosis of asbestosis (Craighead et al. 1982).

25.4.5.2 Malignant Pleuropulmonary Disease

Numerous epidemiologic studies demonstrate an association between asbestos and an excess risk of lung cancer (McDonald 1980). The clinical features—anatomic distribution within the lung and histologic subtypes of asbestos-related lung cancer—are identical to carcinoma in non-exposed individuals. The latency period for asbestos-related lung cancer is lengthy; it peaks at 30–35 years (Selikoff et al. 1980). A linear dose–response relationship exists between exposure and the risk of lung cancer at high-exposure levels (McDonald 1980), but at low-exposure levels Browne (1986) suggests that a threshold exists below which the risk of cancer is not increased. This threshold is thought to be in the range of exposure required to produce asbestosis (Churg 1998b). Available data on asbestos workers indicate that the interaction between cigarette smoke and asbestos in increasing the risk of lung cancer is synergistic, or multiplicative, rather than additive (Greenberg and Roggli 1992). Although persons with asbestosis are at excess risk of developing lung cancer (Doll 1993), investigators are uncertain whether this risk is directly related to the asbestos or to the pulmonary fibrosis. Several studies (Sluis-Cremer and Bezuidenhout 1989; Hughes and Weill 1991) demonstrate an increased rate of lung cancer only in the presence of asbestosis, supporting the view that asbestos is carcinogenic because of its fibrogenicity. Churg (1998b) offers the following approach when evaluating cancer in a given case:

> If asbestosis is present, then the carcinoma is ascribed to asbestos exposure. If the patient smokes or has a history of smoking, then smoking is considered a contributing factor to causation. In the absence of asbestosis, the cancer should not be ascribed to asbestos exposure; the most common causative agent in this instance is cigarette smoke.

Malignant mesothelioma may be idiopathic, or it may occur secondary to ionizing radiation, chemical carcinogens, chronic inflammation and scarring of the pleura, and erionite exposure (Browne 1994; Churg 1998b); however, it is most strongly associated with asbestos exposure. The latency period for asbestos-related mesothelioma is long, peaking at 30–40 years, and virtually never occurring before 15 years (McDonald and McDonald 1987). Asbestos fiber types differ in ability to produce mesothelioma, with the amphiboles posing greater risk than chrysotile (Browne 1994). Characteristically, mesothelioma grows over the surface of the lung in thick sheets or as nodular masses, which may progress to

encase the lung in a hard white rind. Mesotheliomas spread by direct extension into adjacent structures or by lymphatic or hematogenous metastases. Peripheral lung cancers may spread in a similar fashion, which makes the distinction from mesothelioma difficult. Generally, the presence of metastatic disease at clinical presentation favors carcinoma. Mesotheliomas are subclassified by microscopic appearance into epithelioid, sarcomatoid, and biphasic patterns. The diagnosis of diffuse malignant mesothelioma can be very difficult, and it must be distinguished from benign reactive pleural lesions as well as primary and metastatic neoplasms of the pleura.

Frequently, disorders of the lung can lead to cardiovascular dysfunction due to added stress from increased vascular pressures. In the next section, toxicity from metals to the heart is discussed.

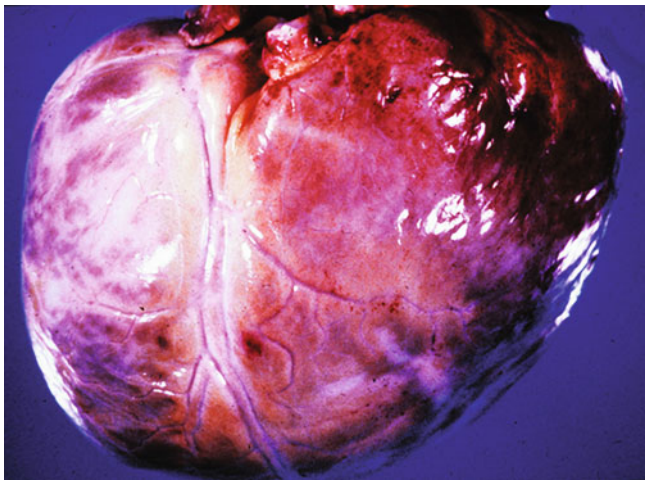


Fig. 25.4 Gross appearance of Keshan cardiomyopathy demonstrating globular configuration of the external aspect of the heart

25.5 Cardiovascular System

Deficiencies of trace elements as well as toxic exposures of metals may be involved in physiologic changes in the cardiovascular system. The clinical, pathologic, and epidemiologic data that support or refute an association between metals and three common cardiovascular disorders—dilated cardiomyopathy, atherosclerosis, and systemic hypertension—are discussed in detail in this section.

25.5.1 Dilated Cardiomyopathy

25.5.1.1 Selenium Deficiency

Selenium is an essential nutrient in trace quantities. It combines with cysteine as a component of selenoproteins, many of which have antioxidant properties. For example, thioredoxin reductase, iodothyronine deiodinases, and glutathione peroxidases are selenium-dependent enzymes (Rayman 2000). Selenoproteins are believed to be especially important in relation to the immune response and cancer prevention; however, the role of selenium in the maintenance of the cardiovascular system is less clear (Rayman 2000).

It has been suggested that two levels of selenium deficiency are involved in the causation of human disease (Rayman 2000). Rare endemic diseases occur where the soil is extremely low in selenium, specifically parts of China. These diseases include Keshan cardiomyopathy and Kashin-Beck disease, a deforming arthritis (See Chap. 16, this volume.) High-prevalence diseases, such as cancer and heart disease, may have as a risk factor a relatively mild deficiency of selenium, such as has been reported in Europe. The role of mildly low selenium intake in the development of reproductive disease, mood disorders, thyroid function, inflammatory disease, and cancer has been recently reviewed (Rayman 2000).

Markedly inadequate dietary intake of selenium in areas of China with endemic selenium deficiency is associated with a form of dilated cardiomyopathy referred to as Keshan disease (Ge and Yang 1993). It has been demonstrated that selenium levels in soil and rocks vary greatly in China and that areas with low levels have higher rates of the disease (Ge and Yang 1993). Intervention studies that have shown a prophylactic effect of sodium selenite provide further evidence that selenium plays a role in Keshan disease. Morphologically, the condition is characterized by multifocal necrosis and replacement fibrosis of the myocardium, which results in acute or chronic heart failure (Ge et al. 1983) (see Figs. 25.4, 25.5 and 25.6). Some patients with Keshan disease show the clinical features of dilated cardiomyopathy, but autopsy studies from China demonstrate distinct pathologic features that separate Keshan disease from sporadic dilated cardiomyopathy (Li et al. 1985). Typical dilated cardiomyopathy does not demonstrate frequent necrosis or fibrosis, which are considered hallmarks of Keshan cardiomyopathy. Ultrastructural observations have shown mitochondrial abnormalities (Ge et al. 1983), which have recently been expanded to include biochemical defects and proteinaceous granular deposits (Yang et al. 1988).

Activities of succinate dehydrogenase, succinic oxidase, and cytochrome *c* oxidase, H(+)-ATPase and its sensitivity to oligomycin, as well as the response of membrane potential to energization by ATP, are decreased, and affected mitochondria had markedly decreased selenium content (Yang et al. 1988). The sole role of selenium in the causation of Keshan disease is debated, and it is generally believed that other cofactors are responsible for the pathogenesis of Keshan cardiomyopathy, including possible genetic and

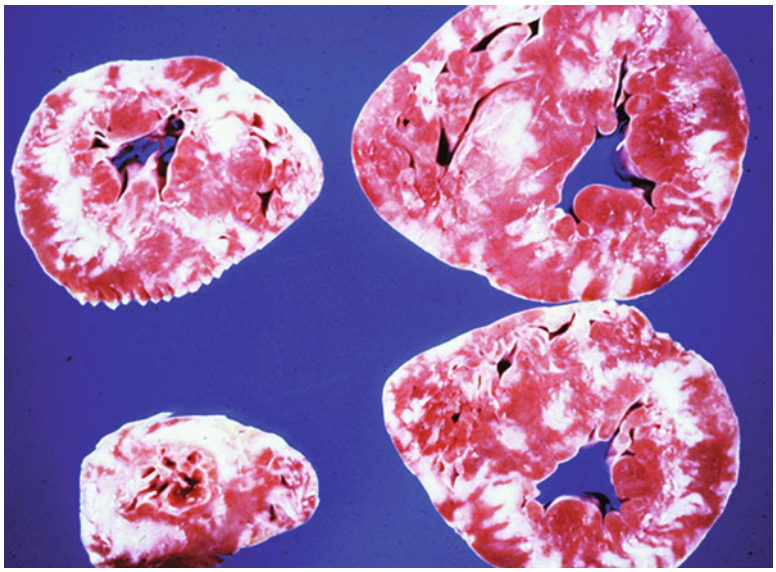


Fig. 25.5 Gross appearance of cross sections of heart muscle demonstrating multiple areas of *white* tan scar tissue

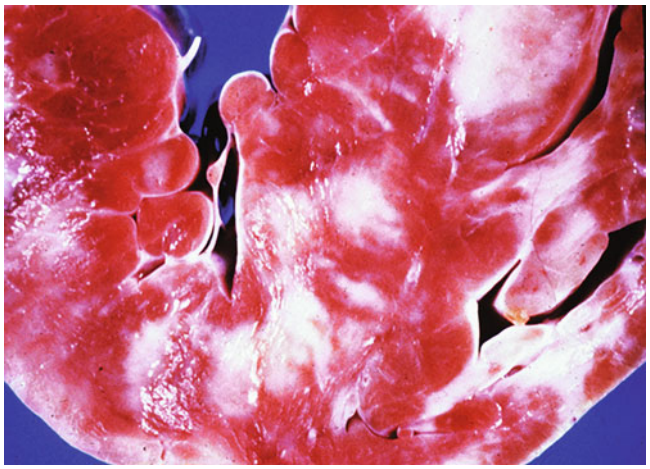


Fig. 25.6 Gross appearance of scar tissue involving the *right* and *left* ventricles of the heart

other environmental factors (Li et al. 1985). Recently, enteroviral infection as a cofactor has been proposed based on the finding of viral genomes in the myocardial tissue of affected patients (See Chap. 16, this volume).

An association between sporadic dilated cardiomyopathy and selenium deficiency is far less clear than the association between selenium and Keshan disease. Serum levels of selenium may be decreased in patients with dilated and postpartum cardiomyopathy in Chinese and African populations as well as Indians (Vijaya et al. 2000), although the data in idiopathic dilated cardiomyopathy are conflicting in Western and Middle Eastern countries (Raines et al. 1999). There have been anecdotal reports of reversible cardiomyopathy due to selenium deficiency in patients with total parenteral nutrition. Selenium deficiency may play a role in the cardiopathy associated with chronic renal failure, but some studies do not support any association (Li and Nan 1989; Chou et al. 1998). The association between dilated cardiomyopathy and dystrophic epidermolysis bullosa does not appear to be mediated by selenium deficiency. The mechanism of cardiomyopathy in selenium deficiency is unclear but is likely related to protection of low-density lipoproteins against oxidative modification, regulation of glutathione peroxidase (Li and Nan 1989; Chou et al. 1998), modulation of prostaglandin synthesis and platelet aggregation, and protection against toxic heavy metals (Oster and Prellwitz 1990; Neve 1996).

Cobalt Toxicity

An epidemic of cardiomyopathy related to cobalt toxicity was reported in 1967 in Canada (Kesteloot et al. 1968). Cobalt ingestion in this epidemic was related to the addition of cobalt sulfate or cobalt chloride to beer as a stabilizer of beer foam. The mechanism of myocyte damage is unclear but may involve increased cardiac vulnerability to oxygen free radicals and may be enhanced in patients with dietary deficiencies. The histologic features of cobalt cardiomyopathy differ somewhat from idiopathic cardiomyopathy, in that greater myofibrillar loss and atrophy and less fibrosis and myocyte hypertrophy are observed in cobalt cardiomyopathy as compared to idiopathic dilated cardiomy-opathy (Centeno et al. 1996) (see Fig. 25.7). Based on heavy metal analysis of cardiac tissues, there is little evidence that elevated cobalt levels are a significant factor in sporadic dilated cardiomyopathy (Centeno et al. 1996).

Mercury-Induced Cardiomyopathy

Dilated cardiomyopathy has been described in Greenlanders with high mercury levels in their blood, presumably from

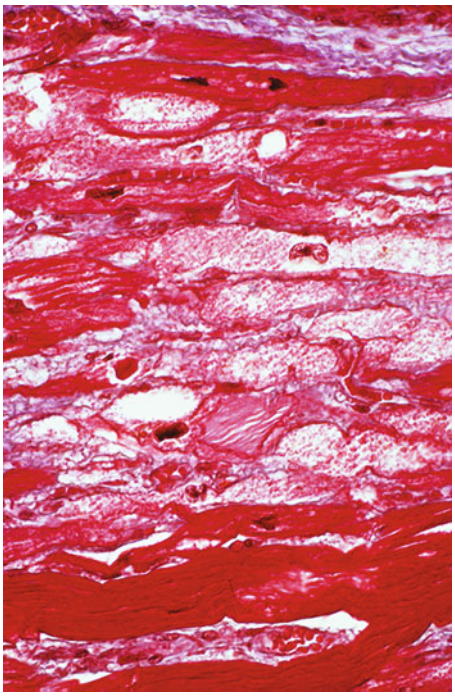


Fig. 25.7 Photomicrograph of myocytes with decreased numbers of myocytes (H&E ×400)

eating seal meat contaminated with high levels of mercury. In addition, evidence suggests that, in cases of sporadic dilated cardiomyopathy, there are thousand-fold elevations of mercury in heart tissue (Frustaci et al. 1999). This latter observation remains to be duplicated, and the mechanism of potential mercury-induced cell damage must yet be elucidated. It is believed that mercury ions may act as calcium antagonists at the actin–myosin junction, inhibiting sarcomere contraction, or that they may disrupt microtubule structure. It has been postulated that a source of mercury in patients with dilated cardiomyopathy may, in addition to occupational exposure, be amalgam fillings, although this remains conjectural.

Iron Overload

Iron overload may adversely affect the heart in patients with hemosiderosis or primary hemochromatosis. The most common cause of death in patients with hemochromatosis is from cardiac failure and arrhythmias followed by cirrhosis and hepatocellular carcinoma. Histologically, these changes are similar to dilated cardiomyopathy with abundant stainable iron present within myocytes. Most patients with hemochromatosis-related cardiomyopathy have a genetic predisposition to iron overload (primary hemochromatosis); significant cardiac failure in patients with hemosiderosis secondary to blood transfusions is rare. There is little evidence that elevated serum iron plays a role in idiopathic dilated cardiomyopathy, although the data are conflicting (Oster 1993; Chou et al. 1998). A recently described mutation in the hemochromatosis gene may contribute to the development of dilated cardiomyopathy that possibly is not mediated by a rise in serum iron (Mahon et al. 2000).

Aluminum Toxicity

Aluminum levels may be elevated in patients on hemodialysis. Although multifactorial, the mechanism of cardiac hypertrophy in patients with end-stage renal disease may involve direct aluminum cardiotoxicity (Reusche et al. 1996). Aluminum may be demonstrated by special stains in the myocardium of such patients. There is little reported evidence that aluminum toxicity plays a role in idiopathic dilated cardiomyopathy.

Arsenic Toxicity

Acute arsenic toxicity may result in a toxic myocarditis, and chronic exposure to arsenic, such as in occupational exposure, may cause chronic cardiomyopathy characterized by small vessel disease and interstitial inflammation with fibrosis (Hall and Harruff 1989).

Miscellaneous

Elevations of other potentially toxic metals have been demonstrated in heart tissue from patients with dilated cardiomyopathy. In addition to cobalt and mercury, these metals include antimony, gold, and chromium (Frustaci et al. 1999). The role of these elements in the pathogenesis of dilated cardiomyopathy remains to be elucidated.

25.5.2 Atherosclerosis

25.5.2.1 Iron Excess

The body has no method of controlling iron excretion; therefore, dietary excesses may lead to iron overload (Schumann 2001). The most likely excesses result from dietary supplements or alcoholic beverages brewed in iron containers (Schumann 2001). Genetic factors are also important in chronic iron overload (hemochromatosis). The medical literature covering the experimental and epidemiological relationship between excess iron stores and coronary artery disease has been recently reviewed (de Valk and Marx 1999). It has long been appreciated that males have a higher incidence of atherosclerosis than females and have greater body iron stores; however, a causative link between iron and atherosclerosis has not been fully established, and a recent prospective study has shown no relationship between iron stores (serum ferritin) and the development of ischemic heart disease (Sempos et al. 2000).

Experimental data support the role of iron in the process of lipid peroxidation, which is hypothesized to be involved in early stages of atherosclerosis. The exact mechanism by

which endothelial cells and macrophages interact with iron and low-density lipoprotein are still unknown. There is speculation that catalytically active iron, in the form of nontransferrin-bound plasma iron and hemoglobin, oxidizes low-density lipoprotein to interact with the macrophage oxidized low-density lipoprotein receptor (Evans et al. 1995).

At a clinical level, the role of iron in lipid oxidation has been questioned. Some studies have failed to show an association between serum ferritin or dietary iron intake with markers of low-density lipoprotein oxidation (Iribarren et al. 1998). Several studies correlating body iron stores with clinically measured atherosclerosis have been positive (Kiechl et al. 1997), which demonstrates a synergistic effect with low-density lipoprotein levels. Other studies have been negative (Rauramaa et al. 1994). A recent review has demonstrated that the majority of studies correlating body iron stores with clinical assessments of atherosclerosis have not shown a significant correlation between clinical indices of coronary, carotid, or aortic atherosclerosis and body iron stores (Meyers 1996).

25.5.2.2 Selenium Deficiency

Like dilated cardiomyopathy, ischemic heart disease has been linked to low levels of serum selenium. The mechanism of selenium deficiency resulting in an increased risk of coronary artery disease is unknown but may be related to increased oxidative stress, as is the case with iron, or increased platelet aggregation (Salonen et al. 1988).

Epidemiological studies have provided some evidence for the role of selenium deficiency in the etiology of atherosclerotic disease. Angiographic studies suggest that low levels of selenium are associated with coronary stenosis (Yegin et al. 1997), and patients with acute myocardial infarction tend to have lower serum selenium compared to controls (Navarro-Alarcon et al. 1999). Plasma, red blood cell, and urine selenium concentrations have been shown to be decreased in patients with acute myocardial infarction. The results of longitudinal studies within populations are conflicting (Huttunen 1997), however, and the risk of decreased levels of selenium appears to correlate with elevated low-density lipoprotein cholesterol. A nested case-control study among participants in the Physicians' Health Study showed no effect of selenium on the risk for acute myocardial infarction (Salvini et al. 1995). Other studies using toenail selenium concentration, which is a better measure of chronic selenium levels, have shown an inverse relationship between selenium and myocardial infarction only in Germany, which has relatively low levels of selenium (Kardinaal et al. 1997).

Salonen et al. (1982) showed an increase in overall cardiovascular morbidity and mortality for individuals with low serum selenium (less than 45 μg/L), whereas Virtamo's group (1985) found no significant associations except for stroke mortality. A large prospective study in a Danish population demonstrated that men with the lowest tertile of blood selenium had a mildly increased risk for developing acute ischemic events of stroke or myocardial infarct (risk ratio of 1.7); this risk remained significant to a low level of probability ($p = 0.05$) when other risk factors were considered. This study also demonstrated a correlation between selenium and traditional risk factors, including tobacco smoking, social class, alcohol consumption, total cholesterol, hypertension, age, and physical inactivity (Suadicani et al. 1992). Other studies have found that selenium levels are affected by cigarette smoking and alcohol use, and that selenium is a negative acute phase reactant and may therefore be secondarily decreased by the inflammatory component of atherosclerosis. The current consensus is that the effect of selenium on ischemic heart disease is small and likely mediated by an association with other risk factors (Rayman 2000).

25.5.2.3 Magnesium

The majority of research related to magnesium and cardiovascular disease addresses magnesium physiology after acute myocardial infarction. Specifically, the role of intravenous magnesium in the treatment of acute myocardial infarction and the effects of magnesium levels or intake on the risk of acute myocardial infarction are controversial areas. Less attention has been paid to the effects of chronic intake of low magnesium because of environmental and natural geological factors and the effect of chronic magnesium supplementation on the incidence of acute coronary events. Low levels of magnesium and fluoride in drinking water were found to be associated with increased risk for acute myocardial infarction in Finland and Sweden. No such effect was found in England, where there was a small protective effect of magnesium in drinking water against death due to ischemic heart disease, but no protection against deaths due to acute myocardial infarction. Oral magnesium supplementation does not decrease the rate of subsequent cardiac events after myocardial infarction, and the level of serum magnesium is not related to morbidity in acute myocardial infarction patients.

25.5.2.4 Calcium

Calcium has a complex relationship with atherosclerosis, and it is frequently deposited in the extracellular matrix of advanced atherosclerotic plaques. There is no clear association between dietary calcium or serum levels of calcium and the risk for development of coronary atherosclerosis. High levels of serum calcium may have a mildly protective effect, because there is a borderline inverse relationship with serum calcium and the extent of coronary atherosclerosis (Narang et al. 1997). However, there is no clear association between calcium in drinking water and the incidence of deaths due to

acute myocardial infarction deaths, as demonstrated in studies in England and Scandinavia (Maheswaran et al. 1999). (See Chap. 14, this volume.) An exception to these results is a study from Finland that demonstrated a protective effect of high levels of calcium in drinking water to fight against acute myocardial infarction in postmenopausal women (Rubenowitz et al. 1999).

25.5.2.5 Other Metals

The effects of trace elements on atherosclerosis are difficult to determine because of the complex interplay between metal levels and traditional risk factors for cardiovascular disease, such as disorders of blood lipids, blood pressure, coagulation, glucose tolerance, and circulating insulin. In addition, the relevance of serum metal levels, metal concentrations within blood cells, and analyses of metal content of portions of atherosclerotic plaque is often unclear.

Like iron, copper ions may exert oxidative stress on the arterial wall. Elevated copper within circulating white cells has been associated with atherosclerosis. Elevated serum copper has been related to peripheral vascular disease, especially when related to serum zinc levels. No consistent data show a link between zinc and atherogenesis. Pathologic studies have demonstrated elevated levels of zinc within fibrous plaques of the aorta (Mendis 1989). There is evidence that zinc requirements of the vascular endothelium are increased during inflammatory conditions such as atherosclerosis, where apoptotic cell death is also prevalent. Theoretically, zinc deficiency may exacerbate the detrimental effects of specific fatty acids and inflammatory cytokines on vascular endothelial functions, but an association between decreased dietary zinc and an increased risk for coronary atherosclerosis has yet to be demonstrated.

A mortality study of mercury miners in Europe demonstrated no increased deaths due to ischemic heart disease (Boffetta et al. 2001); however, in a prospective study among 1,014 men, high hair mercury content was one of the strongest predictors of the 4-year increase in the mean intima-media thickness, which is a measurement of carotid atherosclerosis (Salonen et al. 2000).

A study of chromium welders has suggested that welders exposed to chromium have increased plasma lipid oxidation (Elis et al. 2001). Although arsenic has been only indirectly associated with coronary heart disease, increased serum arsenic may lead to persistent oxidative stress, thereby theoretically leading to atherosclerosis.

25.5.3 Hypertension

25.5.3.1 Lead

The toxic effects of lead poisoning are well known; however, the effects of lower levels of lead exposure are uncertain. Although the finding is controversial, chronic low-level lead exposure has been linked to hypertension in both clinical and experimental studies. The exact pathogenic mechanisms that underlie the actions of lead in the cardiovascular system have yet to be elucidated definitively (Kopp et al. 1988).

There are a number of difficulties in proving a link between low-level exposure to lead (blood levels of <1.45 μm/L or <30 μg/dL) and the development of high blood pressure. A number of possible covariates must be considered in epidemiologic studies, such as body mass index, alcohol consumption, other metals, race, and gender. Measurements of blood levels may not accurately reflect body stores, and there may be discrepant results depending on whether blood lead, bone lead, or urinary lead levels are obtained after chelation testing is performed. Several cross-sectional epidemiologic studies have demonstrated an association between blood lead levels and hypertension in individuals without known occupational exposure and in men with occupational risk. In most of these studies, the association persists after correction for other variables. A similar number of studies have shown no association or only a very weak association between lead and hypertension, after adjustment for age and body mass index, in populations with or without known lead exposure. A study from the United States demonstrated an association only in African-Americans, who had higher lead levels (Sokas et al. 1997). In a study of postpartum women in Los Angeles, an association between lead and hypertension was demonstrated only in immigrants who had, in general, higher levels than nonimmigrants. A study in Germany demonstrated a link only in heavy drinkers of alcohol; results from a cross-sectional analysis of alcoholics in the United States also demonstrate a link between hypertension and blood lead levels. A metaanalysis of pooled data from several studies suggested only a weak positive association between blood pressure and lead exposure (Staessen et al. 1994).

Cross-sectional epidemiologic studies using bone indices have often shown a positive correlation between lead exposure and hypertension with an estimated relative risk of about 1.5 times for individuals with increased skeletal lead stores in men and women. These results, however, are not always confirmed by blood testing in the same patients, and results vary by bone site measured (*e.g.,* patella vs. tibia). Studies estimating body stores of lead in measurements of blood and urine after a chelating challenge have also suggested a link between hypertension and increased lead stores, especially in the presence of renal failure (Sanchez-Fructuoso et al. 1996).

In the United States and Western countries, mild blood pressure elevations due to moderate increases in lead blood levels translate into potentially large numbers of patients dying with coronary artery disease. This link between lead

and coronary artery disease is corroborated by data showing a direct association between blood lead and electrocardiographic changes of left ventricular hypertrophy. There is evidence that occupational exposure to lead is associated with increased risk of dying from hypertension-related illness. A prospective study in Belgium, however, in which patients were followed for 6 years did not show any association between blood levels and hypertension, a finding that refutes a significant link between coronary heart disease and lead exposure (Staessen et al. 1996). The mechanisms by which lead may induce hypertension include increased vascular responsiveness to catecholamines mediated by effects on calcium channels and increased expression of endothelin and generation of reactive oxygen species.

Sources of lead exposure are numerous and include occupational exposure (workers in battery factories, smelters), environmental exposure (traffic exhaust, dust, paint chips, drinking water), cosmetics, food supplements, food preparation utensils, and illegal alcoholic beverages ("moonshine").

25.5.3.2 Mercury

Acute mercury poisoning may result in hypertension, but few data suggest that chronic low-level exposure to mercury is an etiologic factor in essential hypertension. Mortality statistics from mercury miners in several regions across Europe demonstrated no consistent increase in deaths due to hypertension (Boffetta et al. 2001), with the exception of Spain and Ukraine, which had elevated mortality ratios of 2.8 and 9.4, respectively, for hypertension-related diseases.

25.5.3.3 Other Metals

Chronic arsenic exposure from drinking water in contaminated wells has been implicated in a modest increased risk for hypertension in China and Bangladesh. Patients with hypertension, who are not on dialysis, may have elevated levels of serum aluminum, but an etiologic link has yet to be shown between hypertension and aluminum toxicity.

The effects of hypertension can ultimately lead to cardiac failure, resulting in increased venous backpressures. Depending on the severity of the failure, severe chronic congestive heart failure may lead to fibrosis of the liver. Direct toxicity of metal ions on the liver also may lead to injury to the liver and fibrosis of the liver, as is discussed in the final section.

25.6 Hepatotoxicity of Metal Ions

Several metals can injure the liver, but by far the most important are iron and copper. The diseases caused by metals may be genetic or acquired, and the effects can be acute or chronic (see Table 25.2). Acute metal toxicity produces hepatocellular and cholestatic injury:

1. *Hepatocellular injury*—Ferrous sulfate poisoning in children can lead to coagulative degeneration in zone 1 of the hepatic acinus, and phosphorus poisoning induces lytic necrosis in that zone, as well as steatosis. Copper toxicity causes zone 3 necrosis.
2. *Cholestatic injury*—Intrahepatic cholestasis has been reported with acute arsenical toxicity and as an idiosyncratic reaction to the use of gold salts for treatment of rheumatoid arthritis.

In chronic metal toxicity the injuries listed below are seen:

1. *Vascular injury*—Hepatoportal sclerosis is a recognized complication of chronic arsenical toxicity, for example, from ingestion of high levels of arsenic in drinking water (India, Bangladesh).
2. *Chronic hepatitis*—Chronic hepatitis is a stage in the evolution of Wilson's disease; hemosiderin is an accumulation in the liver in chronic hepatitis C which may lead to a poor response to interferon therapy.
3. *Fibrosis and cirrhosis*—Fibrosis and cirrhosis occur in genetic hemochromatosis, Wilson's disease, Indian childhood cirrhosis, acquired copper toxicosis, neonatal hemochromatosis, and sub-Saharan hemosiderosis; the presence of iron in nonalcoholic steatohepatitis may increase the hepatic injury and contribute to the fibrosis.
4. *Granulomas*—Chronic beryllium toxicity (berylliosis) in the past was associated with a sarcoidosis-like disease; exposure of vineyard sprayers to copper sulfate in Portugal has been reported to lead to granulomas in the lungs and liver.
5. *Malignant tumors*—Hepatocellular carcinoma is a dreaded complication of genetic hemochromatosis and, rarely, of Wilson's disease; angiosarcoma has been reported after chronic exposure to arsenic and rarely to iron or copper.

The hepatotoxicity of several metals, in alphabetical order, is discussed in more detail in the following paragraphs.

25.6.1 Arsenic

25.6.1.1 Acute Toxicity

Toxicity from inorganic arsenicals may result from attempts at homicide or suicide; exposure to insecticides, herbicides or rodenticides; industrial exposure; exposure to naturally contaminated materials; or from treatment with arsenical medications. The clinical features, mechanism of action, and other aspects of arsenical toxicity are comprehensively discussed in a review by Schoolmeester and White (1980).

Table 25.2 Hepatotoxicity of metals

Metal	Circumstances	Histopathology	Comments
Acute toxicity			
Iron	Accidental ingestion by children (usually $FeSO_4$)	Zone 1 necrosis	Also severe gastrointestinal injury
Copper	Suicidal or accidental	Zone 3 necrosis	Also cholestasis
Chronic toxicity			
Iron	Genetic hemochromatosis	Hepatic hemosiderosis, fibrosis, cirrhosis, HCC	C282Y and H63D gene mutations
Other (see Table 25.2)			
Copper	Wilson's disease	Copper overload, chronic hepatitis, cirrhosis, fulminant liver failure with necrosis, HCC (rare)	ATP7B gene mutations
	Indian childhood cirrhosis	Copper overload, Mallory bodies, fibrosis, cirrhosis	Excess ingestion of copper in milk from copper or brass containers
	Endemic infantile	Copper overload, cirrhosis	Excess ingestion of copper in milk plus genetic factor
	Tyrolean cirrhosis		
Arsenic	Excess ingestion of arsenic via drinking water, drugs, or exposure to insecticides	Hepatoportal sclerosis, angiosarcoma	

Note: HCC hepatocellular carcinoma

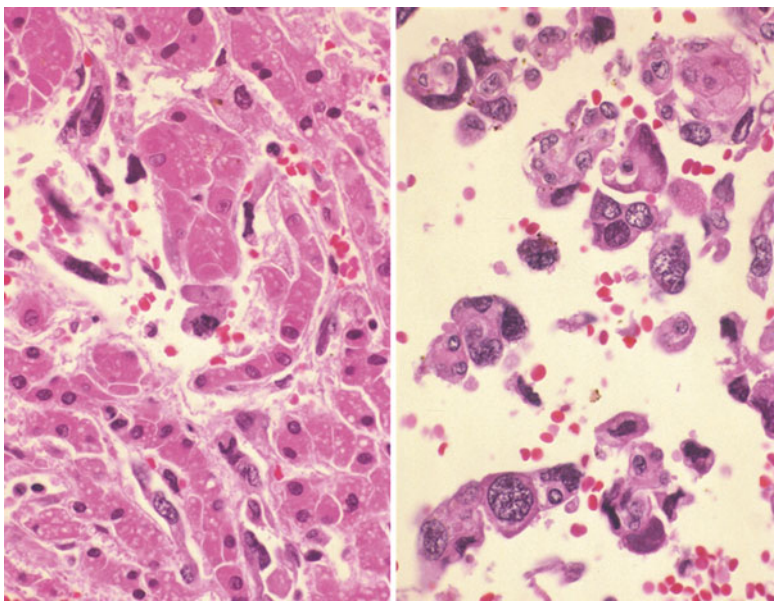


Fig. 25.8 Arsenic-related angiosarcoma. (**a**) The malignant endothelial cells infiltrating the sinusoids have led to disruption of the liver cell plates (**b**); higher magnification of the angiosarcoma cells (**b**) shows marked nuclear pleomorphism and hyperchromasia

Hepatic injury is usually overshadowed by the effects of damage to other organ systems, such as the gastrointestinal, cardiovascular, neurologic, and hematologic ones. Observed histologic changes in the liver have included steatosis, cholestasis, and zone 3 necrosis. A striking increase in hepatocytic mitotic activity was noted in two cases reported by Brenard et al. (1996). Hepatic veno-occlusive disease following severe arsenic poisoning was described by Labadie et al. (1990). Predominantly cholestatic injury has been reported with the antiamebicidal agent carbarsome and the antisyphilitic drug arsphenamine. Acute arsenic and mercury intoxication has also resulted from ingestion of Indian ethnic remedies (Kew et al. 1993) and Chinese herbal balls (Espinoza et al. 1995). Experimental work with arsenates in rabbits was reported to cause periportal focal necrosis and cirrhosis of an unspecified type (Van Glahn et al. 1938).

25.6.1.2 Chronic Toxicity

Chronic arsenical injury to the liver can lead to hepatoportal sclerosis or noncirrhotic portal hypertension (Datta et al. 1979; Mazumder et al. 1988), cirrhosis (Denk et al. 1959; Lunchtrath 1972), systemic arterial disease (Rosenberg 1974), hepatocellular carcinoma (Jhaveri 1959), and angiosarcoma (Falk et al. 1981) (see Fig. 25.8). The chronic exposure was the result of therapy (*e.g.*, the use of Fowler's

solution [potassium arsenite] for the treatment of psoriasis), high levels of arsenic in drinking water, use of arsenical preparation as insecticides in vineyards, or industrial or environmental exposure (*e.g.*, among workers in copper smelters or residents living nearby). The histologic changes that have been described in hepatoportal sclerosis include portal area expansion and fibrosis, an increase in the number of vessels in portal areas, thickening of portal vein branches (intimal thickening and muscular hypertrophy, but no thrombosis), and perisinusoidal fibrosis. The cirrhosis reported with chronic arsenicism has been described as "post-necrotic" or macronodular in type.

25.6.2 Copper

25.6.2.1 Acute Copper Toxicity

Acute copper toxicity may result from: (1) ingestion (either suicidal or accidental) of copper sulfate, (2) its use as an emetic or for the treatment of burns, (3) the release of copper ions from equipment made of copper, (4) acid pH conditions (*e.g.*, malfunctioning hemodialysis equipment or vending machines), or (5) high levels of copper in drinking water (Blomfield et al. 1971; Spitainy et al. 1984; Holtzman et al. 1986). Copper poisoning leads to a hemolytic anemia related to several effects on erythrocytes that include inhibition of glucose-6-phosphate dehydrogenase activity, inhibition of glycolysis, denaturation of hemoglobin (with Heinz body formation), and oxidation of glutathione. The largest series of cases of acute poisoning with copper sulfate, ingested for suicidal purposes, was reported from New Delhi, India (Chuttani et al. 1965). Jaundice appeared on the second or third day in 11 of 48 patients (23.0%). In five patients, the jaundice was deep, the liver was enlarged and tender, the average total serum bilirubin was 11.2 ± 809 mg/dL, the AST averaged 252.4 ± 142 units, and the prothrombin time was markedly prolonged. In the remaining six patients, the jaundice was mild, the serum bilirubin was about 3.0 mg/dL, and the AST averaged 144 ± 80 IU. Histologic changes in biopsy or autopsy material showed zone three necrosis and cholestasis in the deeply jaundiced patients and focal necrosis or no changes in the mildly jaundiced patients. When measured, the serum copper levels have been markedly elevated in acute copper intoxication (Chuttani et al. 1965; Blomfield et al. 1971; Holtzman et al. 1986), as well as in the liver (Blomfield et al. 1971).

25.6.2.2 Chronic Copper Toxicity

Chronic copper toxicity can be hereditary or acquired. In humans, hereditary copper overload is exemplified by Wilson's disease and in animals by the copper toxicosis of the Long-Evans cinnamon rat; both conditions have an autosomal recessive inheritance and are not discussed in this review. Acquired forms of copper overload also occur in humans and animals. Sheep are particularly susceptible to copper intoxication, as they can be chronically exposed to excessive copper by grazing in sprayed orchards or by eating contaminated feed. The changes in sheep livers include steatosis, focal necrosis, swelling of liver and Kupffer cells, and cytochemically demonstrable copper in Kupffer cells (incorporated in lysosomes with lipofuscin), but Mallory bodies have not been observed. An interesting copper toxicosis was reported in North Ronaldsay sheep by Haywood et al. (2001). These sheep are a primitive breed that has adapted to a copper-impoverished environment in the Orkney Islands. Sheep transferred to copper-replete pastures on the mainland developed hepatic copper toxicity resembling that of Indian childhood cirrhosis and idiopathic copper toxicosis.

Acquired Chronic Copper Toxicity

In humans, acquired chronic copper toxicity may result from occupational environmental, or domestic exposure. An example of occupational exposure is the chronic exposure of vineyard workers to fungicide sprays containing copper sulfate. The resultant hepatic injury includes noncaseating granulomas, intraacinar and periportal fibrosis, cirrhosis, and, rarely, angiosarcoma (Pimental and Menezes 1977). The geological environment is the setting for Indian childhood cirrhosis, and a similar disease reported from outside the Indian subcontinent occurred in children who had ingested large quantities of copper in their drinking water.

Indian Childhood Cirrhosis

Indian childhood cirrhosis (ICC) is associated with marked hepatic copper overload. The copper storage has been demonstrated histochemically (by staining tissue sections with orcein for copper-binding protein and the rhodanine method for copper) and by quantitative techniques such as atomic absorption spectrophotometry. It is generally accepted that copper accumulation is directly responsible for the histopathologic lesions in ICC; these include the presence of numerous Mallory bodies, intraacinar and portal inflammation, copper accumulation (beginning in zone 1), periportal ductular proliferation, intraacinar and periportal fibrosis, occlusive lesions of terminal hepatic venules, and, eventually, the development of a micronodular cirrhosis (Smetana et al. 1961; Joshi 1987; Bhagwat et al. 1983) (see Fig. 25.9).

According to Bhagwat et al. (1983), cases of ICC have been reported from 17 countries as well as from the United States (Lefkowitch et al. 1982). The non-Indian cases also are characterized by marked copper overload and histopathologic alterations indistinguishable from those of ICC. In 1983, Tanner et al. (1983) pro-posed that increased

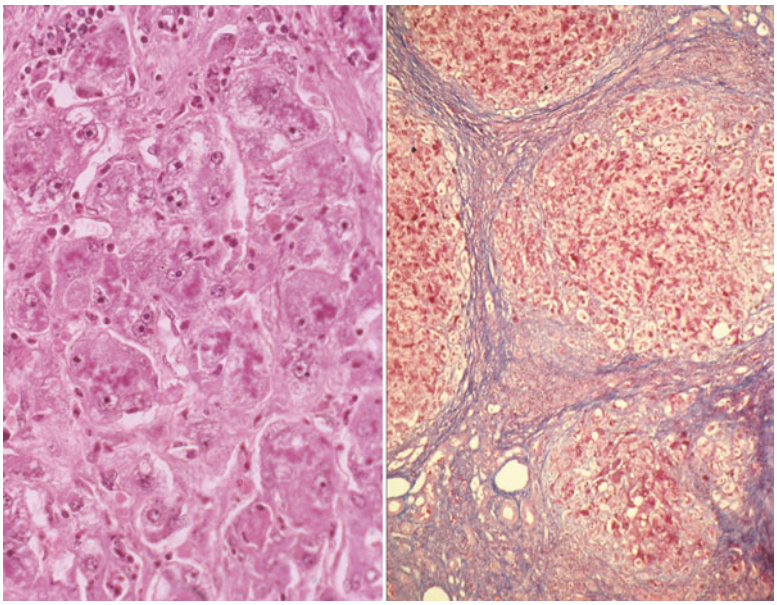


Fig. 25.9 Indian childhood cirrhosis. (**a**) High magnification shows Mallory bodies in liver cells. (**b**) This cirrhotic liver is composed of nodules separated by fibrous septa (Masson)

dietary copper (from copper-contaminated milk stored in brass and copper containers) could be of etiologic significance in ICC. Subsequently, O'Neill and Tanner (1989) demonstrated experimentally that copper (but not zinc) is avidly taken up from brass and bound to casein from which it is completely removable by picolinate chelation. They concluded that milk is an effective carrier of copper from a brass utensil to the infant enterocyte. Lending support to a direct cytopathic effect of copper are recent reports of clinical recovery, improved survival, and reversal of the hepatic histological lesions by penicillamine therapy of infants with ICC (Tanner et al. 1987; Bhusnurmath et al. 1991).

Of great interest are reports of several children who developed an illness, clinically and histopathologically resembling ICC, from the chronic ingestion of well water contaminated with high levels of copper (Müller-Höcker et al. 1985, 1987; Schramel et al. 1988). In all cases, the copper was leached from copper pipes into drinking water delivered to the children's homes. Early exposure to copper appears to be crucial, because siblings exposed after 9 months of age and the parents who drank the same water did not develop the disease (Müller-Höcker et al. 1985; Schramel et al. 1988). Other cases of "non-Indian childhood cirrhosis," also referred to as "idiopathic hepatic copper toxicosis," not specifically related to increased dietary copper intake have been reported (Maggiore et al. 1987; Horslen et al. 1994; Ludwig et al. 1995). To date, most of the affected children have succumbed to liver failure. It is conceivable that some cases of ICC occurring outside the Indian subcontinent are examples of environmental chronic copper toxicity, but other cases are not. The occurrence of the disease in siblings and parental consanguinity in two cases suggest a genetic abnormality, possibly an autosomal recessive disorder (Scheinberg and Sternlieb 1994). Lack of support for excess dietary copper as a cause of non-Indian childhood cirrhosis comes from a study of Scheinberg and Sternlieb (1994) in which no liver-related deaths were found in children less that 6 years of age in three Massachusetts towns with high copper concentrations in drinking water. The study covered 64,124 child-years of exposure. Of relevance to these observations are cases of endemic Tyrolean infantile cirrhosis, which is transmitted by autosomal recessive inheritance but requires an additional risk factor such as excess dietary copper (from cows' milk contaminated with copper leached from untinned copper or brass containers) for its expression (Müller et al. 1996). The disease was eradicated when the untinned copper containers were replaced by modern industrial vessels.

Before leaving this section, it is worth noting that chronic acquired copper toxicity in the adult is very rare. There is one report of chronic toxicity from massive ingestion of copper coins; 275 copper coins were found at autopsy in the stomach of a mentally deranged person (Yelin et al. 1987). Corrosion of the coins led to absorption of copper from the stomach, resulting in liver and kidney injury. The changes in the liver resembled those of ICC and childhood hepatic copper toxicosis. It is important to remember that accumulation of copper occurs in a number of chronic cholestatic conditions, most frequently in primary biliary cirrhosis and primary sclerosing cholangitis. Although demonstration of the stored copper, mainly periportal, is of diagnostic value, it does not appear to be a risk factor for progression of the disease, and attempts at removing the copper by penicillamine therapy have been abandoned.

25.6.3 Iron

Injury to the liver by iron can be either acute or chronic. Genetic hemochromatosis, the most common inherited metabolic disease, is not discussed here; instead, the focus of this section is on acquired iron toxicity, which may be acute or chronic.

25.6.3.1 Acute Iron Toxicity

Accidental overdose of iron-containing drugs and dietary supplements (in tablet or capsule form) is a leading cause of fatal poisoning of children under 6 years of age (Nightingale 1997), but occasional cases have been reported in adults (Monoguerra 1976). Since 1986, U.S. poison control centers have received reports of more than 110,000 incidents of children younger than 6 years accidentally swallowing iron tablets. Almost 17% of pediatric deaths reported to poison control centers between 1988 and 1992 were due to iron poisoning, compared with 12% between 1984 and 1987. Death has occurred from ingesting as little as 200 mg to as much as 5.85 g of iron (Nightingale 1997).

A large number of products available in pharmacies, foodstores, and discount stores contain iron, and the problem is compounded by the attractiveness of dosage forms, high availability, and ambiguous labeling (Krenzelok and Hoff 1979). The Food and Drug Administration (FDA) issued a regulation, effective July 15, 1997, with labeling and packaging requirements to protect children from accidental poisoning from iron-containing drugs and dietary supplements in tablet or capsule form. Under the regulation, these products must display—in a prominent and conspicuous place set off by surrounding lines—the following warning statement: "WARNING: Accidental overdose of iron-containing products is a leading cause of fatal poisoning in children under 6. Keep this product out of reach of children. In case of accidental overdose, call a doctor or poison control center immediately" (Nightingale 1997). In addition, the agency requires most products that contain 30 mg or more of iron per dosage, such as iron tablets or capsules for pregnant women, to be packaged as individual doses (for example, in blister packages). The FDA has concluded that this packaging will limit the number of unit doses that a child may consume once access is gained to the product, thus significantly reducing the likelihood of serious injury.

Although the morbidity of acute iron poisoning is high, the mortality in two large series was low—none in the series of 66 patients reported by James (1970), and 5.16% of the 172 cases reported by Westlin (1966). The typical victim of iron poisoning vomits within 10 min to 1.5 h after swallowing a toxic dose of iron; consequently, it is often difficult to state what the toxic dose is. In a severe case that does not prove fatal, the three stages of poisoning are:

- First stage—This stage lasts approximately 6 h and begins with abdominal pain, nausea, and vomiting and goes on to hematemesis, melena, and subsequently shock and possibly coma.
- Second stage—This stage is one of apparent recovery during which toxic iron compounds are formed.
- Third stage—This stage begins about 24 h after ingestion of the toxic dose. Acidosis and hyperglycemia develop, and there may be convulsions, coma, bleeding, and evidence of hepatic damage. Invariably, edema and necrosis of the gastric mucosa are found in fatal cases. If the child recovers, there may be fibrosis of the gastric mucosa leading to pyloric stenosis.

Gandhi and Robarts (1962) reviewed 11 such cases. The interval between the time of ingestion of the ferrous sulfate and the symptoms of obstruction ranged from 13 to 40 days. Pyloric stenosis was present in five patients. "Hourglass" stricture of the stomach was found in the other six children, two of whom had pyloric stenosis as well; one child also had a penetrating gastric ulcer that involved the liver.

Nearly 90% of deaths from iron poisoning occur in the first 48 h after ingestion. Pathologic findings include widespread necrosis of the upper gastrointestinal tract (see Fig. 25.10), thrombosis of intestinal vessels, periportal (zone 1) necrosis of hepatic cells, hemorrhagic pneumonia, and pericardial effusion. Liver cells in zone 1 show coagulative degeneration with nuclear pyknosis or karyorrhexis, and there is variable drop out of cells (Pestaner et al. 1999) (see Fig. 25.11). The inflammatory response is typically minimal, but there may be a sprinkling of neutrophils and lymphocytes. Histochemical stains for iron often show positive staining of the coagulated cells and damaged sinusoids.

25.6.3.2 Chronic Iron Toxicity

Numerous conditions, other than genetic hemochromatosis, can lead to chronic accumulation of iron in the liver. Among these are the hereditary anemias listed in Table 25.3. In transfusion-dependent β-thalassemia major, the iron overload is severe and, over a period of months or years, can lead to congestive heart failure, hypothyroidism, diabetes mellitus, hepatomegaly, fibrosis, and cirrhosis (Modell 1975; Schafer et al. 1981). In transfusional siderosis, iron initially accumulates in reticuloendothelial cells but eventually storage also occurs in parenchymal cells, beginning in zone 1. A large amount of iron also is found in the spleen, in contrast to genetic hemochromatosis (Oliver 1959).

Iron overload has been reported in a number of inherited metabolic diseases, including hereditary tyrosinemia, Zellweger's syndrome, congenital atransferrinemia, aceruloplasminemia, Wilson's disease after penicillamine therapy (Shiono et al. 2001), porphyria cutanea tarda, and

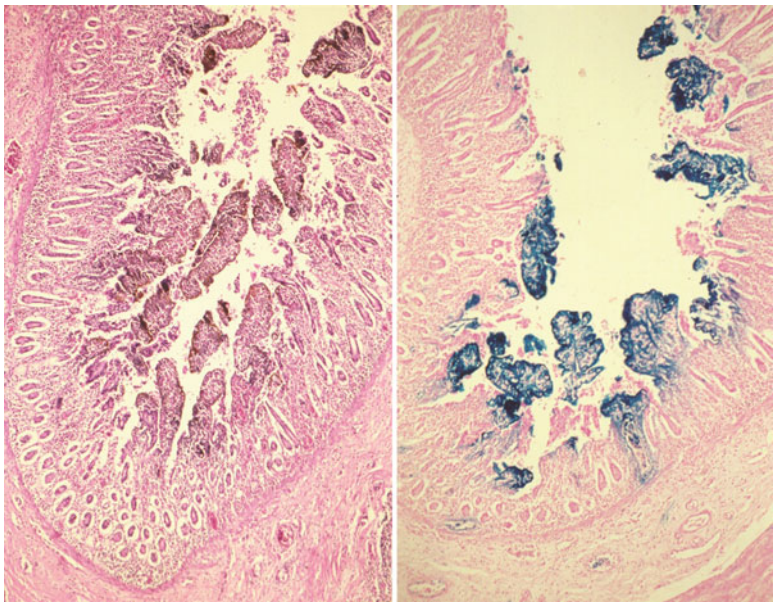


Fig. 25.10 Acute ferrous sulfate toxicity: (**a**) small intestine showing marked necrosis of the villi; (**b**) iron encrustation (*blue*) of the necrotic intestinal villi (Prussian blue stain)

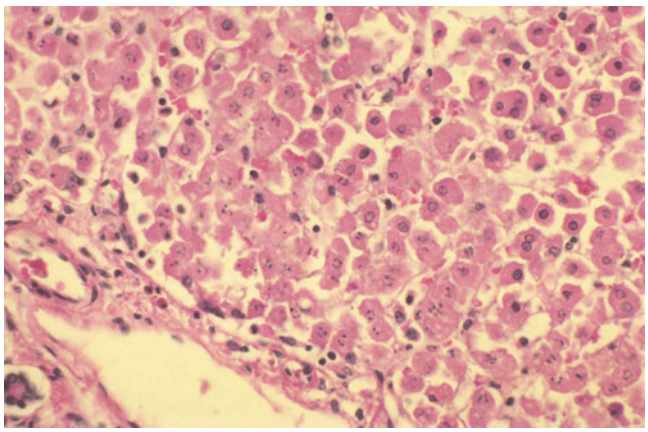


Fig. 25.11 Acute ferrous sulfate toxicity; section of the liver of the same patient illustrated in Fig. 25.10 shows coagulative degeneration and dissociation of liver cells in zone 1. A portal area is present in the *lower left* corner

alpha-1 antitrypsin (AAT) deficiency. In the last condition, no association between genetic hemochromatosis and AAT deficiency was found in one study (Fargion et al. 1996). In another study, the association of AAT deficiency and genetic hemochromatosis led to an earlier onset of cirrhosis in genetic hemochromatosis, but it did not increase the risk of hepatocellular carcinoma (Elzouki et al. 1995). In porphyria cutanea tarda, 60–70% of patients have mild to moderate iron overload, and approximately 10% have increases in the range of genetic hemochromatosis; the major cause is the presence of mutations of the HFE gene (Bonkovsky and Lambrecht 2000). Hepatic iron accumulation can occur in a number of acquired diseases of the liver (Table 25.3). The ones discussed briefly here are chronic hepatitis C, nonalcoholic and alcoholic steatohepatitis, cirrhosis, and post-portacaval shunt surgery.

Hepatic iron accumulation occurs in chronic hepatitis C. In one study from Italy, 10% of patients with chronic hepatitis C had an elevated hepatic iron content (Riggio et al. 1997). Stainable iron in sinusoidal cells and portal tracts reflects an increased hepatic iron concentration, and elevated serum iron values have been linked to poor responses of chronic hepatitis C to interferon therapy. In one study, hepatic iron accumulation was found to be significantly associated with histologic activity and cirrhosis (Hezode et al. 1999). Barton et al. (1995) have suggested that the percentage of portal areas staining for iron should be included in the pathology report of liver biopsies; these investigators found that patients who respond to interferon therapy have less than 40% portal areas positive for iron. Enhanced iron accumulation and lipid peroxidation were improved by interferon therapy in one study (Kageyama et al. 2000). Pirisi et al. (2000) believed that interface hepatitis (piecemeal necrosis) and its sequelae (sinusoidal capillarization and microshunting) are major factors in iron deposition in chronic hepatitis C. HFE mutations are not responsible for the iron deposition but could favor the progression of virus-induced damage independently from interference with iron metabolism. The topic of iron and its relationship to chronic viral hepatitis is reviewed in detail by Bonkovsky et al. (1996, 1999).

Patients with heavy alcohol consumption and genetic hemochromatosis (GH) have a higher prevalence of cirrhosis than those who do not. Ludwig et al. (1997) found that subjects with GH who drank more than 60 g of alcohol per day were approximately nine times more likely to develop cirrhosis than those who drank less than that amount. The range of the hepatic iron concentration associated with

Table 25.3 Causes of hepatic iron overload not related to genetic hemochromatosis

Etiology	Ref.
Hereditary anemias	
Sickle cell anemia	Searle et al. (2001)
Thalassemia major	
Sideroblastic anemia	
Hereditary spherocytosis	Barry et al. (1968)
Chronic liver disease	
Chronic hepatitis C	Hezode et al. (1999) and Ganne-Carrié et al. (2000)
Alcoholism and alcoholic liver disease	Kageyama et al. (2000)
Nonalcoholic steatohepatitis	George et al. (1998), Younossi et al. (1999) and Bonkovsky and Lambrecht (2000)
Cirrhosis or diverse etiology	Deugnier et al. (1997) and Ludwig et al. (1997)
Post-portacaval shunt	Bonkovsky et al. (1996)
Increased oral or parenteral iron loading	
African iron overload (dietary overload with genetic factor)	Gordeuk et al. (1986, 1992), Gangaidzo and Gordeuk (1995) and Mandishona et al. (1998)
African-American iron overload (similar to African iron overload)	Barton et al. (1995a, b)
Medicinal iron ingestion	Hennigar et al. (1979)
Transfusional siderosis	Oliver (1959) and Schafer et al. (1981)
Hemodialysis	Gokal et al. (1979)
Inherited metabolic diseases	
α-1 Antitrypsin deficiency with genetic hemochromatosis	Elzouki et al. (1995); Fargion et al. (1996)
Porphyria cutanea tarda	Bonkovsky and Lambrecht (2000)
Hereditary tyrosinemia	
Zellweger's syndrome	
Congenital atransferrinemia	Searle et al. (2001)
Aceruloplasminemia	Andrews (1999)
Wilson's disease, especially after penicillamine therapy	Shiono et al. (2001)
Neonatal hemochromatosis?	Bonkovsky et al. (1996)

cirrhosis (in the absence of other cofactors) was 237–675 μmol/g dry weight.

Iron overload in cirrhosis was studied in 447 native livers by Ludwig et al. (1997) and Deugnier et al. (1997). It is very common in many types of nonbiliary cirrhosis and appears to be acquired and to occur rapidly once cirrhosis develops (Ludwig et al. 1997). Quantitative iron determinations cannot be relied on to differentiate such cases from GH. Patients with nonalcoholic cirrhosis can accumulate iron (and zinc) after end-to-side portacaval shunting, but quantitative iron analyses have shown that such patients have far less iron than patients with GH, and they are unlikely to have tissue injury resulting from the iron overload (Adams et al. 1994).

Nonalcoholic steatohepatitis (NASH), now sometimes referred to as nonalcoholic fatty liver disease, may be accompanied by iron overload. In one study of 65 patients from the United States, iron accumulation, when present, was not associated with increased overall mortality, liver-related mortality, or cirrhosis (Younossi et al. 1999). Bonkovsky et al. (1999) found increased prevalence of HFE mutations in 57 subjects with NASH. Those with C282Y mutations had significantly higher levels of serum ALT and greater hepatic fibrosis than those without. In another study of 51 patients with NASH from Australia, the C282Y mutation was found to be responsible for most of the mild iron overload, and it was significantly associated with hepatic damage in these patients (George et al. 1998). It was suggested that phlebotomy therapy to remove the increased iron stores could delay or even reverse the liver damage.

African iron overload (formerly referred to as Bantu siderosis) is now rare in South Africa but remains a public health issue in sub-Saharan Africa (Gordeuk et al. 1986). It appears to be caused by an interaction between the amount of dietary iron (consumed in traditional beer brewed in steel drums) and a gene distinct from any HLA-linked gene (Gordeuk et al. 1992). African iron overload is considered a risk factor for hepatocellular carcinoma in black Africans (Gangaidzo and Gordeuk 1995; Mandishona et al. 1998). Iron overload in African-Americans (that is unexplained by dietary or medicinal iron excess, transfusions, or sideroblastic anemia) is considered to be similar to that in sub-Saharan Africans (Barton et al. 1995a, b).

Miscellaneous effects of iron overload that should be briefly mentioned include impaired cellular immune

function and increased susceptibility to bacterial infections (Bonkovsky et al. 1996), as well as hepatocellular carcinoma (HCC). The occurrence of HCC in patients with genetic hemochromatosis and cirrhosis is too well recognized to require comment, but iron may play a putative role in HCC that occurs rarely in a noncirrhotic liver (Turlin et al. 1995).

Acknowledgements The authors would like to express their appreciation to Drs. Elena Ladich, Todor Todorov, Leonor Martinez, Elizabeth Meze, and Linda Murakata for their valuable contributions during the preparation of this chapter.

Further Reading

Adams PC, Bradley C, Frei JV (1994) Hepatic iron and zinc concentrations after portacaval shunting for nonalcoholic cirrhosis. Hepatology 19:101–105

American Thoracic Society (1986) The diagnosis of nonmalignant diseases related to asbestos. Am Rev Respir Dis 134:363–368

Andrews N (1999) Disorders of iron metabolism. N Engl J Med 341:1886–1995

Barchowsky A, Dudek EJ, Treadwell MD, Wetterhahn KE (1996) Arsenic induces oxidant stress and NFKB activation in cultured aortic endothelial cells. Free Radic Biol Med 21:783–790

Barnhart S, Thornquist M, Omenn G et al (1990) The degree of roentgenographic parenchymal opacities attributable to smoking among asbestos-exposed subjects. Am Rev Respir Dis 141:1102–1106

Barrett JC, Lamb PW, Wang TC, Lee TC (1989) Mechanisms of arsenic-induced cell transformation. Biol Trace Elem Res 21:421–429

Barry M, Scheuer PJ, Sherlock S et al (1968) Hereditary spherocytosis with secondary haemochromatosis. Lancet 2:481–485

Barton AL, Banner BF, Cable EE et al (1995a) Distribution of iron in the liver predicts the response of chronic hepatitis C infection to interferon therapy. Am J Clin Pathol 103:419–424

Barton JC, Edwards CQ, Bertoli LF et al (1995b) Iron overload in African Americans. Am J Med 99:616–623

Becklake M (1983) Occupational lung disease: past record and future trend using the asbestos case as an example. Clin Invest Med 6:305–337

Belton JC, Benson NC, Hanna ML, Taylor RT (1985) Growth inhibition and cytotoxic effects of three arsenic compounds on cultured Chinese hamster ovary cells. J Environ Sci Health 20A:37–72

Bhagwat AG, Walia BN, Koshy A, Banerji CK (1983) Will the real Indian childhood cirrhosis please stand up? Cleve Clin Q 50:323–337

Bhusnurmath SR, Walia BNS, Singh S et al (1991) Sequential histopathologic alterations in Indian childhood cirrhosis treated with D-penicillamine. Hum Pathol 22:653–658

Blomfield J, Dixon SR, McCredie DA (1971) Potential hepatotoxicity of copper in recurrent hemodialysis. Arch Intern Med 128:555–560

Boffetta P, Sallsten G, Garcia-Gomez M, Pompe-Kirn V, Zaridze D, Bulbulyan M, Caballero JD, Ceccarelli F, Kobal AB, Merler E (2001) Mortality from cardiovascular diseases and exposure to inorganic mercury. Occup Environ Med 58:461–466

Bonkovsky HL, Lambrecht RW (2000) Iron-induced liver injury. Clin Liver Dis 4:409–429

Bonkovsky HL, Banner BF, Lambrecht RW, Rubin RB (1996) Iron in diseases other than hemochromatosis. Semin Liver Dis 16:65–82

Bonkovsky HL, Jawaid Q, Tortorelli K et al (1999) Non-alcoholic steatohepatitis and iron; increased prevalence of mutations of the HFE gene in non-alcoholic steatohepatitis. J Hepatol 31:421–429

Brain J, Valberg P (1979) Deposition of aerosol in the respiratory tract. Am Rev Respir Dis 120:1325–1373

Brenard R, Laterre PF, Reynaert M et al (1996) Increased hepatocytic mitotic activity as a diagnostic marker of acute arsenic intoxication. J Hepatol 25:218–220

Browne K (1986) A threshold for asbestos related lung cancer. Br J Ind Med 43:556–558

Browne K (1994) Asbestos-related disorders. In: Parkes W (ed) Occupational lung disorders. Butterworth-Heinemann, Oxford

Cavanagh JB, Nolan CC (1994) The neurotoxicity of organolead and organotin compounds. In: de Wolff FA (ed) Handbook of clinical neurology, vol 20. Elsevier, Amsterdam, pp 129–150

Cavigelli M, Li WW, Lin A, Su B, Yushioka K, Karin M (1996) The tumor promoter arsenite stimulates AP-1 activity by inhibiting a JNK phosphatase. EMBO J 15:6269–6279

Centeno JA, Pestaner JP, Mullick FG, Virmani R (1996) An analytical comparison of cobalt cardiomyopathy and idiopathic dilated cardiomyopathy. Biol Trace Elem Res 55:21–30

Chan P, Huff J (1997) Arsenic carcinogenesis in animals and in humans: mechanistic, experimental, and epidemiological evidence. Environ Carcinog Ecotoxicol Rev 15:83–122

Chen CJ, Chen CW, Wu MM, Kuo TL (1992) Cancer potential in liver, lung, bladder, and kidney due to ingested inorganic arsenic in drinking water. Br J Cancer 66:888–892

Choie DD, Richter GW (1972) Lead poisoning: rapid formation of intranuclear inclusions. Science 177(55):1194–1195

Chou HT, Yang HL, Tsou SS, Ho RK, Pai PY, Hsu HB (1998) Status of trace elements in patients with idiopathic dilated cardiomyopathy in central Taiwan. Chung Hua I Hsueh Tsa Chih (Taipei) 61:193–198

Chung C, Tsai R, Chen G, Yu H, Chai C (1998) Expression of bcl-2, p53 and Ki-67 in arsenical skin cancers. J Cutan Pathol 25:457–462

Churg A (1998a) Non-neoplastic disease caused by asbestos. In: Churg A, Green F (eds) Pathology of occupational lung disease. Williams & Wilkins, Baltimore

Churg A (1998b) Neoplastic asbestos-induced disease. In: Churg A, Green F (eds) Pathology of occupational lung disease. Williams & Wilkins, Baltimore

Churg A, Colby T (1998) Diseases caused by metals and related compounds. In: Churg A, Green F (eds) Pathology of occupational lung disease. Williams & Wilkins, Baltimore

Churg A, Warnock M (1981) Asbestos and other ferruginous bodies. Am J Pathol 102:447–456

Churg A, Wright J, Vedal S (1993) Fiber burden and patterns of asbestos-related disease in chrysotile miners and millers. Am Rev Respir Dis 148:25–31

Chuttani HK, Gupta PS, Gulati S, Gupta DN (1965) Acute copper sulfate poisoning. Am J Med 39:849–854

Craighead J, Abraham J, Churg A et al (1982) The pathology of asbestos-associated diseases of the lungs and pleural cavities: diagnostic criteria and proposed grading schema (report of the pneumoconiosis committee of the college of American pathologists and the national institute for occupational safety and health). Arch Pathol Lab Med 106:544–596

Datta DV, Mitra SK, Chhuttani PN, Chakravarti RN (1979) Chronic oral arsenic intoxication as a possible aetiological factor in idiopathic portal hypertension (non-cirrhotic portal fibrosis) in India. Gut 20:378–384

de Valk B, Marx JJ (1999) Iron, atherosclerosis, and ischemic heart disease. Arch Intern Med 159:1542–1548

Denk R, Holzmann H, Lange HJ, Greve D (1959) Uber Arsenspatschaden bei obduzierten Moselwinzern. Med Welt 11:557–567

Deugnier Y, Turlin B, Le Quilleue D et al (1997) A reappraisal of hepatic siderosis in patients with end-stage cirrhosis: practical

implications for the diagnosis of hemochromatosis. Am J Surg Pathol 21:669–675
1999 Diseases of immunity. In: Cotran R, Kumar V, Collins T (eds) Pathologic basis of disease. W. B. Saunders, Philadelphia, pp 188–259
Doll R (1993) Mortality from lung cancer in asbestos workers 1955 (classical article). Br J Ind Med 50:485–490
Dupres J, Mustard J, Uffen R (1984) Report of the royal commission on matters of health and safety arising from the use of asbestos in Ontario. Queen's Printer for Ontario, Toronto
Elis A, Froom P, Ninio A, Cahana L, Lishner M (2001) Employee exposure to chromium and plasma lipid oxidation. Int J Occup Environ Health 7:206–208
Elzouki A-NY, Hulterantz R, Stäl P et al (1995) Increased PiZ frequency for α-1 antitrypsin in patients with genetic hemochromatosis. Gut 36:922–926
Epler G, McLoud T, Gaensler E (1982) Prevalence and incidence of benign asbestos pleural effusion in a working population. JAMA 247:617–622
Espinoza EO, Mann MJ, Bleasell B (1995) Arsenic and mercury in traditional Chinese herbal balls (letter). N Engl J Med 333:803–804
Evans PJ, Smith C, Mitchinson MJ, Halliwell B (1995) Metal ion release from mechanically disrupted human arterial wall. Implications for the development of atherosclerosis. Free Radic Res 23:465–469
Falk H, Caldwell GG, Ishak KG et al (1981) Arsenic-related hepatic angiosarcoma. Am J Ind Med 2:43–50
Fargion S, Bissoli F, Fracanzani AL et al (1996) No association between genetic hemochromatosis and α-1 antitrypsin deficiency. Hepatology 24:1161–1164
Fraser R, Muller N, Colman N et al (1999a) Pleural effusion. In: Fraser and Pare's diagnosis of diseases of the chest. W. B. Saunders, Philadelphia
Fraser R, Muller N, Colman N et al (1999b) Inhalation of inorganic dust (pneumoconiosis). In: Fraser and Pare's diagnosis of diseases of the chest. W. B. Saunders, Philadelphia
Frustaci A, Magnavita N, Chimenti C, Caldarulo M, Sabbioni E, Pietra R, Cellini C, Possati GF, Maseri A (1999) Marked elevation of myocardial trace elements in idiopathic dilated cardiomyopathy compared with secondary cardiac dysfunction. J Am Coll Cardiol 33:1578–1583
Gandhi RK, Robarts FH (1962) Hourglass stricture of the stomach and pyloric stenosis due to ferrous sulphate poisoning. Br J Surg 49:613–617
Gangaidzo IT, Gordeuk VK (1995) Hepatocellular carcinoma and African iron overload. Gut 37:727–730
Ganne-Carrié N, Christidis C, Chastang C et al (2000) Liver iron is predictive of death in alcoholic cirrhosis: a multivariate study of 229 consecutive patients with alcoholic and/or hepatitis C virus cirrhosis: a prospective followup study. Gut 46:277–282
Ge K, Yang G (1993) The epidemiology of selenium deficiency in the etiological study of endemic diseases in china. Am J Clin Nutr 57:259S–263S
Ge K, Xue A, Bai J, Wang S (1983) Keshan disease—an endemic cardiomyopathy in China. Virchows Arch A Pathol Anat Histopathol 401:1–15
George DK, Goldwurm S, MacDonald GA et al (1998) Increased hepatic iron concentration in non-alcoholic steatohepatitis is associated with increased fibrosis. Gastroenterology 114:311–318
Ghadially FN (1979) Ultrastructural localization and *in situ* analysis of iron, bismuth, and gold inclusions. CRC Crit Rev Toxicol 6 (4):303–350
Gibbs A (1996) Occupational lung disease. In: Hasleton P (ed) Spencer's pathology of the lung. McGraw-Hill, New York
Gokal R, Millard PR, Weatherall DJ et al (1979) Iron metabolism in haemodialysis patients. Q J Med 48:369–391
Good PF et al (1992a) Selective accumulation of aluminum and iron in the neurofibrillary tangles of Alzheimer's disease: a laser microprobe (LAMMA) study. Ann Neurol 31:286–292
Good PF, Olanow CW, Perl DP (1992b) Neuromelanin-containing neurons of the substantia nigra accumulate iron and aluminum in Parkinson's disease: a LAMMA study. Brain Res 593:343–346
Gordeuk VR, Boyd RD, Brittenham GM (1986) Dietary iron overload in rural Sub-Saharan Africa. Lancet 1:1310–1313
Gordeuk V, Mukübi SJ, Hasstedt SJ et al (1992) Iron overload in Africa: interaction between a gene and dietary iron content. N Engl J Med 326:95–100
Greenberg S, Roggli V (1992) Carcinoma of the lung. In: Roggli V, Greenberg S, Pratt P (eds) Pathology of asbestos-associated disease. Little Brown & Company, Boston
Hall JC, Harruff R (1989) Fatal cardiac arrhythmia in a patient with interstitial myocarditis related to chronic arsenic poisoning. South Med J 82:1557–1560
Hanada K, Hashimoto I, Kon A, Kida K, Mita R (1998) Silver in sugar particles and systemic Argyria. Lancet 351:960
Haywood S, Müller T, Müller W et al (2001) Copper-associated liver disease in north ronaldsay sheep: a possible animal model for non-wilsonian hepatic copper toxicosis of infancy and childhood. J Pathol 195:264–269
Helwig E (1951) Chemical (beryllium) granulomas of the skin. Mil Surg 109:540–558
Hennigar GR, Greene WB, Walker EM et al (1979) Hemochromatosis caused by excessive vitamin iron intake. Am J Pathol 96:611–624
Hezode C, Cazeneuve C, Coué O et al (1999) Liver iron accumulation in patients with chronic active hepatitis C: prevalence and role of hemochromatosis gene mutations and relationship with hepatitis histological lesions. J Hepatol 31:979–984
Holtzman NA, Elliot DA, Heller RH (1986) Copper intoxication: report of a case with observations on ceruloplasmin. N Engl J Med 275:347–352
Horslen SP, Tanner MS, Lyon TDB et al (1994) Copper associated childhood cirrhosis. Gut 35:1497–1500
Hsu CH, Yang SA, Wang JY, Yu HS, Lin SR (1999) Mutational spectrum of p53 gene in arsenic-related skin cancers from the blackfoot disease endemic area of Taiwan. Br J Cancer 80:1080–1086
Hughes JM, Weill H (1991) Asbestosis as a precursor of asbestos related lung cancer: results of a prospective mortality study. Br J Ind Med 48:229–233
Hume L, Roe M (1992) The biodurability of chrysotile asbestos. Am Mineral 77:1125–1128
Hunter D, Russell DS (1954) Focal cerebral and cerebellar atrophy in a human subject due to organic mercury compounds. J Neurol Neurosurg Psychol 17:235–241
Hutchinson J (1888) On some examples of arsenic-keratosis of the skin and arsenic-cancer. Trans Pathol Soc Lond 39:352–363
Huttunen JK (1997) Selenium and cardiovascular diseases—an update. Biomed Environ Sci 10:220–226
Iribarren C, Sempos CT, Eckfeldt JH, Folsom AR (1998) Lack of association between ferritin level and measures of LDL oxidation: the ARIC study. Atherosclerosis risk in communities. Atherosclerosis 139:189–195
James JA (1970) Acute iron poisoning: assessment of severity and prognosis. J Pediatr 77:117–119
Jensen S, Jernelov A (1969) Biological methylation of mercury in aquatic organisms. Nature 223:753–756
Jhaveri SS (1959) A case of cirrhosis and primary carcinoma of the liver in chronic arsenical intoxication. Br J Ind Med 16:248–250
Jingbo PI, Hiroshi Y, Yoshito K, Guifan S, Takahiko Y, Hiroyuki A, Claudia HR, Nobuhiro S (2002) Evidence for induction of oxidative stress caused by chronic exposure of Chinese residents to arsenic

contained in drinking water. Environ Health Perspect 110 (4):331–336
Johnson RA, Baker SS, Fallon JT, Maynard EP 3rd, Ruskin JN, Wen Z, Ge K, Cohen HJ (1981) An occidental case of cardiomyopathy and selenium deficiency. N Engl J Med 304:1210–1212
Joshi VV (1987) Indian childhood cirrhosis. Perspect Paediatr Pathol 11:175–192
Kageyama F, Kobayashi Y, Kawasaki T et al (2000) Successful interferon therapy reverses enhanced hepatic iron accumulation and lipid peroxidation in chronic hepatitis C. Am J Gastroenterol 95:1041–1050
Kardinaal AF, Kok FJ, Kohlmeier L, Martin-Moreno JM, Ringstad J, Gomez-Aracena J, Mazaev VP, Thamm M, Martin BC, Aro A, Kark JD, Delgado-Rodriguez M, Riemersma RA, van't Veer P, Huttunen JK (1997) Association between toenail selenium and risk of acute myocardial infarction in European men. The EURAMIC study. European antioxidant myocardial infarction and breast cancer. Am J Epidemiol 145:373–379
Kesteloot H, Roelandt J, Willems J, Class JM, Joosens JV (1968) An inquiry into the role of cobalt in the heart disease of chronic beer drinkers. Circulation 37:854–864
Kew J, Morris C, Aihie A et al (1993) Arsenic and mercury intoxication due to Indian ethnic remedies. Br Med J 306:506–507
Kiechl S, Willeit J, Egger G, Poewe W, Oberhollenzer F (1997) Body iron stores and the risk of carotid atherosclerosis: prospective results from the Bruneck study. Circulation 96:3300–3307
Kitchin KT (2001) Recent advances in arsenic carcinogenesis: modes of action, animal model systems, and methylated arsenic metabolites. Toxicol Appl Pharmacol 172(3):249–261
Koch P, Bahmer FA (1999) Oral lesions and symptoms related to metals used in dental restorations: a clinical, allergological, and histologic study. J Am Acad Dermatol 42:422–430
Kopp SJ, Barron JT, Tow JP (1988) Cardiovascular actions of lead and relationship to hypertension: a review. Environ Health Perspect 78:91–99
Krenzelok EP, Hoff JV (1979) Accidental iron poisoning: a problem of marketing and labeling. Pediatrics 63:591–596
Labadie H, Stoessel P, Callard P, Beaugrand M (1990) Hepatic veno-occlusive disease and perisinusoidal fibrosis secondary to arsenic poisoning. Gastroenterology 99:1140–1143
Lansdown A (1995) Physiological and toxicological changes in the skin resulting from the action and interaction of metal ions. Crit Rev Toxicol 25:397–462
Lefkowitch JH, Honig CL, King ME, Hagstrom JWC (1982) Hepatic copper overload and features of Indian childhood cirrhosis in an American sibship. N Engl J Med 307:271–277
Li Y, Nan BS (1989) Correlation of selenium, glutathione peroxidase activity and lipoperoxidation rates in dilated cardiomyopathy. Chin Med J 102:670–671
Li JH, Rossman TC (1989) Inhibition of DNA ligase activity by arsenite: a possible mechanism of its comutagenesis. Mol Toxicol 2:1–9
Li GS, Wang F, Kang D, Li C (1985) Keshan disease: an endemic cardiomyopathy in China. Hum Pathol 16:602–609
Ludwig J, Farr GH, Freese DK, Sternlieb I (1995) Chronic hepatitis and hepatic failure in a 14-year-old girl. Hepatology 22:1874–1879
Ludwig J, Hashimoto E, Poroyko MD et al (1997) Hemosiderosis in cirrhosis: a study of 447 native livers. Gastroenterology 112:888–890
Lunchtrath H (1972) Cirrhosis of the liver in chronic arsenical poisoning of vintners. Ger Med Mon 2:127–128
Maggiore G, De Giacomo C, Sessa F, Burgio GR (1987) Idiopathic hepatic copper toxicosis in a child. J Pediatr Gastroenterol Nutr 6:980–983
Maheswaran R, Morris S, Falconer S, Grossinho A, Perry I, Wakefield J, Elliott P (1999) Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England. Heart 82:455–460
Mahon NG, Coonar AS, Jeffery S, Coccolo F, Akiyu J, Zal B, Houlston R, Levin GE, Baboonian C, McKenna WJ (2000) Haemochromatosis gene mutations in idiopathic dilated cardiomyopathy. Heart 84:541–547
Maloney M (1996) Arsenic in dermatology. Dermatol Surg 22:301–304
Mandishona E, MacPhail AP, Gordeuk VR et al (1998) Dietary iron overload as a risk factor for hepatocellular carcinoma in black Africans. Hepatology 27:1563–1566
Markesbery WR (1997) Oxidative stress hypothesis in Alzheimer's disease. Free Radic Biol Med 23:134–147
Marsh D (1979) Organic mercury: methylmercury compounds. In: Vinken PJ, Bryun GW (eds) Handbook of clinical neurology, vol 36. North-Holland, Amsterdam, pp 73–81
Mazumder DNG, Chakraborty AK, Ghosh A et al (1988) Chronic arsenic toxicity from drinking tube well water in rural West Bengal. Bull WHO 66:499–506
McDonald JC (1980) Asbestos and lung cancer: has the case been proven? Chest 78:374–376
McDonald A, McDonald J (1987) Epidemiology of malignant mesothelioma. In: Antman K, Aisner J (eds) Asbestos-related malignancy. Harcourt Brace Jovanovich, Orlando
Mena I (1979) Manganese poisoning. In: Vinken PJ, Bryun GW (eds) Handbook of clinical neurology, vol 36. North-Holland, Amsterdam, pp 217–237
Mena I et al (1967) Chronic manganese poisoning: clinical picture and manganese turnover. Neurology 17:128–136
Mendis S (1989) Magnesium, zinc, and manganese in atherosclerosis of the aorta. Biol Trace Elem Res 22:251–256
Meyers DG (1996) The iron hypothesis—does iron cause atherosclerosis? Clin Cardiol 19:925–929
Modell CB (1975) Transfusional haemochromatosis. In: Kief H (ed) Iron metabolism and its disorders. Excerpta Medica, Amsterdam, pp 230–239
Monoguerra AS (1976) Iron poisoning: report of a fatal case in an adult. Am J Hosp Pharmacol 33:1088–1090
Morgan A (1980) Effect of length on the clearance of fibers from the lung and on body formation. In: Wagner J (ed) Biologic effects of mineral fibers. International Agency for Research on Cancer, Lyon
Morgan A, Holmes A (1980) Concentrations and dimensions of coated and uncoated asbestos fibres in the human lung. Br J Ind Med 37:25–32
Morgan A, Holmes A (1985) The enigmatic asbestos body: its formation and significance in asbestos-related disease. Environ Res 38:283–292
Morgan W, Seaton A (1975) Occupational lung diseases. W. B. Saunders, Philadelphia
Muehrcke RC, Pirani CL (1968) Arsine-induced anuria: a correlative clinicopathological study with electron microscopic observations. Ann Intern Med 68(4):853–866
Müller T, Feichinger H, Berger H, Muller W (1996) Endemic Tyrolean infantile cirrhosis: an ecogenetic disease disorder. Lancet 347:877–880
Müller-Höcker J, Meyer U, Wiebecke B, Hübner G (1985) Copper storage disease of the liver and chronic dietary copper intoxication in two further German infants mimicking Indian childhood cirrhosis. Pathol Res Pract 183:39–45
Müller-Höcker J, Weib M, Meyer U et al (1987) Fatal copper storage disease of the liver in a German infant resembling Indian childhood cirrhosis. Virchows Arch (A) 411:379–385
Murphy G (1995) Skin. In: Craighead J (ed) Pathology of environmental and occupational disease. Mosby, St. Louis, pp 437–453

Murphy M, Hunt S, McDonald GSA et al (1991) Intrahepatic cholestasis secondary to gold therapy. Eur J Gastroenterol Hepatol 3:855–859

Nakamuro K, Sayato Y (1981) Comparative studies of chromosomal aberration induced by trivalent and pentavalent arsenic. Mutat Res 88:73–80

Narang R, Ridout D, Nonis C, Kooner JS (1997) Serum calcium, phosphorus and albumin levels in relation to the angiographic severity of coronary artery disease. Int J Cardiol 60:73–79

Navarro-Alarcon M, Lopez-Garcia de la Serrana H, Perez-Valero V, Lopez-Martinez C (1999) Serum and urine selenium concentrations in patients with cardiovascular diseases and relationship to other nutritional indexes. Ann Nutr Metab 43:30–36

Needleman HL (1988) The persistent threat of lead: medical and sociological issues. Curr Probl Pediatr 18:697–744

Needleman HL, Gunnoe C, Leviton A (1979) Deficits in psychologic and classroom performance in children with elevated dentine lead levels. N Engl J Med 300:689–695

Needleman HL, Schell A, Bellinger D (1990) The long-term effects of exposure to low doses of lead in childhood. An 11th year follow-up report. N Engl J Med 322:83–88

Neve J (1996) Selenium as a risk factor for cardiovascular diseases. J Cardiovasc Risk 3:42–47

Newland MC et al (1989) Visualizing manganese in the primate basal ganglia with magnetic resonance imaging. Eur Neurol 11106:251–258

Nightingale SL (1997) Action to prevent accidental iron poisonings in children. JAMA 277:1343

O'Neill NC, Tanner MS (1989) Uptake of copper from brass vessels in bovine milk and its relevance to Indian childhood cirrhosis. J Pediatr Gastroenterol Nutr 9:167–172

Olanow CW et al (1994) Manganese-induced neurotoxicity. In: Mizuno Y, Calne DB, Horowski R (eds) Advances in research on neurodegeneration. Vol II. Etiopathogenesis. Birkhauser, Boston, pp 53–62

Olanow CW et al (1996) Manganese intoxication in the rhesus monkey: a clinical, imaging, pathologic, and biochemical study. Neurology 46:492–498

Oliver RAM (1959) Siderosis following transfusions of blood. J Pathol Bacteriol 77:171–194

Oster O (1993) Trace element concentrations (Cu, Zn, Fe) in sera from patients with dilated cardiomyopathy. Clin Chim Acta 214:209–218

Oster O, Prellwitz W (1990) Selenium and cardiovascular disease. Biol Trace Elem Res 24:91–103

Patrizi A, Rizzoli L, Vincenzi C, Trevisi P, Tosti A (1999) Sensitization to thimerosal in atopic children. Contact Dermat 40:94–97

Perl DP, Brody AR (1980) Alzheimer's disease: X-ray spectrographic evidence of aluminum accumulation in neurofibrillary tangle-bearing neurons. Science 208:297–299

Perl DP, Good PF (1992) Aluminum and the neurofibrillary tangle: results of tissue microprobe studies. In: Williams RMP (ed) Ciba symposium: aluminum in biology and medicine. Wiley, New York, pp 217–236

Pestaner JP, Ishak KG, Mullick FG, Centeno J (1999) Ferrous sulfate toxicity: a review of autopsy findings. Biol Trace Elem Res 70:1–8

Pimental JC, Menezes AP (1977) Liver disease in vineyard sprayers. Gastroenterology 72:275–289

Pirisi M, Scott CA, Avellini C et al (2000) Iron deposition and progression of disease in chronic hepatitis C. Role of interface hepatitis, portal inflammation, and HFE missense mutations. Am J Clin Pathol 113:546–554

Raines DA, Kinsara AJ, Eid Fawzy M, Vasudevan S, Mohamed GE, Legayada ES, Al-Rawithi S, El-Yazigi A (1999) Plasma and urinary selenium in Saudi Arabian patients with dilated cardiomyopathy. Biol Trace Elem Res 69:59–68

Ramirez P, Eastmond DA, Laclette JP, Ostrosky-Wegman P (1997) Disruption of microtubule assembly and spindle formation for the induction of aneuploid cells by sodium arsenite and vanadium pentoxide. Mutat Res 386:291–298

Rauramaa R, Vaisanen S, Mercuri M, Rankinen T, Penttila I, Bond MG (1994) Association of risk factors and body iron status to carotid atherosclerosis in middle-aged eastern Finnish men. Eur Heart J 15:1020–1027

Rayman MP (2000) The importance of selenium to human health. Lancet 356:233–241

Reusche E, Koch V, Friedrich HJ, Nunninghoff D, Stein P, Rob PM (1996) Correlation of drug-related aluminum intake and dialysis treatment with deposition of argyrophilic aluminum-containing inclusions in CNS and in organ systems of patients with dialysis-associated encephalopathy. Clin Neuropathol 15:342–347

Riggio O, Montagnese F, Fiore P et al (1997) Iron overload in patients with chronic viral hepatitis: how common is it? Am J Gastroenterol 92:1298–1301

Roberts G (1971) The pathology of parietal pleural plaques. J Clin Pathol 24:348–353

Roggli V (1989) Pathology of human asbestosis: a critical review. In: Fenoglio-Preiser C (ed) Advances in pathology. Year Book Medical Publishers, Chicago

Roggli V (1992) Asbestos bodies and nonasbestos ferruginous bodies. In: Roggli V, Greenberg S, Pratt P (eds) Pathology of asbestos-related disease. Little Brown & Company, Boston

Roggli V (1994) The pneumoconioses: asbestosis. In: Saldana M (ed) Pathology of pulmonary disease. J. B. Lippincott, Philadelphia

Roggli V, Pratt P (1983) Numbers of asbestos bodies on iron-stained sections in relation to asbestos body counts in lung tissue digests. Hum Pathol 14:355–361

Roggli V, Pratt P (1992) Asbestosis. In: Roggli V, Greenberg S, Pratt P (eds) Pathology of asbestos-associated diseases. Little Brown & Company, Boston

Roggli V, Shelburne J (1994) Pneumoconioses, mineral and vegetable. In: Dail D, Hammar S (eds) Pulmonary pathology. Springer, New York

Rosenberg HG (1974) Systemic arterial disease and chronic arsenicism in infants. Arch Pathol 97:360–365

Rubenowitz E, Axelsson G, Rylander R (1999) Magnesium and calcium in drinking water and death from acute myocardial infarction in women. Epidemiology 10:31–36

Russell M, Langley M, Truett A, King L, Boyd A (1997) Lichenoid dermatitis after consumption of gold-containing liquor. J Am Acad Dermatol 36:841–844

Salonen JT, Alfthan G, Huttunen JK, Pikkarainen J, Puska P (1982) Association between cardiovascular death and myocardial infarction and serum selenium in a matched-pair longitudinal study. Lancet 2:175–179

Salonen JT, Salonen R, Seppanen K, Kantola M, Parviainen M, Alfthan G, Maenpaa PH, Taskinen E, Rauramaa R (1988) Relationship of serum selenium and antioxidants to plasma lipoproteins, platelet aggregability and prevalent ischaemic heart disease in eastern Finnish men. Atherosclerosis 70:155–160

Salonen JT, Seppanen K, Lakka TA, Salonen R, Kaplan GA (2000) Mercury accumulation and accelerated progression of carotid atherosclerosis: a population-based prospective 4-year follow-up study in men in eastern Finland. Atherosclerosis 148:265–273

Salvini S, Hennekens CH, Morris JS, Willett WC, Stampfer MJ (1995) Plasma levels of the antioxidant selenium and risk of myocardial infarction among U.S. Physicians. Am J Cardiol 76:1218–1221

Sanchez-Fructuoso AI, Torralbo A, Arroyo M, Luque M, Ruilope LM, Santos JL, Cruceyra A, Barrientos A (1996) Occult lead intoxication as a cause of hypertension and renal failure. Nephrol Dial Transplant 11:1775–1780

Schafer AI, Cheron RG, Dluchy R et al (1981) Clinical consequences of acquired transfusional iron overload in adults. N Engl J Med 304:319–324

Scheinberg IH, Sternlieb I (1994) Is non-Indian childhood cirrhosis caused by excess dietary copper? Lancet 344:1002–1004

Schoolmeester WL, White DR (1980) Arsenic poisoning. South Med J 73:198

Schramel P, Müller-Höcker J, Meyer U et al (1988) Nutritional copper intoxication in three carcinoma infants with severe liver cell damage (features of Indian childhood cirrhosis). J Trace Elem Electrolyte Health Dis 2:85–89

Schumann K (2001) Safety aspects of iron in food. Ann Nutr Metab 45:91–101

Searle J, Leggett BA, Crawford DHG, Powell LW (2001) Iron storage disease. In: MacSween RNM, Burt AD, Portmann BC, Ishak KG, Scheuer PJ, Anthony PP (eds) Pathology of the liver. Churchill Livingstone, London, pp 257–272

Selikoff IJ, Hammond EC, Seidman H (1980) Latency of asbestos disease among insulation workers in the United States and Canada. Cancer 46:2736–2740

Sempos CT, Looker AC, Gillum RE, McGee DL, Vuong CV, Johnson CL (2000) Serum ferritin and death from all causes and cardiovascular disease: the NHANES II mortality study. National health and nutrition examination study. Ann Epidemiol 10:441–448

Shiono Y, Wakusawa S, Hayashi H et al (2001) Iron accumulation in the liver of male patients with Wilson's disease. Am J Gastroenterol 96:3147–3151

Shiraki H (1979) Neuropathological aspects of organic mercury intoxication, including Minamata disease. In: Vinken PJ, Bryun GW (eds) Handbook of clinical neurology, vol 36. North-Holland, Amsterdam, pp 83–145

Simeonova PP, Luster MI (2000) Mechanisms of arsenic carcinogenicity: genetic or epigenetic mechanisms. J Environ Pathol Toxicol Oncol 19:281–286

Sluis-Cremer GK, Bezuidenhout BN (1989) Relation between asbestosis and bronchial cancer in amphibole asbestos miners (see comments). Br J Ind Med 46:537–540

Smetana HF, Hadley GG, Sirsat SM (1961) Infantile cirrhosis. An analytic review of the literature and a report of 50 cases. Pediatrics 28:107–127

Sokas RK, Simmens S, Sophar K, Welch LS, Liziewski T (1997) Lead levels in Maryland construction workers. Am J Ind Med 31:188–194

Spitainy KC, Brondum J, Vogt RL et al (1984) Drinking-water-induced copper intoxication in a Vermont family. Pediatrics 74:1103–1106

Staessen JA, Bulpitt CJ, Fagard R, Lauwerys RR, Roels H, Thijs L, Amery A (1994) Hypertension caused by low-level lead exposure: myth or fact? J Cardiovasc Risk 1:87–97

Staessen JA, Roels H, Fagard R (1996) Lead exposure and conventional and ambulatory blood pressure: a prospective population study. PheeCad investigators. J Am Med Assoc 275:1563–1570

Styblo M, Drobna Z, Jaspers I, Lin S, Thomas DJ (2002) The role of biomethylation in toxicity and carcinogenicity of arsenic: a research update. Environ Health Perspect 110(5):767–771

Suadicani P, Hein HO, Gyntelberg F (1992) Serum selenium concentration and risk of ischaemic heart disease in a prospective cohort study of 3000 males. Atherosclerosis 96:33–42

Tanner MS, Kantarjian AH, Bhave SA, Pandit AN (1983) Early introduction of copper-contaminated animal milk feed as a possible cause of Indian childhood cirrhosis. Lancet 2:992–995

Tanner MS, Bhave SA, Prodham AM, Pandit AN (1987) Clinical trials of penicillamine in Indian childhood cirrhosis. Arch Dis Child 62:118–1124

Tchounwou PB, Wilson BA, Schneider J, Ishaque A (2000) Cytogenetic assessment of arsenic trioxide toxicity in the mutatox, Ames II, and CAT-Tox assays. Met Ions Biol Med 6:89–91

Tchounwou PB, Wilson BA, Ishaque A, Schneider J (2001) Atrazine potentiation of arsenic trioxide-induced cytotoxicity and gene expression in human liver carcinoma cells (HepG2). Mol Cell Biochem 222:49–259

Timbrell V, Pooley F, Wagner J (1970) Characteristics of respirable asbestos fibers. In: Shaprio H (ed) Pneumoconioses: proceedings of the international conference, Johannesburg. Oxford University Press, London

Tsai S, Wang T, Ko Y (1999) Mortality for certain diseases in areas with high levels of arsenic in drinking water. Arch Environ Health 54:186–193

Turlin B, Juguet F, Moirand R et al (1995) Increased liver iron stores in patients with hepatocellular carcinoma developed on a noncirrhotic liver. Hepatology 22:446–450

Van Glahn WC, Flinn FB, Keim WF (1938) Effect of certain arsenates on the liver. Arch Pathol 25:488–505

Van Heck E, Kint A, Temmerman L (1981) A lichenoid eruption induced by penicillamine. Arch Dermatol 117:676

Vijaya J, Subramanyam G, Sukhaveni V, Abdul Latheef SA, Gupta SR, Sadhasivaiah G, Salam NM (2000) Selenium levels in dilated cardiomyopathy. J Indian Med Assoc 98:166–169

Virtamo J, Valkeila E, Alfthan G, Punsar S, Huttunen JK, Karvonen MJ (1985) Serum selenium and the risk of coronary heart disease and stroke. Am J Epidemiol 122:276–282

Wagner J, Newhouse M, Corrin B et al (1988) Correlation between fibre content of the lung and disease in east London factory workers. Br J Indust Med 45:305–308

Wei M, Wanibuchi H, Morimura K, Iwai S, Yoshida K, Endo G, Nakae D, Fukushima S (2002) Carcinogenicity of dimethylarsinic acid in male F344 rats and genetic alterations in induced urinary bladder tumors. Carcinogenesis 23(8):1387–1397

Westlin WF (1966) Deferoxamine in the treatment of acute iron poisoning. Clinical experiences with 172 children. Clin Pediatr 5:531

Wright J, Churg A (1998) Diseases caused by gases and fumes. In: Churg A, Green F (eds) Pathology of occupational lung disease. Williams & Wilkins, Baltimore

Yang FY, Lin ZH, Li SG, Guo BQ, Yin YS (1988) Keshan disease—an endemic mitochondrial cardiomyopathy in China. J Trace Elem Electrolyte Health Dis 2:157–163

Yegin A, Yegin H, Aliciguzel Y, Deger N, Semiz E (1997) Erythrocyte selenium–glutathione peroxidase activity is lower in patients with coronary atherosclerosis. Jpn Heart J 38:793–798

Yelin G, Taff ML, Sadowski GE (1987) Copper toxicity following massive ingestion of coins. Am J Forensic Med Pathol 8:78–85

Younossi ZM, Gramlich T, Bacon BR et al (1999) Hepatic iron and nonalcoholic fatty liver disease. Hepatology 30:847–850

Zenker F (1867) Iron lung: sclerosis pulmonum. Dtsch Arch Klin Med 2:116

Zielhuis R (1977) Public health risks of asbestosis. Pergamon Press, Elmsford

Toxicology

26

Tee L. Guidotti

Contents

Geosciences and chemistry and the scientific discipline of toxicology itself have been on parallel, often intertwined paths for many years. Issues related to toxic substances from natural sources, such as arsenic, lead and other metals, and from the contamination of soil and groundwater, have been recognized from the historical beginnings of the discipline.

Toxicology obviously plays a central role in medical geology. The scientific principles of toxicology apply to medical geology in three broad areas: clinical toxicology, risk assessment, and hazard control and monitoring. Clinical toxicology is the recognition, diagnosis, and management of human toxicity and in environmental medicine reflects the outcome of environmental chemical exposures. Toxicology plays an essential role in risk assessment both in characterizing the potential toxicity of a chemical hazard, the first step in the process, and in providing the conceptual framework upon which an estimate of safe exposure is derived for regulatory purposes. People may encounter chemical hazards in the course of daily life or in their jobs that are natural or that result form some disturbance of naturally-occurring hazards. Hazard control is more in the domain of occupational and environmental health and is necessarily based on an understanding of the physicochemical characteristics of the chemical hazard. Here toxicology provides the essential information needed to design a control system and to set priorities for control.

Toxicology engages other sciences as well as "medical geology", pharmacology, medicine, and chemistry. Toxicologists often work closely with epidemiologists, who use statistical methods to determine the distribution of diseases, their risk factors, and health characteristics of populations rather than individuals. Because the delineation of "safe" levels of exposures assumes a socially determined level of acceptable risk (which is implicit in the definition of "safety""), toxicology has been adapted, together with epidemiology, and applied in the form of "risk assessment" to

T.L. Guidotti (✉)
Medical Advisory Services, 1700 Research Blvd Street 240, Rockville, MD 20850, USA
e-mail: tee.guidotti@gmail.com

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_26,

provide guidance to regulatory bodies. Risk assessment, as the term is used here, is the identification and characterization of the level of risks resulting from exposure to hazards, including the uncertainties.

Toxicology has a long and colorful history. Historically, toxicology developed in Europe as a forensic science, because poisons were used in "trials by ordeal" (God was thought to protect the innocent from the effects of poison), for executions, and, notoriously, for murder. It later became a subdiscipline of pharmacology, the science of medications, as the mechanisms of drug effects were elucidated. Many important drugs have been derived from classical toxins and toxic agents were extensively used in the early modern days of physiology and neuroscience to identify basic mechanisms of the body. From its early preoccupation with particularly toxic chemicals, from which it gained its essential definition as the science of poisons, toxicology has expanded its scope to include biological mechanisms of toxicity and host defences (or "resistance") against toxicity. In the twentieth century, momentum for its development as an independent discipline has come (in roughly historical order) from food safety, chemical warfare, defense, product safety (especially cosmetics and food additives but also industrial chemicals), radiation biology, pesticide research, concern for environmental quality, environmental medicine, recent refinements in methodology of epidemiology and risk assessment, materials science and biocompatibility, molecular genetics and carcinogenesis research, and immunology. In recent years, toxicology has become highly specialized in the area of risk assessment, identifying the level of hazard associated with a particular chemical exposure and the limits of acceptably safe exposure. These issues go far beyond characterizing the effects of poisons, since most of the chemicals of modern concern are not classically "poisons", in the sense of being potentially lethal at low doses.

For convenience in terminology, toxicologists often refer to all substances not normally present in the body and which are introduced from an outside source as "xenobiotics" (from the Greek xeno-, meaning foreign). Xenobiotics may be drugs, food constituents, natural chemical exposures, or anthropogenic environmental chemical exposures.

In this Chapter, the basic principles of toxicology will be presented briefly, followed by a general framework for clinical toxicology, a general framework for toxicology as applied to risk assessment and to hazard control, and finally a longer discussion of the interpretation of trace element analysis, which is a practical matter of great concern in medical geology. The science of toxicology can be divided into toxicokinetics, the study and description of how xenobiotics enter and are handled by the body, and toxicdynamics, the study and description of what the xenobiotic does to the body. (see also Chaps. 9, 22and 24, this volume).

26.1 Toxicokinetics

Regardless of their effect or origin, the behavior of xenobiotics in the body can be described by accepted terms and general models reflecting the mechanisms by which exposure occurs and of how the body handles the chemical, which collectively is called "toxicokinetics". From the standpoint of evolutionary biology, it is supposed that most of these mechanisms developed in response to natural selection pressures reflecting either of two biological needs: to detoxify and excrete harmful substances ingested in foods (especially in spoiled or putrefied foodstuffs) and to metabolize endogenous chemical compounds (such as steroid hormones).

Toxicokinetics is the toxicological analogy to pharmacokinetics and is based on identical concepts. Four terms describe the disposition of xenobiotics, whether a drug or a toxic agent: absorption, distribution, metabolism, and excretion. Modelled together, the terms describe the entry, local and overall accumulation, transformation and removal from the body of the xenobiotic. Because tissue levels depend on transport of the xenobiotic to the target organ and the degree to which the xenobiotic partitions or is sequestered into the tissue, the kinetics of the xenobiotic determines the presentation of the xenobiotic or its metabolite to the target organ at the receptor level, where the toxic effect occurs. Figure 26.1 is an illustration of the principles of toxicokinetics.

26.1.1 Absorption

Xenobiotics may enter the body through any of several "portals" or routes of entry. In environmental medicine, the most significant portals of entry include ingestion, absorption through skin and inhalation. Ingestion results from eating contaminated foods or placing objects such as cigarettes in the mouth in a situation where the object or the hands may have been contaminated. Splashes into the eyes are more often associated with local eye irritation and only rarely with absorption and systemic toxicity. Other routes of exposure, such as intravenous infusion or implantation, are artificial and seldom seen outside of medical care and experimental studies.

The toxicity of the xenobiotic may or may not involve the organ of first contact or site of entry. For example, carbon monoxide enters the body by the inhalation route but causes no toxicity whatever to the lung. Other chemicals may cause local toxicity without significant absorption into the body, such as strong irritants applied to the skin. These routes of entry are not mutually exclusive. Inhalation of poorly

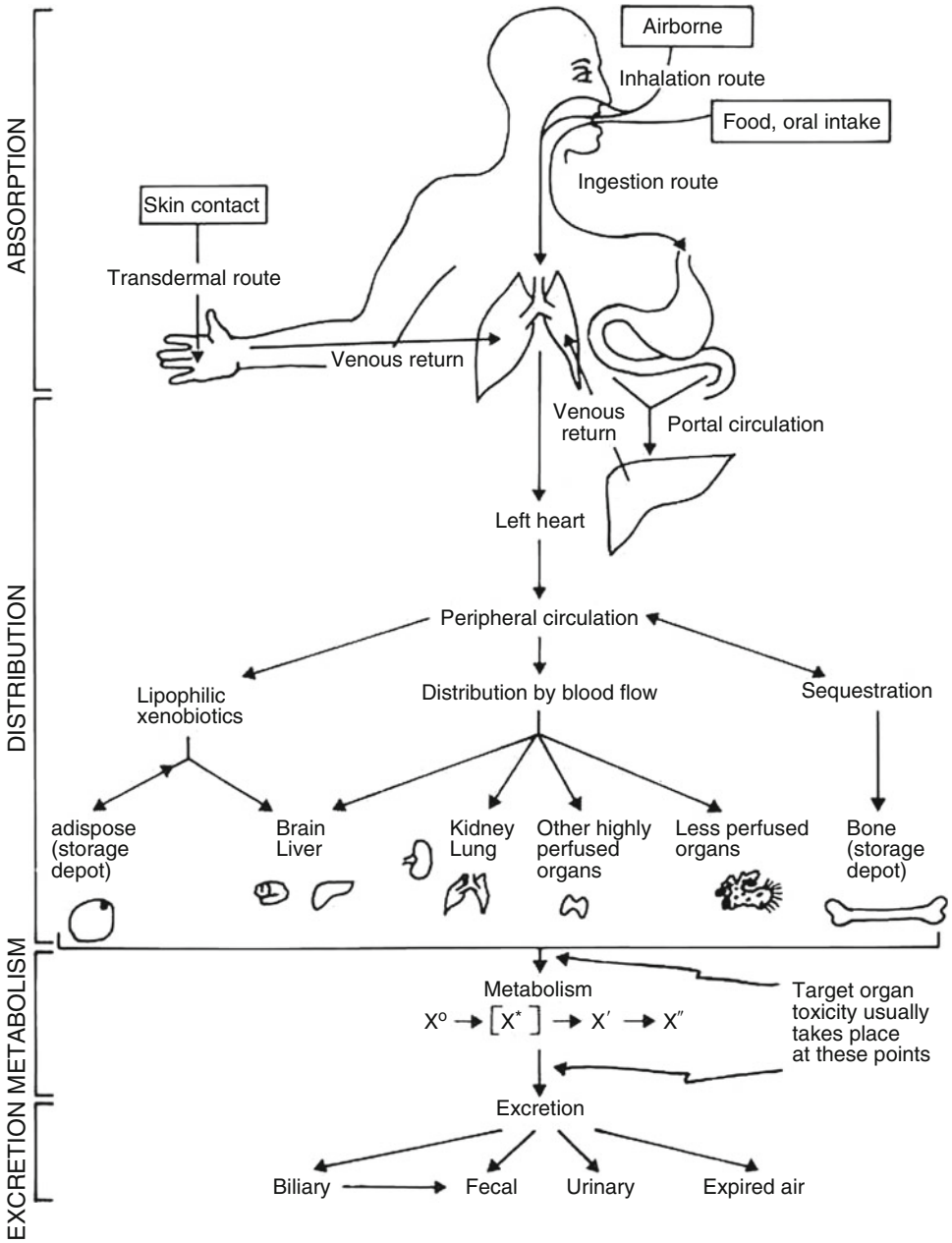


Fig. 26.1 Principles of toxicokinetics: absorption, distribution, metabolism, excretion (Reproduced, with permission, from: Guidotti (1994). By permission of Elsevier/Mosby)

soluble dusts such as silica, for example, may result in ingestion of the same material because of clearance from the lung bringing the material up the primary particle clearance mechanism of the lung, called the mucociliary escalator to the throat, where it is swallowed or expectorated.

The rate at which a xenobiotic enters the bloodstream is determined by absorption across the barrier presented by the given route of exposure. Absorption of xenobiotics across membranes is determined for the most part by the chemical and physical properties of the agent. In general, lipid-soluble (lipophilic) substances are absorbed more readily than water-soluble substances across barriers such as skin. The rate of absorption is the most important determinant of the peak levels that will be reached in plasma. For many toxic substances, this is the prime determinant of acute toxicity.

The skin is sufficiently permeable to be a major route of entry of many chemicals into the body, particularly those that are readily lipid-soluble. Absorption across the skin is highly variable, depending on skin characteristics and the solubility of the xenobiotic in fat. Most transdermal absorption occurs directly across the superficial layers of the skin, the stratum corneum, which consists of nonliving, keratinized cells, and

the other living cell layers of the epidermis, to be absorbed in the capillary bed of the dermis. Some chemicals applied to the skin may gain entry through a short cut, passing more rapidly through hair follicles and sebaceous gland ducts. When the skin is injured with open wounds or abrasion, or in the presence of a skin rash, absorption across the skin is much faster. Transcutanenous absorption is a problem in the toxicology of pesticides, solvents, and halogenated hydrocarbons generally. Some agents may be significantly metabolized by enzyme systems in the skin, but most gain entry into the bloodstream unchanged.

Exposure by inhalation is relatively efficient absorption and the lung itself is vulnerable to damage from inhaled xenobiotics. The lungs are the organ of gas exchange and are positioned in the circulation just before the heart. The organ receives venous blood from the body, oxygenates it, and returns it to the heart which pumps it out via arteries. Thus, blood reaching the lungs is initially low in oxygen and consists of mixed blood from many tissues, but the oxygen tension in lung tissue itself is very high, which makes the organ susceptible to damage-causing free radicals derived from oxygen. Exposure by inhalation results in relatively efficient absorption of gases if the gas can penetrate to the deep lung where gas exchange occurs, the alveolar-capillary bed. Whether the gas will penetrate efficiently depends on its solubility in water, reflecting clearance rates in the bronchial tree. Once having penetrated to the alveolar level, however, gases are readily absorbed into the blood stream across the alveolar-capillary membrane by simple passive diffusion. Absorption across the alveolar membrane in the lung is usually very efficient and complete and entry into the bloodstream is limited only by the xenobiotic's solubility in plasma, an aqueous medium. Particles, on the other hand, are subject to a number of host defense mechanisms in the respiratory tract that limit the efficiency of penetration to the alveolar level. Once there, their size prevents them from passing directly into the bloodstream and they must dissolve or be digested by macrophages (scavenging cells) before their constituent chemical contents can be absorbed and enter the bloodstream. Particles may contribute to systemic toxicity if they are composed of a soluble material, such as lead or polycyclic aromatic hydrocarbons. For this reason, inhalation of toxic gases is usually associated with acute systemic toxicity or vascular injury to the lung (resulting in pulmonary edema, a very dangerous condition of fluid accumulation in the lungs). Particle deposition in the lung is usually associated with localized pulmonary effects and chronic systemic toxicity if the particle is absorbed. Increasingly, inhalation toxicologists are learning about the effects of nanomaterials, both natural and synthetic. Particles at nanoscale (on the order of 100 nm or so) have every different characteristics than larger pieces of matter of the same composition and how they affect the lung and body cannot be extrapolated from the physical and chemical characteristics of the bulk material.

Ingestion is an important route of exposure for water and food and sometimes soil, and is therefore of interest in medical geology. Absorption through the gastrointestinal (GI) tract depends for many organic compounds on pH (because passage is increased when they are in a nonionized state) and therefore on location in the GI tract: the stomach is acid and the small intestine is basic. There are specialized transport mechanisms in the GI tract, among them facilitated diffusion to absorb glucose and a divalent-metal ion transporter that increases absorption of metals such as calcium and iron, as well as electrochemically similar ions. The GI route of exposure is unique in another important respect. Absorbed xenobiotics do not pass directly into the systemic circulation, as they do by transcutaneous and inhalation exposure, to be returned to the heart (via the lungs). Rather, veins draining the stomach and intestine conduct the blood to the liver by a specialized circuit (the "portal circulation"). The liver then metabolizes many xenobiotics before they pass into the systemic circulation and stores many xenobiotics. The veins draining the liver conduct blood to the main vein of the lower body and into the systemic circulation. Thus, when a xenobiotic is ingested, it may produce a toxic effect on the GI tract, it may produce a toxic effect on the liver, it may be metabolized, sometimes to a more toxic product, and it may pass in an altered form into the general circulation. (see also Chaps. 6, 9 and 25, this volume.

26.1.2 Distribution

Once the xenobiotic is absorbed and enters the bloodstream, it is transported to the capillary level in tissues of the body where it becomes available for uptake by the target organ. After one pass through the circulation the xenobiotic is uniformly mixed in arterial blood regardless of its route of entry. When a bolus is absorbed, the peripheral tissues are therefore presented with an increasing concentration in the blood which peaks and then declines as the xenobiotic is distributed to tissues throughout the body and removed by metabolism, excretion, or storage.

When a xenobiotic is dissolved in plasma, some fraction of the total usually binds to circulating proteins, particularly albumin (an abundant, soluble protein which binds many organic compounds as well as calcium, copper and zinc). Metals may also be bound to specialized proteins in the plasma, such as ceruloplasmin (copper) and transferrin (iron. Binding occurs quickly and an equilibrium is established between the fraction of the xenobiotic bound to plasma protein, which cannot leave the vascular space, and that dissolved in the plasma, which is free to diffuse or be

taken up by tissues. As the concentration of free xenobiotic falls in plasma, some molecules will separate from their binding sites until a new equilibrium is reached. Binding therefore acts as both a storage and distribution mechanism, maintaining a more even blood concentration than would otherwise be the case and reducing the peak concentration that would otherwise be presented to tissues.

Bound xenobiotics may be displaced by other xenobiotics. Some drugs, such as barbiturates or sulfonamides, compete with others for binding sites and may increase the concentration of free xenobiotic in the plasma and therefore increase toxicity. As a practical matter, therefore this phenomenon is of greatest significance in drug-related toxicology as a mechanism of drug interaction and overdose and is seldom a consideration in environmental toxicology.

The persistence of free (unbound) xenobiotic in the bloodstream is an important determinant of the duration of its action and the penetration that may occur into tissues less avid in their uptake of the particular agent. However, the most important determinant of uptake by the target organ is the uptake of the xenobiotic from plasma into the tissue.

Uptake of a xenobiotic by an organ from the plasma depends on the blood flow to the organ and the avidity of the tissue for the material. Special transport mechanisms exist at the cellular level for some xenobiotics. As mentioned above in the context of absorption into the body, absorption of a xenobiotic from the bloodstream into the tissue depends importantly on the solubility of the xenobiotic in fat (how "lipophilic" it is); lipophilic agents will be accumulated in adipose tissue or lipid-rich organs such as the nervous system or in liver. Where the physicochemical properties of the organ attract and bind metals, as in bone, a metal xenobiotic will be sequestered and will accumulate over time.

Entry into some tissues is restricted by special barriers to passage, such as the blood–brain barrier and the placenta. In most cases, however, delivery of a xenobiotic depends on the blood supply to a tissue relative to its weight. When the xeniobiotic is neither particularly lipohilic nor sequestered nor preferentially taken up by some organ-specific mechanism, it is largely distributed on the basis of blood flow to the target organ; organs with greater perfusion will tend to accumulate the xenobiotic because of the increased total amount presented to it. The lung, a very lightweight organ, is the only organ of the body to receive 100% of the cardiac output at a tissue level. (The heart, functioning as a pump, moves blood in bulk but is itself nourished by a much smaller coronary artery system.) Not surprisingly, the lung is a principal target organ for blood-borne as well as airborne xenobiotics. The liver and kidneys each receive massive fractions of the cardiac output and are therefore presented with circulating xenobiotics in quantity. The brain also receives a disproportionate fraction of the cardiac output but is partly protected by the blood–brain barrier; this barrier works well for most polar xenobiotics but is permeable to lipophilic compounds.

Between the stomach and the liver there are blood vessels called "portal veins" that bring nutrient-rich blood directly, without passing through the general circulation, to the liver for processing before distribution throughout the body by way of the general circulation. The portal circulation, however, also delivers ingested xenobiotics at high concentrations from the stomach and small intestine directly to liver cells often causing liver toxicity. The liver is the principal metabolic site for xenobiotics, as it is for nutrients. The portal delivery of xenobiotics therefore provides an opportunity for metabolism of some xenobiotics to take place before the xenobiotic enters the general circulation. Xenobiotics metabolized in the liver may even be taken up and reprocessed through biliary excretion and reabsorption through the enterohepatic circulation, such as many pesticides and metals, including mercury.

Some tissues have an affinity for xenobiotics with certain characteristics. Organs with a high adipose or lipid content accumulate much larger concentrations of highly lipophilic xenobiotics, such as organohalide pesticides or the PCBs, than occurs in plasma or in other organs. When an obese individual in whose adipose tissue is stored a high level of a fat-soluble toxic chemical rapidly loses weight as a result of dieting, food deprivation, unaccustomed exercise, or cachexia, the xenobiotic may be mobilized and a rapidly climbing circulating level of the agent may rise to toxic levels. In general, however, the principal significance of adipose and intracellular lipid is as a storage depot, in that the blood concentration comes into an equilibrium with release from the tissue where it is stored, remaining fairly constant for the remaining life of the individual. The xenobiotic can rarely be effectively purged from the body in this situation because of the extent of the storage, although strategies exist to steadily reduce the body burden over time by vigorous removal from plasma to force mobilization. Another important implication of storage in fatty tissues is accumulation in breast tissue and subsequent excretion into breast milk. This is the major route of exposure to a variety of xenobiotics for newborns who breast feed. Metals such as lead are sequestered in bone; mobilization from depots in bone by chelating agents may substantially increase blood levels and create a risk of renal toxicity. (see also Chap. 25, this volume).

26.1.3 Metabolism

Many xenobiotics are substrates for intracellular enzyme systems, most of which appear to have evolved as mechanisms for clearing endogenous mainly steroid, hormones or foreign substances taken in with food. These

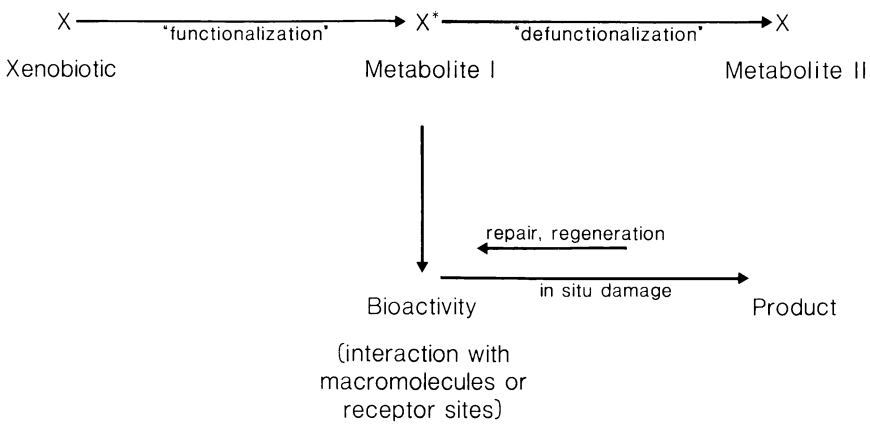


Fig. 26.2 Metabolism of organic xenobiotics often follows a pattern of chemical "functionalization" in which the intermediate product may be more reactive and toxic than the original xenobiotic, followed by de-activation and defunctionalization or conjugation to make the product more water soluble (From Zenz (1994). By permission of Elsevier/Mosby)

enzyme systems transform the xenobiotic in a series of steps from the original compound to a series of stable metabolites, often through intermediate unstable compounds. For many xenobiotics there are many pathways of metabolism, resulting in numerous metabolites. These transformations may have the effect of either "detoxifying", by rendering the agent toxicologically inactive, or of "activation", by converting the native agent into a metabolite that is more active in producing the same or another toxic effect. An active xenobiotic may be transformed into an inactive metabolite, effectively removing the agent from the body in its toxicologically active form. However, an inactive precursor may also be transformed into an active metabolite.

In general, the enzyme systems available for the metabolism of xenobiotics tend to convert nonpolar, lipid-soluble compounds into polar water-soluble products that are more easily excreted in urine or bile. The general pattern consists of two phases. These are illustrated in Fig. 26.2.

Phase I of the metabolic process involves the attachment of functional chemical groups to the original molecule. This usually results in activation, especially in the very important "mixed function oxidase" (MFO) system, and results in a metabolite capable of interacting with macromolecules, such as DNA in the early steps of carcinogenesis. The mixed function oxidase system requires a great deal of metabolic energy and is closely linked with the cytochrome oxidase system, which provides it. Because the particular cytochrome most closely linked with the system has a spectral absorption peak at 450 nm, there is frequent reference in the literature to "P450" as an indicator of MFO activity.

Most important of the metabolizing systems, the mixed function oxidase (MFO) system also is known by other names: aryl hydrocarbon hydroxylase, arene oxidase, epoxide hydroxylase, cytochrome oxidase. It is a complex of membrane-associated enzymes closely linked to the cytochrome P450 system (and other cytochromes) that acts on organic compounds with aromatic or double bonds. The system attacks these bonds, creating first an epoxide and then an alcohol and in the process first activating the compound and then deactivating it and rendering it more easily excreted. The MFO system is virtually ubiquitous in the body but activity is particularly concentrated in liver and lung, and can be found and conveniently studied in circulating lymphocytes. The MFO system has a huge capacity and acts on a wide variety of substrates. It also has the property of being inducible; when presented with suitable substrate, the cell synthesizes more MFO enzymes, increasing the capacity of the system, and preparing itself for a greater load. The degree of inducibility and the level of baseline activity in a given tissue is genetically determined, so that at any one time MFO activity in a particular tissue reflects heredity combined with exposure in the recent past.

Phase II involves the removal or conversion of chemical groups in such a way as to render the molecule more polar and therefore more easily excretable by the kidney (and less easily diffused back across the renal tubular epithelium after filtration). In the process, the activated xenobiotic metabolite from Phase I usually becomes inactivated. This process frequently involves "conjugation", the attachment of a functional group such as sulfonate or glucoronic acid that makes the molecule much more hydrophilic.

The most complicated metabolic pathways are those for organic compounds. Metals may also be metabolized, however. The methylation of mercury and arsenic, especially, plays a major role in their toxicity. The methylation pathway of arsenic is species specific and this is thought to be the reason why arsenic is a carcinogen in humans but not in animals.

26.1.4 Excretion

The xenobiotic or its metabolites would remain and accumulate within the body if there were no mechanisms for excretion. Elimination is the term used for removal of the xenobiotic from the bloodstream, whether by excretion or metabolism or sequestration (storage).

The kidney is the major route of excretion for most xenobiotics. Those that are water-soluble may be filtered or excreted unchanged. The reserve capacity of the kidney is

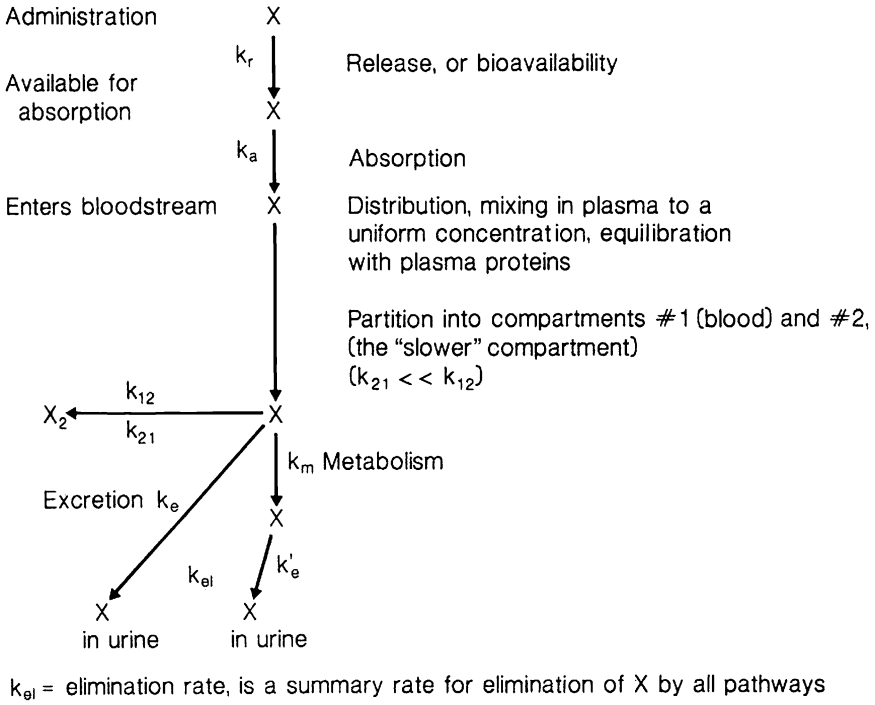


Fig. 26.3 Kinetics of elimination are determined by the behavior of the xenobiotic in different toxicokinetic phases and can be modeled (From Zenz et al. (1994). By permission of Elsevier/Mosby)

very great and this mechanism is rarely saturated in healthy people but individuals with renal insufficiency may show accumulation and persistence of the xenobiotic and, consequently, prolonged and more severe toxicity. Other xenobiotics may be metabolically transformed into more water-soluble metabolites before renal clearance occurs. Xenobiotics that are themselves nephrotoxic may injure the kidney and reduce their own clearance, enhancing their own toxicity by causing further accumulation.

The liver, besides being an important metabolizing organ, secretes some xenobiotics into bile, including heavy metals such as lead and mercury. These may recirculate by the enterohepatic circulation, persisting in the body much longer than otherwise, or they may pass out of the body in feces. Forced biliary excretion is not presently possible but interruption of the enterohepatic circulation by binding agents such as cholestyramine is a practical clinical intervention to hasten excretion and reduce the body burden of xenobiotics excreted in the bile and reabsorbed in the gut; this was first demonstrated for kepone. Although hepatotoxic agents may interfere with their own excretion by the liver, they are more likely to interfere with metabolism and as a practical matter this effect is rarely significant.

Volatile gases are readily excreted by the lung through passive diffusion from the blood, crossing the alveolar-capillary barrier in "reverse" direction. Gases that are poorly soluble in blood, such as ethylene, are rapidly and efficiently eliminated by this route. Those that are readily soluble in blood, such as chloroform, are less efficiently eliminated and may be detectable in expired air for days or even weeks.

Xenobiotics and their metabolites are also eliminated by various minor routes that matter little with respect to reduction of the total body burden but may have toxicological implications. Lipid-soluble agents may be secreted in breast milk; this is a major route of exposure of neonates and young children to substances such as the organohalides, including PCBs. Water-soluble agents are excreted in saliva and tears and are filtered through sweat glands, the latter functioning much like miniature nephrons. Lipid-soluble agents may also be found in cerumen and sebum. These minor elimination pathways permit noninvasive monitoring techniques for the detection of the agent but are rarely reliable enough to quantify exposure.

26.1.5 Kinetics

Metabolism and excretion define the rates of elimination, the change in the concentration of the xenobiotic in the plasma with time. Elimination may occur (1) because the xenobiotic is excreted, (2) because it is converted to something else by metabolism, or (3) because it is stored somewhere inaccessible to the bloodstream. The description of the rates of elimination of the agent is an important tool in understanding its behavior in the body. Each phase of the kinetics of a xenobiotic is governed by rates determined by properties of the agent and characteristics of the biological system, as illustrated in Fig. 26.3. Each rate is described by a rate constant (k) that determines how rapidly the process proceeds.

Rate constants are described by their "order". A zero-order rate constant describes an elimination curve in which the rate is limited intrinsically by the fixed ability of the body to eliminate the agent, regardless of its concentration. In practice, the only important example of this is, ironically, the most common metabolizing system of toxicological concern: alcohol dehydrogenase, which metabolizes ethanol and other alcohols. Regardless of how much alcohol a person ingests, elimination will occur at a fixed rate regardless of the dose or plasma concentration. (The kinetics of the enzyme are not truly zero-order and it behaves as a first-order elimination curve at low levels of alcohol consumption; the capacity of the enzyme is simply saturated quickly as a person consumes more than a minimal amount.)

A first-order rate constant describes a process in which the rate of elimination is independent of the dose and the concentration of the xenobiotic in plasma is proportional to the concentration of the agent in tissue, the most common situation. The concentration of the xenobiotic in plasma decreases over time by a constant proportion per unit time. This is called a "one-compartment" model because the agent behaves as if it is restricted to one compartment of the body, the vascular space. In reality, it may not remain in the vascular space but equilibrates freely with it from its tissue depots. Water-soluble xenobiotics usually show first-order kinetics, except for alcohols.

A multi-compartment, or "multiexponential", function of elimination suggests that the agent equilibrates in more than one compartment and is eliminated at different rates from each. The rate of elimination varies with the concentration in plasma and the initial dose and is biphasic. The elimination will not fit a simple exponential decay (or straight line on a logrithmic scale) but will be described by a more complex equation with two rate constants, a "fast" rate constant and a "slow" one. Organohalides typically show two-compartment kinetics because of their storage and slow release from fatty tissue. (The term "second order" is not used because it would imply that elimination rate is a function of the square of the concentration, which is not the case.) Increasingly, the behavior of such xenobiotics is modeled using "physiologically-based pharmacokinetic" (PBPK) models, so-named because they were first worked out for drugs. Metals often have multiple compartments and complicated toxicokinetics.

First-order kinetics are most common for water-soluble xenobiotics. In such systems, elimination of the agent, being proportional to concentration, results in an exponential decay or reduction in plasma concentration over time. The period required for the plasma concentration to drop by half is called the "half life" ($t_{1/2}$). The $t_{1/2}$ can be calculated easily and accurately and is related to the elimination rate for first order systems by the equation: $t_{1/2} = 0.693/k_{el}$.

26.2 Toxicodynamics

The mechanisms of toxic injury are too numerous to accommodate simple classification or generalizations. There are, however, a few general principles that are useful in understanding toxic effects.

26.2.1 Mechanisms of Toxicity

Xenobiotics exert toxic effects by interfering with the normal functions of the body. These effects occur at the molecular and cellular level. Thus, an understanding of normal function and biochemistry is essential for understanding toxicodynamics. The toxic effect is an interaction between the xenobiotic and the cellular and biochemical mechanism, usually involving the interaction between the xenobiotic and a macromolecule, as illustrated in Fig. 26.2. This has not always been understood: for centuries, poisons were considered to be a special class of chemicals and the toxicity of poisons were understood to be intrinsic properties of the chemical, or magic.

Certain organs of the body are in harm's way. Because they may be the first to encounter a toxic exposure, receive a large blood flow, are highly metabolically active, actively metabolize xenobiotics themselves, concentrate toxic substances or their metabolites or have biochemical characteristics that render them vulnerable, the liver, kidney, lungs, skin and bladder are particularly susceptible to toxic effects.

Although there are as many potential mechanisms of toxic effects as there are reactions in biochemistry and functions in physiology, there are a few processes that play special roles. They include the following:

- Inflammation. The body has natural mechanisms to repair and limit damage. Many xenobiotics are irritating to human tissues and induce local inflammation. For others, inflammatory reactions may contribute to systemic toxicity. A particularly important phenomenon associated with inflammation is the production of reactive oxygen and nitrogen free radicals, which cause intracellular damage as a byproduct of inflammation.
- Immune responses. The body also has natural mechanisms to respond to specific foreign substances ("antigens") or cells, by producing antibodies or by mobilizing special cells that destroy the foreign material, and in the process set off inflammatory responses. These immune responses require that the body has seen the antigen first or that it is persistent, and are triggered by subsequent exposure to low levels of the antigen. When the immune response is dysfunctional, it may cause

allergies, diseases collectively called "immunopathies" and self-directed autoimmune diseases.
- Carcinogenesis. Cancer is the prototype for "stochastic" or probabilistic toxic effects, in which the response depends on the probability of an interaction rather than the magnitude of exposure and degree of response. This is discussed below. There is some evidence that certain other classes of toxic response, such as neurotoxicity, follow similar patterns.
- Endocrine mimics. Many xenobiotics interact with hormonal receptors, sometimes by simulating the effect of hormones and sometimes by inhibiting them.

26.2.2 Exposure-Response Relationships

The exposure-reponse relationship is a concept fundamental to an understanding of toxicology. Paracelsus, the great medieval toxicologist, said "It is the dose that makes the poison" and thereby established that poisons were not a mystically benighted form of matter but that all chemicals have toxic properties that become apparent as increasing quantities are consumed or absorbed. It follows from this simple observations that there may be "safe" levels of exposure to even the most toxic substances, a much more controversial assertion. Obviously, there are several dimensions to this seemingly straightforward concept. There are three distinct varieties of the exposure-response relationship that need to be distinguished conceptually. These are the toxicological dose–response relationship, the clinical dose- or exposure-response relationship, and the epidemiological exposure-response relationship. These are illustrated in Fig. 26.1.

"Dose" is generally understood to mean the total quantity of a toxic substance administered; "exposure" is generally considered to be the level of concentration available for absorption by any or all routes at or over a given period of time. Thus, dose is best understood as total or cumulative exposure over a relevant time period. If the dose is given all at once, the dose–response relationship is most meaningful, as it is when the toxic substance is accumulated in the body. If the exposure takes place over a prolonged period of time, the internal dose at any given time tends to vary and it is more useful to think of an exposure-response relationship. When a xenobiotic accumulates and persists in the body, such as lead over a period of weeks or dioxin and pesticides over a period of months and years, cumulative exposure approximates dose in toxicological terms. When a xenobiotic does not readily accumulate and is quickly eliminated, cumulative exposure over a long time period does not equate to effective dose in toxicological terms, although there may be cumulative effects if each exposure produces permanent injury. For example hydrogen sulfide, a highly toxic gas that is often a problem in medical geology because of its association with volcanoes, vents, and petroleum and gas, is acutely toxic with a toxicity that is a high exponent of exposure (on the order of concentration to the fifth power) and does not accumulate; for hydrogen sulfide, duration of exposure does not matter much but the concentration at first contact is critical.

The most fundamental building block of toxicology is the dose-reponse relationship demonstrable in the laboratory, often called the "toxicological" dose- or exposure-response relationship. The fundamental principle is that the physiological response depends on the amount of the agent presented to the tissue. In a given individual, exposure to an increasing amount of a toxic substance leads to the progressive appearance of new and usually more severe health problems leading to death, a sort of stepladder to lethality.

This gives rise to another type of dose- or exposure-response relationship, which might be termed the "clinical" exposure-response relationship. At a given level of exposure, often referred to clinically (if colloquially) as a "threshold", one can usually expect a given constellation, or "toxidrome", of symptoms and signs. This clinical exposure-response relationship depends importantly on the strength of the host defenses of the individual (which can be very variable) and whether the individual has an acquired condition or genetically-determined phenotype that renders him or her more vulnerable than others. In a given exposure situation, one person may show one symptom and another a different symptom, based on personal susceptibility. At relatively low levels of lead toxicity, some patients show elevated uric acid levels because of reduced renal clearance and develop gout, most do not. As well, the detection of the expected clinical response depends on the sensitivity of clinical examination and laboratory tests. Clinical tests are often inadequate for early detection of equivocal cases because they were designed for making specific diagnoses in people known to be sick in a way that strongly suggested a particular type of disease.

The third type of exposure-response relationship relates exposure levels to the frequency of the response in a population. If one is interested in what personal characteristics of those exposed render them vulnerable to a toxic effect or in how frequently a response is associated with a given level of exposure in a population, one may do a "nose-count" of the observed response among individuals exposed. This is the essential method of epidemiology and yields what is usually called the "epidemiological" exposure-response relationship. To be meaningful, however, the outcome must be experimentally or clinically detectable. This removes from study many types of response that cannot be directly measured and which are usually considered "subclinical"

or "adaptive responses". In this system, a "threshold" means the level of exposure associated with the first appearance of an excess of the health outcome representing the toxic response. It is this threshold for response that generates the most controversy in regulatory policy. However, interpretation of this type of threshold depends on understanding the basis for selecting and detecting the health outcome.

At higher levels of exposure, the exact shapes of these exposure-response relationships are not critical and the general relationship is usually obvious. At lower levels of exposure, however, interpretation of the population response is very dependent on an interpretation of the general mechanism of the toxic effect and extrapolation to low exposures is very sensitive to the biological model applied.

A particularly important, if confusing, term in toxicology is "threshold", which means the level of exposure at which an effect is first observed. The existence of thresholds for certain types of response (particularly carcinogenicity) are controversial and arguments surrounding identification of a threshold for response frequently neglect to specify the type of threshold under consideration.

26.2.3 Interaction

Some xenobiotics interact with others to produce disproportionate effects. For example, exposure to sulfur oxides in the presence of particulate air pollution in combination causes worse lung irritation than would be predicted by the individual effects of each added together. This is because the sulfur oxide adsorbs onto the surface of the particle and is carried deeper into the lung than it would be as a gas.

Interaction may be positive (often called "synergy") or negative (often called "inhibition"). Different models of interaction are applied to the interpretation of data. When the effects of two or more are additive, no interaction occurs. This may suggest that both are acting by the same pathway or mechanism and are simply adding their proportionate share to the total magnitude of the effect. When the effects multiply, this is strong evidence for extensive interaction, and suggests that the xenobiotics are acting by different pathways that potentiate each other's effects. When the effects are less than additive, it is evidence that in some way the effects of one exposure are reducing or blocking the effects of the other or are acting by a similar pathway that has only limited capacity.

Toxicologists are very concerned that exposure to mixtures, such as cigarette smoke or heavily contaminated groundwater, could present the potential for numerous interactions and unpredictable effects. In practice, except for drugs relatively few examples of interaction producing significant health effects have been documented. Some of the most important involve cigarette smoking and persistent carcinogens, including asbestos, silica and radon daughters, which greatly increase the risk of lung cancer compared to the sum of the separate risks of either smoking alone or exposure to the other carcinogen alone.

26.2.4 Carcinogenesis

Much of environmental toxicology is oriented toward the etiology and prevention of cancer. This emphasis is not misplaced, as cancer is a leading cause of death and results in the largest total loss of quality-adjusted years of life, among all causes, in developed countries.

Carcinogenesis is not a straightforward, deterministic process. Rather, at each step in the sequence there is a finite probability of events leading to the next step. Chemical carcinogenesis is thus a stochastic, or probabilistic process, like a roulette wheel or radioactive decay, not a certain prediction based on chemical structure and properties. In any one individual, an exposure may increase the odds of getting cancer but do not make it certain in absolute terms that this will happen.

Chemical carcinogens are demonstrable by their effect in increasing the frequency of cancers observed in exposed subjects as compared to unexposed. They may produce malignant tumors that are often different in tissue type and wider in diversity than those usually observed among unexposed subjects, produce malignant tumors at characteristic or unusual sites in the body, and produce these malignant tumors earlier in the life span of subjects than they would otherwise be seen. Often, however, chemical carcinogens produce malignancies identical in tissue type, location, and onset to those seen in unexposed populations and the only clue is an increased frequency of cancers in exposed groups.

Recent advances in research on carcinogenesis, especially the identification of the oncogene, may have identified new and rather complicated mechanisms but the effect has been to simplify our understanding by providing common pathways and unitary, comprehensible mechanisms by which many causes may act. The principal categories of causes of cancer are fairly conclusively identified as heredity, chemical exposure, viral infection, and radiation exposure. Other categories of causes may be identified, but these appear to be primary.

The contemporary model for induction of cancer by chemicals that is most consistent with available evidence for most chemicals and for radiation is that called the "two-stage model of carcinogenesis". (The model is insufficient to account for some other types of cancer induction - these are discussed below.) The two-stage model assumes the introduction of a carcinogen into the body (or the metabolic activation of a procarcinogen) and its distribution in the body in such a way as to be presented to a tissue at levels

in which it is likely to be taken up intracellularly and to react with cellular constituents, most specifically nuclear DNA. Transformation of the cell does not occur with every interaction between a carcinogen and DNA, however. Only in a relatively small fraction of such interactions will the critical sites on DNA be affected, resulting in a probabilistic phenomenon. When it occurs, the process is called "initiation" because those cancers that may ultimately result are initiated at this step. In many cases, things presumably go no further than initiation. The mechanism for much, if not most, initiation activity is oncogene activation.

Among these relatively few but numerically many interactions with DNA are a handful that may cause the cell to behave in a manner more appropriate to a primitive, embryo-like state and these are thought to be the mechanism for transforming normal cells into neoplastic cells. Oncogenes are capable of being activated by chemical exposure. They are latent within the genetic structure of all humans (and probably all advanced life) and at least some probably play a physiological role in normal embryonic and fetal growth and development. Activated in the absence of regulation, however, the oncogenes trigger malignant transformation of the cell, causing a previously differentiated cell to regress to a more primitive state abnormal for that stage of the life of the organism.

Next in sequence is the growth of a clone of transformed cells from a single cell altered in its growth characteristics to a small focus in situ. The transformed cell, having been altered in its DNA blueprint, does not necessarily begin to multiply at once. Rather, it may be held in check by host factors or cell-specific factors, such as the need for further DNA reorganization or oncogene activation to take place. The abnormal cell may rest for a very long time, contributing the greater part of the latency period before appearance of the clinically evident tumor. Additional exposures may trigger the conversion of the initially abnormal cell into a transformed or preneoplastic cell capable of giving rise to a tumor. This process may be facilitated by exposure to chemicals that also have genotoxic potential, either simultaneously or after the action of the primary carcinogen. This ancillary process is called "cocarcinogenesis", implying that the second or combination exposure may not be the initiator but participates in the genotoxic cell events and either leads to expression of the critical event, resulting in oncogene activation, or overrides mechanisms that would otherwise inhibit oncogene activation and cell transformation. In general, the same chemicals that are primary carcinogens are likely also to be cocarcinogens. The distinguishing feature is not which chemical reaches the DNA first or which exposure preceded which, but which chemical actually participated in the critical event that specifically altered the DNA in such a way as to activate the oncogene.

Exposure may occur at this stage to chemicals that are capable of triggering proliferation by removing the inhibitory factors that are suppressing the transformed cell. This is called "promotion" and it is the second stage in the two-stage model of carcinogenesis. Promoters are sometimes primary carcinogens themselves and probably act through genetic mechanisms, such as the polycyclic aromatic hydrocarbons, but others are weakly or not carcinogenic and presumably act by nongenetic mechanisms. The most well known are the plant-derived phorbol esters (specifically tetradecanoyl phorbol acetate, TPA), constituents of croton oil that are chemically extremely complex and seem to act at least in part pharmacologically by activating certain specific receptors on the cell surface. Chlorinated hydrocarbon species are often potent promoters, including the PCBs, DDT, PBBs, and certain dioxins. They seem to act by nongenetic means and have variable primary carcinogenic activity, depending on the species.

By whatever mechanism, promotion results in deregulation and progression of the neoplasm by proliferation into a clone of cells. The transformed cell has now become a cancer cell with the essential features of a malignancy: unresponsiveness to regulation, loss of contact inhibition, potential for sloughing and migration of cells, and the potential for inducing growth of new nutrient blood vessels.

Because this all takes time at each step, there is a delay between the initiation (commonly assumed to be at first exposure) and earliest clinical presentation of a tumour. This is called the "latency period". For most chemically-induced cancers it is on the order of 20–30 years but may be as long as 50 or more (in the case of mesothelioma and asbestos) or as short as 5 years (for radiation or radiomimetic exposures and some bladder carcinogens).

Not all chemically-induced cancers act by this genetic mechanism. "Epigenetic" refers to the actions of cancer-inducing agents and exposures that do not directly interact with DNA. At least some probably act by inducing intracellular free radicals that damage DNA in a nonspecific manner, however. Asbestos is an example. Others are more obscure in their mechanisms. None are adequately explained by the conventional two-stage model of carcinogenesis but subsequent refinements in theory will almost certainly result in a unitary model demonstrating a final common mechanism for most cancers.

Metal-induced carcinogenesis occurs by a variety of mechanisms and often strongly depends on the chemical composition, redox state and solubility: arsenic (lung, bladder and skin), beryllium (lung), cadmium (lung), chromium (hexavalent ion: lung), nickel (subsulfide: lung).

26.3 Interpretation of Trace Element Levels

By definition, trace elements are normally present in very low concentrations in the body. A good working definition of trace elements in medicine would include an element that is present in minute amounts and that is not under tight physiological regulation. Iron and magnesium, for example, are not true trace elements in the body because substantial quantities are normally present and concentrations are kept within a normal range by mechanisms of biological control, reflecting the importance of these elements in metabolic and physiological processes. Other trace elements are not kept under control by the body and therefore blood and tissue concentrations vary widely. Even trace elements that are essential elements for various metabolic functions (e.g. cobalt, manganese, selenium) may not be physiologically regulated because they are present only in minute quantities. Most trace elements are metals, although two are metalloids (arsenic and selenium).

Three biological fluids are assayed on a routine basis: serum, whole blood, and urine. These analyses typically report on a profile of several metals. The state of the art for trace element analysis in body fluids is a highly accurate and sophisticated technology called inductively-coupled plasma mass spectroscopy (ICP-MS) to analyze the concentration of minute amounts of metals in a sample. It is an exceptionally accurate method for low concentrations of trace elements, in the nmol/l or μmol/l range. Metals that are present at higher concentrations, such as iron, overwhelm the system and are better determined using traditional methods. The technology works equally well for biological fluids and for water samples and is used routinely for both.

The clinical interpretation of trace element analysis has lagged behind the technology available to measure the elements in body fluids. It is easy to automate the determination of a dozen or so trace elements (metals or semimetals such as arsenic and selenium) in blood or urine. This is commonly done commercially by laboratories that offer their services in screening for purported disease through "alternative" health practitioners that do not necessarily practice in the mainstream of modern medicine. However, if health practitioners do not know what they are looking for or do not understand the principles of toxicology, they can be seriously misled. The use of trace element analysis by alternative medical practitioners and its conduct by questionable laboratories have somewhat discredited the method and is a frequent source of confusion in medical diagnosis.

Interpretation can be difficult to interpret and requires knowledge of toxicokinetics. Trace element analysis is best used for specific applications, such as establishing levels of exposure, epidemiological studies, occupational and environmental health surveillance, biomonitoring of populations, and to confirm an association following a compatible diagnosis. Except for lead, trace element analysis is not well suited for screening individual subjects because the results are not easy to interpret, there is wide variation and a skewed distribution in the population, and even elevated levels carry limited prognostic value. Just because a level of, say, selenium is higher in one person than is usually seen in the general population does not mean that that person is suffering from selenium toxicity or needs to do anything about it. Some metals, such as manganese, do not lend themselves at all to analysis and results correlate very poorly with evidence of toxicity.

Trace element analysis reflects absorption from all sources, including diet (the major source for most of them), occupational exposure (a major source for elevated levels), smoking (a particularly important source for cadmium), hobbies, housing quality (the major source for exposure to lead in children being peeling paint chips), medication, and local soil-containing dust (another important source of lead). It may also reflect characteristics of the host in retaining or accumulating the trace element, as in the case of patients on dialysis (aluminium), receiving parenteral nutrition (in which nutrition is provided directly into the bloodstream through a catheter) with inborn errors of metabolism (such as Wilson's disease, which involves copper but is not caused by excessive intake), or with impaired excretion (lead).

The essential "trace elements" are copper, cobalt, manganese, molybdenum, selenium and zinc. A balanced and regular diet generally keeps these trace elements roughly constant and consistent in well-nourished populations, but there is no homeostatic mechanism that does so in the individual. Deficiencies in the uptake or metabolism could result to sustained imbalance of trace elements. In addition, an excess intake could result in disease. Other trace elements, such as arsenic cadmium, lead and mercury, have no known human body function and high exposure to these elements could result in both immediate and delayed adverse health effects.

Testing is performed on serum (the liquid fraction of blood) for metals that are carried in the blood in dissolved for or that are bound onto proteins that circulated in the blood. These include aluminium, antimony, barium, beryllium, copper, manganese, nickel, selenium, vanadium, and zinc. Serum samples cannot be used for lead or arsenic and are inaccurate for some trace elements such as bismuth.

Testing is performed on whole blood for metals that are present in serum and also for those that concentrate in the red cell fraction of blood. Concentrations in serum and whole blood for those metals present in serum are usually similar but in some cases, as for copper, the serum assay result may be somewhat higher because there is less in the red cells that make up approximately 40% of the volume of whole blood. Those metals that accumulate preferentially in red cells are

not accurately reflected in serum concentrations and only whole blood concentrations are valid for these metals: cadmium, cobalt, molybdenum, lead, and thallium.

Urine is tested to determine the excretion of metals over a 24-h period for aluminium, antimony, arsenic, barium, beryllium, bismuth, cadmium, manganese, selenium, lead, thallium, vanadium, and zinc. This is usually the most accurate reflection of the total body burden of the metal. Many factors affect excretions of metals over short periods: state of hydration, renal function, intake with foods, short-term exposures from other sources, renal blood flow. Over a longer period of time, however, these variations even out and excretion is then generally directly related to the average serum concentration during this period (in equilibrium with red cell concentration, in the case of metals that accumulate in red cells). For this reason, spot urine samples are not very useful in determining whether excessive exposure has occurred, although their toxicological value can be improved somewhat if the results are normalized to the urinary creatinine level (a marker of kidney function). Compliance is much more difficult to achieve in collecting a 24-h urine specimen but it is preferable for an accurate and valid result.

Hair analysis has been very popular among some practitioners because of its convenience in sampling. However, hair is a difficult matrix for trace element determinations because it is solid and so difficult to process, and is contaminated by dust, hair products, and other environmental exposures. Concentrations may also vary along the length of the hair. As a consequence, hair analysis is not recommended for trace element analysis except in highly controlled settings, when collecting body fluids is not possible or practical, and when forensic studies are conducted (for example, assessing changes over time along the length of the hair).

The reported value is compared against a reference range which is usually derived from the authoritative literature, matched against the experience of the local laboratory. In recent years, data have become available from large cross-sectional surveys, such as the National Health and Nutrition Examination Survey in the United States. Reference ranges for occupational exposure are often benchmarked against standard reference works. Trace element studies on populations must be interpreted with knowledge of the log-normal distribution pattern and suitable transformation for statistical analysis. Always, the toxicokinetics of the trace element defines the meaning of the results obtained.

Chelation therapy is a potentially dangerous form of treatment in which trace elements (usually lead, mercury or arsenic) are eluted from tissue by chemical agents that coordinate with and trap the metal ion, solubilizing it and making it easier to excrete. The abuse of chelation treatment is intimately linked with misunderstanding of trace elements and the significance of laboratory levels. Greater care and understanding in interpretation of trace element analysis will lead to safer interventions and less potentially dangerous over-utilization of chelation. The danger, even in treatment for overt toxicity from metals, is that mobilization of the metals also enhances their toxicity, particularly to the nervous system and kidney.

See Also the Following Chapters. Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 9 (Biological Responses of Elements) • Chapter 22 (Environmental Epidemiology) • Chapter 21 (Environmental Medicine) • Chapter 24 (Environmental Pathology)

Further Reading

Guidotti TL (1994) Principles of occupational toxicology (Chapter 7). In: Zenz C, Nickerson B, Horvath EP (eds) Occupational medicine: principles and practical applications, 3rd edn. Mosby-Year Book, Chicago, pp 70–84

Guidotti TL (2010) The Praeger handbook of occupational and environmental medicine. Praeger ABC-Clio, Santa Barbara

Guidotti TL, McNamara J, Moses MS (2008) The interpretation of trace element analysis in body fluids. Indian J Med Res 128:524–532

Hoffman H, Phillips S (2010) Clinical practice of biological monitoring. OEM Press, Beverley Farms

Klaasen C (ed) (2007) Casarett & Doull's toxicology: the basic science of poisons. McGraw Hill, New York (A simplified and condensed version of this standard textbook is available from the same publisher under the title Casarett & Doull's Essentials of Toxicology)

US Agency for Toxic Substances and Disease Registry. Toxicological Profiles (series). Atlanta GA, U.S. Public Health Service, date of publication varies with profile. Web access: http://www.atsdr.cdc.gov/toxpro2.html

Zenz C, Nickerson B, Horvath EP (eds) (1994) Occupational medicine: principles and practical applications, 3rd edn. Mosby-Year Book, Chicago, pp 70–84

Speciation of Trace Elements

27

Bernhard Michalke and Sergio Caroli

Contents

27.1 Introduction

The determination of trace elements has assumed a place of prominence in the life sciences. Elements present at even minimal concentrations in biological and environmental matrices can have a significant influence on vital functions, depending on the amount present. The study of, for example, pathophysiological processes in the human body requires the determination of elements at concentrations measured in $\mu g L^{-1}$, $n g g^{-1}$, and even $p g g^{-1}$. The higher concomitant amounts of organic and inorganic components make it difficult to determine the presence of trace elements. Moreover, it is a complex process that progresses from an initial trace element analysis to the final statement of biological implications, one that requires close collaboration between the analytical chemist and life scientist. Furthermore, it should be kept in mind that the concept of *zero tolerance* for potentially toxic elements has been replaced by the more scientific notions of *safe ranges of exposure* and *range of safe intake*.

The research field of element speciation analysis has its roots and basics in (trace) element analysis and was rapidly developed within the recent decades: Toxicological investigations clearly indicated that different compounds (species) of the same element can exhibit largely different adverse or essential effects (e.g. Cr(III) is considered essential while Cr(VI) is classified as carcinogenic). Analytical chemists have come to realize that, in general, the total concentrations of chemical elements cannot provide information about their mobility, bioavailability, and eventual impact on ecological systems and biological organisms. Only knowledge of the chemical species of an element can provide information regarding possible chemical and biochemical reactions and thus lead to a greater understanding of toxicity or essentiality. Therefore the determination of the element content itself is not sufficient for adequate risk assessment. On the other hand the significant progress and improvement in the commercially available analytical

B. Michalke (✉)
Helmholtz Center Munich, German Research Center for Environmental Health, Munich, Germany
e-mail: bernhard.michalke@helmholtz-muenchen.de

S. Caroli
Community Reference Laboratory, Istituto Superiore di Sanità, Rome, Italy

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_27,

instrumentation, in particular mass spectrometric detectors, has opened many new applications for elemental speciation analysis with lower limits of detection. Initially, the main focus of speciation analysis was on organometallic compounds, like alkyl tin species, which were recognised as health risks due to their high toxicity and which could easily be determined with existing GC-MS instrumentation after derivatisation.

But now, the scope and focus changed significantly. Nowadays metallo-protein complexes (e.g. Cd-metallothionein, Mn-transferrin), metal containing carbohydrate metabolites (e.g. Arsenosugars, Selenosugars), metal complexes with low molecular mass organic ligands (e.g. Mn-citrate, Pt-nucleotide adducts) and different oxidation levels of the same element (e.g. Cr(III) and Cr(VI)) are investigated. The majority of these species is polar and thermo-labile and therefore liquid chromatographic separation has gained high relevance in speciation analysis. The increasing use of metallo-drugs in clinical therapy (e.g. Pt compounds for cancer treatment), the large variety of mineral supplements available without prescription (e.g. Se and Cr enriched yeast) and advances in legislation (e.g. limits for alkyl tin species) led to a high demand for reliable methods for the determination of element species in a broad range of matrices.

Until now, already established separation and detection methods had to be combined in novel ways and modified according to particular speciation problems (For specific details on various analytical techniques, see Chap. 31, this volume). Combination and hyphenation of separation techniques and element- or molecule-selective detection systems are generally the approaches of choice for speciation analysis; however, further methodological developments are still necessary, primarily for hyphenation and quality-control strategies: In spite of significant advances in instrumentation, research, method development and validation there are still a lot of challenges in speciation analysis. This is mainly due to the broad variety in the chemical properties of elemental species. Organometallic compounds with low polarity of the metal—carbon bond (semi-metals As, Se, Sn) exhibit fairly high stability during sample preparation and analysis and thus allow mostly species identification and quantification. More care is required in the case of transition metal complexes with high(er) polarity of the organo-metal coordination bond. In case the native conditions of the sample are preserved (in particular the pH) several complexes like Cd-metallothionein and Fe(III)-transferrin have high stability constants (up to 10^{20} mol^{-1}). However, many low molecular mass complexes like Mn-citrate and V-glutathione have fairly low stability constants (<10,000). As a consequence, there is a high risk of species transformation in case of sample dilution, sample interaction with the mobile or stationary phase and separation of the metal-complex from excess apo-ligand during chromatography.

In the case of element species with low stability the use of direct methods, that means species selective detection techniques without the need for species separation, would be of great advantage. However, the available direct methods nuclear magnetic resonance (NMR) and X-ray absorption near edge spectroscopy (XANES) are typically far less sensitive than the hyphenated methods based on HPLC coupled to elemental mass spectrometry and therefore often not applicable to trace element speciation in real sample matrices. Consequently, hyphenated techniques are predominantly applied due to their high sensitivity and high flexibility. However, apart from the instrumental speciation analysis itself investigations on quality control have shown that changes in the original species information can easily occur already during sampling, sample preparation and storage.

27.1.1 Definitions of Terms Related to Speciation

The use of concepts and terms related to chemical speciation in recent years still reflects a certain degree of inconsistency within the scientific community. In recognition of the importance of standard terminology from the viewpoint of both interdisciplinary communication and constructive interaction with decision makers, the International Union of Pure and Applied Chemistry (IUPAC) has undertaken a collaborative effort of three of its divisions and reached a consensus on some basic definitions that can be used by specialists in the discipline of speciation (Templeton et al. 2000). One of the major conclusions of this working group was that the term "speciation" should be restricted to the distribution of an element among well-defined chemical forms. A clear distinction was also made between *speciation* and *fractionation*. As it is quite crucial for those working in this field to fully abide by such definitions, they are reported below verbatim:

- *Chemical species* (of a chemical element)—Specific form of an element defined as to isotopic composition, electronic or oxidation state, and/or complex or molecular structure.
- *Speciation analysis* (in analytical chemistry)—Analytical activities of identifying and/or measuring the quantities of one or more individual chemical species in a sample.
- *Speciation of an element*—Distribution of an element among defined chemical species in a system.
- *Fractionation*—Process of classification of an analyte or a group of analytes from a certain sample according to physical (*e.g.*, size, solubility) or chemical (*e.g.*, bonding, reactivity) properties.

The analytical activity of identifying and measuring species includes a clear identification of the species (elements and possibly the binding partners) and exact quantification in representative samples, as well as quality control (Caroli 1996; Michalke 1999a). If the identification and quantification of a chemical species cannot be performed, the analytical procedure can only lead to operationally defined species characterization. Operational or functional chracterizations are not considered to be real chemical speciation analysis, as the true identification of the single species is missing.

27.1.2 Operationally and Functionally Defined Characterization of Species (Groups)

Operationally and functionally defined characterization of species has to be distinguished from chemical speciation analysis (Ure et al. 1993). The former gives a characterization of molecule groups (not single species) that show a similar behavior during an analytical procedure (operation) such as extraction. Characterization of the molecule groups is strongly dependent on the selected analytical procedure, and usually the original species information (identity) is lost. In this sense, fractionation is regarded as an operationally defined characterization. The functionally defined characterization of species (groups) provides information about the function of species groups (not singly identified species) in organisms and their impact on living systems (Caroli 1996; Mota and Simaes Gonçalves 1996). Neither is considered to be a real chemical speciation analysis, as identification of a single species does not actually take place. However, often such analytical attempts are the sole possibility to obtain element speciation information of complex samples because individual species are not accessible to the available analytical instrumentation (e.g. in tissue samples), extraction is necessary with operationally defined leaching of species (groups) or species are instable even when gentle analytical methods are applied.

27.1.3 Need for Element Speciation

The quality and quantity of the relevant element species in a matrix, rather than the total element concentration, are greatly responsible for the mobility, bioavailability, and ecotoxicological or toxicological impact of an element (Florence 1989; Templeton et al. 2000). Elements usually interact as parts of macromolecules (*e.g.*, proteins, enzymes, hormones) or according to their oxidation state; therefore, only current knowledge about a species provides a reason for further assessing whether or not it is toxic, without (known) impact at a specific concentration, or essential. Problem-related speciation analysis appears to play a key role in effectively assessing the risk posed by elements in the environment. Also, from a health viewpoint, the consequences of a trace element's essentiality, depletion, or toxicity can be better determined, and the development of diagnoses and potential therapies is improved. Some examples are provided in the following text to illustrate how different species of a given element can have different impacts (see Table 27.1). It must be realized that, in speciation analysis, the analytical procedure generally interacts with the separation and detection of a species. These interactions usually shift equilibria among the species and possibly change some of the species themselves. The nature and extent of these alterations, as well as a critical discussion of the results achieved, should therefore be an essential part of well-conducted speciation analysis.

27.1.4 Useful Fields for Speciation

Speciation is particularly relevant to the environmental field, as well as to biology and medicine. Food chemistry and nutrition, in turn, can also greatly benefit from the speciation approach, which can act as an interface between these two fields. The nature and amount of manmade species are altering natural species formation and equilibria; consequently, trace element mobility and bioavailability may be influenced and modified. Bioavailability is directly linked to biochemical mechanisms within the organism. Thus, the fate of a species—such as adsorption to membranes, transport and incorporation into larger molecules (*e.g.*, enzymes), and enrichment or excretion—may be modified and result in an unbroken path from environmentally changed species to toxicity, deficiency, or growth in biological systems.

27.1.5 Species Impact and Mechanism in Biological Systems

A necessary prerequisite for an elemental species to interact with an organism is that the species must be able to cross the cell membrane and participate in biochemical paths and reactions (Morrison 1989). Several intake mechanisms are known, as detailed in Table 27.2. These uptake mechanisms result in an enrichment of element species in the organism by a factor of 10^2–10^5. In some cases, toxic concentrations are reached even when the original species concentration in the environment is low (Morrison 1989). The uptake and subsequent metabolization of element species is obviously dependent on the nature of the species itself, as are the consequences of that uptake and metabolization:

Table 27.1 Impact of various species of elements

Element	Impact of species
Arsenic	As(III) and As(V) are toxic; arsenobetaine is nontoxic
Chromium	Cr(III) is essential; Cr(VI) is highly toxic and cancer-promoting
Copper	Ionic Cu(II) is toxic in aquatic systems; humic complexes of Cu are almost nontoxic
Iodine	Thyroid hormones influence an extended range of biochemical reactions in organisms and play an active role in immune defense. Triiodothyronine (T_3) shows about a fivefold effectivity compared to thyroxin (T_4), but it comprises only about 20% of the total iodine hormones
Iron	Absorption capacity for Fe(II) is lower compared to Fe(III), but only Fe(II) is effective against Fe deficiency. This is important for supplementation; however, Fe(III) is utilized efficiently following reduction by ceruloplasmin
Mercury	Inorganic Hg salts are less dangerous than methylated forms; these are more toxic and can be enriched (*e.g.*, up to 10,000-fold in fish)
Platinum	Pt(0) is nearly completely insoluble in water. After emission from car catalysts as Pt(0), Pt species transformation occurs, and solubility in water as well as its availability are significantly increased

Source: Data from Florence (1989) and Lustig et al. (1998)

Table 27.2 Cell-entering mechanisms for elemental species

Diffusion	Diffusion is dependent on the size and lipophilic nature of the element species. It is fast and efficient for lipophilic molecules and is associated with high toxicity. Ionophores, which have such an increased lipophilic nature, may form complexes with (lipophobic) metals. These "excluded metals" are then transported across the cell membrane by the ionophores. In the cell, the metals are set free again and recomplexed by proteins or other ligands
Active transport by ATPase-driven ion pumps	The ATPase-driven uptake mechanism has been proven for some essential elements such as Cu^{2+}, Zn^{2+}, and Ni^{2+}
Carrier/shuttle transport	This transport mechanism is typically shown by proteins and hormones
Uptake via ion-selective channels/active transport by electrical potential	Ion-selective channels have been tested for cations such as Ca^{2+} and K^+. Transportation across the membrane is dependent on the D_{μ}^{-H+}-membrane potential. A potential higher than −70 mV opens voltage-gated channels

Source: Morrison (1989), with permission

- *Immediate excretion without any interaction*—This action is considered beneficial for species having toxic potential and adverse for essential element species.
- *Interaction with the organism and participation in metabolic paths*—This result is considered to be beneficial for essential element species and adverse when toxification and a reduction in enzymatic selectivity and turnover rate occur. The replacement of an essential element by another one in the reactive center of the enzyme can sometimes cause enzyme damage.
- *Intracellular toxicity*—Such toxicity often appears when intracellular species transformations occur. The displacement of essential elements in the reactive enzyme centers results in inactive enzyme–metal complexes (*i.e.*, new species). Conversely, metal exchange at a protein can also be a detoxification reaction (*e.g.*, via metallothioneins [MTs]).
- *Metallothionein transcription*—Metallothioneins, 7- to 10-kDa proteins with about 30% S amino acids (~30% cystein) have a high affinity to metals such as cadmium, copper, mercury, and zinc. In organisms experiencing such a metal load, a genetic MT transcription is induced that increases the concentration of the "offered" ligand to the toxic metals. The generated MT-metal complex is excreted via kidneys, thus protecting the reactive centers of enzymes.
- *Availability of many elements*—Elements must be available for the organism as well-defined species and at suitable quantities to guarantee a normal health status. A good example of this is chromium, which is essential in the trivalent oxidation state and highly toxic and carcinogenic in the hexavalent oxidation state.

27.1.6 Reference Values and Ranges

The achievement of reference concentration values and intervals for elements in biological and environmental matrices is of paramount importance in the detection of imbalances that can adversely affect human health and ecosystems. An exhaustive overview of problems and applications related to this issue was published several years ago (Caroli et al. 1994). In this context, "normal values" are provided as tolerance limits for those elements that may be undetectable in human organs. Elements essential for life, on the other hand, are homeostatically regulated, and their concentrations are expected to fluctuate within narrow limits for each species under normal conditions.

Although doing so is still far from feasible, there is no doubt that for chemical speciation determining reference figures will become even more crucial than obtaining knowledge regarding the total amount of a particular element.

27.2 Speciation Analysis: Sampling, Storage, Preparation, Separation, Detection

The direct determination of trace elements in samples is an important problem in technology, industry, and research, because decomposition and preconcentration procedures, as well as the storage of trace analytes in solutions, are often sources of concern. The accuracy of analytical results can, in fact, be threatened by these pretreatments. Only a few methods exist for direct analytical determinations in solids. In many cases, the detection power and reproducibility of spectral analyses are inadequate to meet the needs of analysts. This less-than-ideal performance is of particular concern in speciation analysis, which requires a series of carefully planned steps among which chemical and/or physical pretreatments of the material being tested are almost always mandatory. In this context, sampling and sample preparation are of prime importance. Without proper sampling and sample treatment procedures, there is little chance that any speciation analysis will be able to provide reliable data upon which human health or environmental decisions can be safely based. Thiers and Vallee (1957) stated that, unless the complete history of samples is known with certainty, the analyst is well advised not to spend time analyzing them. The container in which the sample is stored is a potential source of contamination, as is the sample pretreatment procedure or manipulation or the analyst. Volatilization is another source of error. If unexpected changes to the form of the element occur, such as oxidation state, extent of chelation, or organometallic state, then clearly the species has been changed, and the original species identity and amount cannot be ascertained, thus defeating the purpose of the experiment. On the other hand, highly reliable data can be obtained with careful evaluation of the potential chemical changes in the sampling and sample preparation process.

27.2.1 Sampling

The sampling step is all the more critical in speciation analysis and usually shows uncontrolled and irreversible interferences in the species equilibria (Dunemann and Begerow 1995). Sampling should be designed to preserve the original information about native species; however, existing techniques are often inadequate for the problem at hand. They must be adapted to the actual situation with regard to the element species of interest in the given matrix (Kersten and Förstner 1989). Several problems have been identified; for example, wall adsorption effects have been described, as well as contamination from the sampler and alteration of biological or chemical equilibria or oxidation by atmospheric oxygen (Caroli 1996). When sampling natural water, unintended contamination of the probe frequently occurs in the surface microlayer, which is usually enriched by trace metal species. This contamination can be avoided by taking water samples 0.5–1.0 m below the surface. Also, materials such as dust and airborne particulate matter may be available in only very limited amounts (0.02–0.05 g); therefore, their analysis requires the use of methods with high detection power.

In biological samples, contamination from syringes, metal scalpels, and other metal tools can alter the species pattern of other elements. Species alterations due to bacterial activity have also been observed (Dunemann and Begerow 1995). The extent to which the sample is representative of a chemical species is often not confirmed. Each of the thousands of possible major biological and environmental materials suffers from different matrix effects; even the urine of a given individual can differ substantially in its concentration of salts from 1 day to another. An overview of the sampling problems encountered under critical conditions in an extreme environment has recently been published by Caroli et al. (2000), or, more specifically for cerebrospinal fluid—here standing generally for biological fluids—by Michalke and Nischwitz (2010).

27.2.2 Sample Storage and Processing

In most cases the required analytical instrumentation for trace element speciation analysis is not available at the sampling site and/or at the sampling time. Therefore, storage becomes necessary which should be as short as possible to avoid transformation of element species. For replicate measurements on a given sample and for many applications, longer storage times are necessary. Proven storage techniques typically used for trace element analysis may be inappropriate for element speciation studies: For example, acidification cannot be used for element speciation studies because pH changes affect species composition and thus alter native speciation information.

Unreliable data can be found in the analytical literature due to the ubiquitous nature of certain elements and unaware analysts. Such elements are present in gloves, rubber stoppers, and anticoagulants. Among the many potential sources of contamination are dust, dirt, cosmetics, disinfectants, talc and dust on gloves, and metallic corrosion products. Laboratory dust can contain up to 0.3% aluminum, 0.3% calcium, 0.3% iron, 0.8% potassium, 0.2%

magnesium, 0.3% sodium, 0.2% lead, 2% sulfur, and 0.2% zinc. Preservatives are also rich in elements; heparin contains calcium and zinc, and formalin contains iron, manganese, and zinc. These contaminants are likely to disturb not only trace element analysis but also element speciation. Such contaminations shift species pattern and may rearrange complexes in a sample. Even when the element species of interest is not the contaminating element, a shifting of species equilibria or changes in complexes are possible. Drying of samples can result in the loss of element species (predominantly volatile species). Problems of subsampling must also be considered. In most regulations and directives, analytical methods and sampling are only briefly addressed. Sampling and analysis should be described as accurately as possible to profit from the excellent spectrometric methods available.

Careful storage and processing of samples is required to avoid contamination or destabilization. Clean-room conditions and precleaned vials, among other precautions, must be used throughout. Sample preparation should be as simple as possible to reduce possible alterations. For this reason, storage time should be kept to a minimum, preferably at 4 °C. For long-term storage, freeze-drying or shock freezing at −80 °C is recommended; however, when the freeze-drying procedure is used, the process should allow for control of the sample temperature. Even then, volatile species may be lost. No acid additions or other pH-changing agents are allowed, nor are repetitive, slow thaw/freeze cycles. Glassware can have ion-exchange properties; thus it is less suitable than polymer materials (Urasa 1996). Before storing water samples, particulate matter must be removed to avoid species condensation and a shift in elemental speciation. On the other hand, colloids must not be removed, as they are considered to be a species group or fraction themselves. Mercury is a well-known example of the problems that can be encountered. The possibility of random errors in the determination of mercury due to migration phenomena is primarily associated with the sampling, pretreatment, and storage of samples. Mercury concentrations in water samples stored for a long time are strongly affected by several physical phenomena (*e.g.*, sorption and desorption), dissolvation, and passage through the walls of the container. The errors due to these processes are, however, in most cases independent of the analytical method used.

For biological samples, storage and sample preparation should be as short as possible, preferably at 4 °C. Sample preparation for liquid samples can be kept to a minimum prior the final speciation analysis. However, problems exist for all solid samples, either biological or soil/sediments, as species have to be extracted before separation. Extraction protocols must be chosen carefully to avoid species transformation and a shift of the analysis to an "operationally defined" fractionation approach. Species extractions typically face the contradiction of maintained species stability versus extraction efficiency: The stronger the attack the higher is the extraction efficiency but also the higher is the probability of species transformations (and vice versa) (Quevauviller et al. 1996). This can obviously cause species alteration and bring subsequent analysis into the field of "operationally defined" approaches. As Pickering (1981) wrote: "The gained results often are bound to errors and limitations, finally leading to wrong or misunderstanding data."

With respect to biological tissue a series of extraction methods was compared, like enzymatic extraction or aqueous extraction, HCl leaching and combinations of above approaches, showing different efficiency for different purposes (Potin-Gautier et al. 1997; Crews et al. 1996; Michalke et al. 2002). Wolf et al. (2002) found that low temperature and inert gas atmosphere are inevitable. This was confirmed by Nischwitz et al. (2003), who successfully elaborated a species preserving protocol under Ar-atmosphere, which recently was improved with respect to extraction efficiency while maintaining species stability (Diederich and Michalke 2010).

For soil and sediments the usual extractions are "single extractions" when HPLC-ICP-MS analysis is performed subsequently. Single extractions use leaching agents like H_2O, $NaNO_3$, NH_4NO_3, KNO_3, $CaCl_2$, CH_3COONH_4, EDTA, CH_3COOH. Pellegrino et al. (2000) compared 12 different extraction methods for the example of organotin species in mussel samples concerning the influence of organic solvents, pH or of complexing agents. They found that acidic conditions of 1 mol L^{-1} acid are necessary for extracting all of the investigated tin compounds. Tropolone improved extraction of mono- and di-substituted Sn-species. A promising single extraction method was described for Sn-species using low-power-short-time microwave assisted leaching. Acetic acid or diluted nitric acid was used as extractant, applying microwave power at 30 W, 60 °C for few minutes. The stability of organotin species was preserved and extraction efficiency reached 100% (Rodriguez-Pereiro et al. 1997).

Sequential extraction procedures attack the sample by utilizing consecutive leaching agents of increasing strength and ability to interact with the sample matrix. Typical steps are summarized in Table 27.3. Representative extraction procedures have been described in a number of papers (Zeien and Bruemmer 1989; Ure et al. 1993). These procedures try to leach the elements from the different compounds and complexes in a stepwise fashion. Often, however, the selectivity is not high enough. Problems may arise because of the pH dependence of the extraction step. Sample matrices can also alter pH and therefore selectivity.

Table 27.3 Example of steps in sequential extraction

Steps	Remarks
1. Exchangeable, adsorbed	–
2. Carbonate fraction	Subsequent fractions are usually not interfered
3. Reducible fraction	This step is usually accomplished by using hydroxylamine–HCl/acetic acid; however, selectivity is doubtful
4. Oxidizable fraction	Possible binding mechanisms are adsorption, chelating, and complexation; differentiation from step 3 is often impossible
5. Residual fraction	For example, aqua-regia

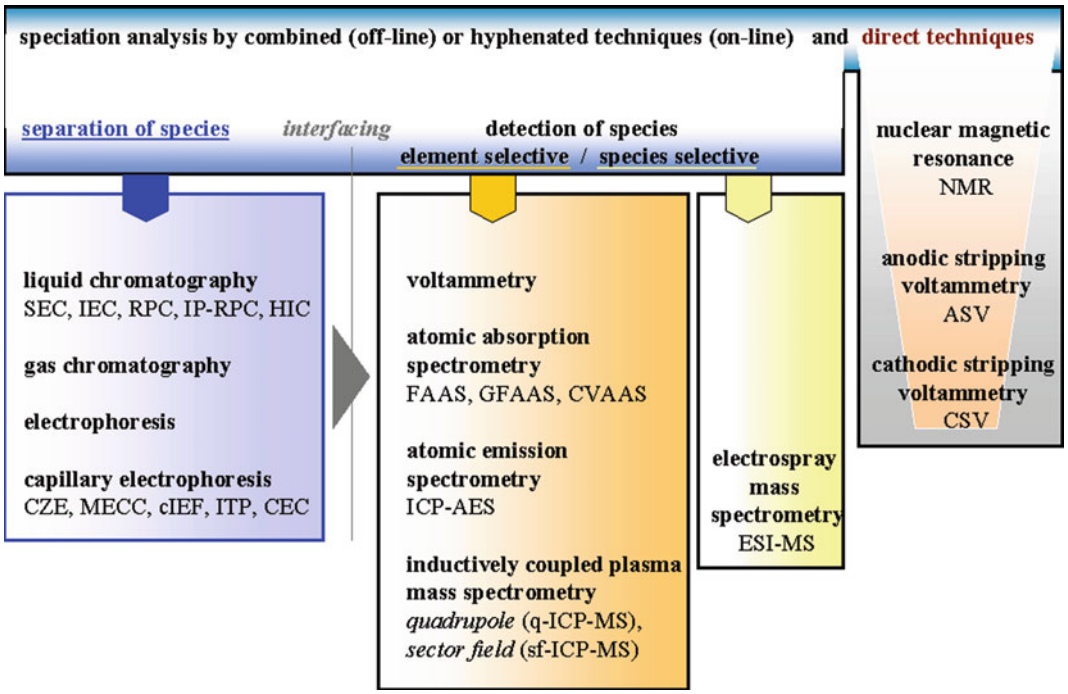


Fig. 27.1 Two speciation approaches are shown: one using direct methods, the other employing coupled techniques that include separation, interfacing/coupling, and selective detection. The latter are commonly used for element speciation due to flexibility and sensitivity

27.2.3 Speciation Approaches: Direct Speciation Methods or Combined (Hyphenated) Techniques

After sampling, storage, and sample preparation, species are identified and analyzed. Direct speciation approaches can provide full information about the species in a sample without any additional (separation) method; that is, they can directly quantify the species. Such methods include nuclear magnetic resonance (NMR) or X-ray absorption near edge spectroscopy (XANES). Structural information and in specific cases elemental information are gained by NMR, being thus an ideal speciation technique. However, the sensitivity is in most cases too low for real samples at biologically or ecologically relevant concentration levels. The most usual approach for element speciation is based on combined (or hyphenated) systems. Here, species are selectively separated and then the elements in the various chemical forms are selectively detected. For species identification and increased quality control, molecule-selective detection is coupled to separation devices. These combinations provide extended flexibility and a broad applicability. Disadvantages include the increased complexity and thus increased risk for failure of the systems. Also, species equilibria can be drastically altered during separation due to dilution, and some components can be removed from the chemical equilibrium. Species transformation and destruction are likely consequences; therefore, the total separation time should be short compared to the transformation rate of species. Figure 27.1 gives an overview of the various speciation methods. In the following section, separation mechanisms are described and their features are taken into account from the viewpoint of element speciation. The hyphenation of separation techniques to element detectors is also discussed.

27.2.3.1 Liquid Chromatography

One of the most important advantages of liquid chromatography (LC) is the ample selection of separation mechanisms and the use of various mobile and stationary phases which provide nearly all the necessary techniques for separation of element species; therefore, specific problem-related speciation analysis is often possible that meets the requirements for species stability and efficient separation. Preservation of original species information is at least as necessary as good separation and influences the choice of separation mechanisms and reagents. Many stationary phases and buffers or organic modifiers can denature native species. Chelating eluents or ion exchangers may cause recomplexation of free or labile bound metal species (Dunemann and Begerow 1995). A more general disadvantage is seen in the existence of a stationary phase, as compared, for example, to capillary electrophoresis (CE); the large surface allows adsorption effects, contamination, or miscellaneous alterations of the species to occur (Harms and Schwedt 1994).

The mobile phases, too, can cause alterations of species, even if they may be very effective in separation. Buffers can stabilize biomolecules, but they may also alter species equilibria. Complexing tendencies or input from metal contamination may also occur (Arnaud et al. 1992).

The commonly used LC separation techniques in speciation are size exclusion chromatography (SEC), ion-exchange chromatography (IEC), reversed-phase liquid chromatography (RPLC), ion-pairing chromatography (IPC) and increasingly now hydrophilic interaction chromatography (HILIC). Advantages of SEC are seen in the fact that samples of unknown molecular masses are characterized in a mass-calibrated chromatographic system. SEC is a gentle method of performing chromatographic separation and normally does not result in a loss of element species or on-column alterations, although the column exhibits limited peak capacity. For complex multicomponent samples, complete resolution of the peaks is normally not achieved. Szpunar (2000b) reported shifting retention times for some compounds beyond the final elution volume. The stationary phase is not totally uncharged; for example, electrostatic effects have been observed when analyzing cadmium species. Adsorption, hydrophobic interactions, and species-specific affinities or H-bridging have also been observed. IEC shows high separation efficiency and wide applicability, thus solving many speciation problems. The relative retention of the ionic species is determined by three variables: namely, pH, ionic strength of the mobile phase, and nature of the ion exchanger. On the other hand, hydrophobic interactions between the sample and the non-ionic carbon backbone of the stationary phase cause organic ions to be retained in a way typical for RPLC (Mikes 1988). The pore size of the resin particles is an important parameter for the success of the separation. Often, loosely bound metal ions are lost or replaced by other metals originating from the buffer (Quevauviller et al. 1996); thus, IEC is a good fit for the separation of covalent bound element species in different valence states, such as Cr(III)/Cr(VI) or Sb(III)/Sb(V)/methyl-Sb (Lintschinger et al. 1998). Another frequent application is for covalent bound Se species (Xu et al. 2008; Michalke and Berthele 2011) or a high number of arsenicals (Goessler et al. 1997; Nischwitz et al. 2006).

In RPLC the retention mechanism is based on an apolar stationary phase and the relative hydrophobicity of the analyte. The predominant advantage is the wide analytical spectrum available; this very effective separation technique provides high resolution of species, and the flexibility offered by the multiple mobile phases allows the addition of ion-pairing reagents for analysis of ionized molecules. Obtaining results is usually easy and fast. The normally high reproducibility is a significant advantage in speciation analysis (Mikes 1988). In HILIC *polar* stationary phases are used with buffer systems being identical to RPLC. However, in contrast to RPLC aqueous eluents are strongest. Gradients typically begin with high acetonitrile content (>70%) and end with high aqueous buffer amount. Therefore, this technique is called *inverse RPLC*. HILIC is an interesting alternative for the separation of all polar and hydrophilic compounds, irrespective of charge and molecular size. Separation problems can be solved which were unsolvable for RPLC.

However, in practice, the stationary phase can exhibit ion-exchange properties or undesired adsorption effects, especially for basic analytes. At pH values higher than 4, these basic analytes can be adsorbed tightly. Usually, two different eluents are necessary, at least one of them containing a considerable amount of organic modifier. The polar eluent often shows high complexing tendencies. Organic solvents and acids may easily change element species such as protein–metal complexes. The structure of proteins may be unfolded, and complex-bound elements are subsequently released. Released metals are likely to be recomplexed by other ligands, which results in species transfer reactions; hence, only those analytes with no loosely bound metals may be separated by RPLC technique. Species with covalent element-ligand bondings are best suited to this method; therefore, RPLC is often used in parallel or in combination with other smoother techniques as a part of the so-called orthogonal speciation concepts (Szpunar 2000a).

When performing RPLC, care should be taken to keep the pH below 7.5 to avoid hydrolysis of the stationary phase. Hyphenation to an element-selective technique, such as inductively coupled plasma–mass spectrometry (ICP–MS), causes further problems, particularly when RPLC is coupled online to ICP–MS. These problems are detailed in

Sect. 27.2.3.3. The use of ion-pairing reagents allows RPLC to analyze ionic species. Additional advantages of this method are self-evident; for example, flexibility and variability are drastically increased. Also, very efficient separation is possible for a wide range of analytes, and the separation conditions may be tailored for the specific separation task.

Problems and limitations of ion-pairing RPLC are also increased in element speciation and hyphenation to, for example, ICP–MS. Again, organic solvents and acids might be used that could easily change element species. The structure of proteins may be destroyed, and complex-bound elements can be released to be recomplexed by other ligands; thus, species transfer reactions are likely. Covalent element-ligand bondings are best suited for analysis by this method. Applications of RPLC include separation of metallo-proteins, e.g. metallothionein isoforms and superoxide dismutase (Nischwitz et al. 2003). The majority of applications uses ion-pair reagents like trifluoroacetic acid or tetraalkylammonium hydroxides in the mobile phase to increase the retention of ionic compounds, e.g. selenoamino acids (Letsiou et al. 2007), trivalent and hexavalent chromium (Wolf et al. 2007) or Pt-nucleotide complexes (Harrington et al. 2010).

But the use of ion-pairing reagents intensifies the problems seen in RPLC and analyte stability during separation becomes more difficult to maintain with ion pairing than with RPLC alone.

27.2.3.2 Capillary Electrophoresis

One of the most powerful separation techniques, capillary electrophoresis provides a very efficient separation of species and is often superior to LC separation techniques. CE is able to separate positive, neutral, and negative ions in a single run with high separation efficiencies. A single CE instrument can even offer several different modes of separation: capillary zone electrophoresis (CZE), micellar electrokinetic chromatography (MEKC), isoelectric focusing (IEF), isotachophoresis (ITP), and capillary electrochromatography (CEC). All of these modes rely on the application of high voltage (Kuhn and Hofstetter 1993); however, the separation principles are quite different, and they each provide a variety of characterization and identification mechanisms for elemental species. The latter point is of great importance, as species identification is rarely achieved by a single method; rather, it requires multidimensional strategies. Furthermore, species integrity is thought to be affected less than when HPLC is used. Limitations are seen in the very small sample volume used, thus giving rise to concerns with regard to the representativeness and homogeneity of the sample and instrumental detection limits (Michalke 1999b).

27.2.3.3 Interfacing LC to ICP–MS

Combining separation methods with element-selective detection methods leads to hyphenated systems. These online systems give easier and faster results; therefore the risk for contamination or loss is reduced. On the other hand, collected fractions allow for several quality control checks and application of the orthogonal identification concepts (Michalke and Schramel 1997). Interfacing LC with ICP–MS is achieved by a nebulizer. Several pneumatic nebulizers are available, such as the Meinhard (MN), cross-flow (CFN), microconcentric (MCN), ultrasonic (USN), direct injection (DIN), and hydraulic high pressure (HHPN) nebulizers. Nebulization efficiency ranges from around 1 to 5% for the MN and CFN to quite high values for the DIN and USN (Dunemann and Begerow 1995). The MCN and DIN are used for low flow rates between about 30 and 150 μL min^{-1} and are believed to achieve nebulization efficiency up to 100%. These are suitable systems for interfacing with microbore LC.

The major problems with online hyphenation include:

1. The salt load of eluents can cause problems, such as crusting and changes in ionization.
2. The use of high amounts of organic solvents in RPLC cools the plasma and increases the reflected radiofrequency (RF) power. This results in plasma extinction already at relatively low organic solvent concentrations. Further, the ionization characteristics of the argon plasma are altered, which affects the sensitivity of the detector for element species. The extremely high carbon intake induces polyatomic interferences and carbon precipitation on the torch and cones. The conductivity of carbon can cause flashovers from the coil to the carbon-coated torch.
3. Sample transfer to ICP–MS; This may be influenced by dead volume of the interface, or affected by peak broadening and flow rate.

Developments to address the first issue listed above are typically based on novel column technology that provides high separation efficiency even at low (buffer) salt concentrations. The easiest way to overcome the problems caused by the second issue is post-column dilution of the eluent; however, this results in dilution of the separated analytes. The most common method to stand high organic modifier concentrations is to reduce the evaporation pressure of the modifier by cooling the transfer line and/or the spray chamber below 10 °C. Methanol concentrations up to 60–80% in the eluent are tolerated at flow rates of about 1.5 mL min^{-1}. For improved burning of carbon from MeOH, the nebulizer gas is added with small amounts of oxygen. Desolvating systems are also used that can tolerate methanol concentrations up to 100% (Lustig et al. 1998); however, one should be aware that some species may be removed in the

desolvator. Further aspects of nebulizer behavior are discussed by Montaser and Golightly (1992).

27.2.3.4 Interfacing CE to ICP–MS

Much effort has been devoted to interfacing CE with ICP–MS. Nowadays, few commercial interfaces are available, each of them being based on low-flow-nebulizers. Often also laboratory made approaches are in use, typically constructed around commercial low-flow nebulizers. It has been demonstrated that such hyphenated CE techniques can provide sub-μg L^{-1} detection limits for the analysis of many types of environmental samples and are also capable of multiple-element monitoring of various metal functionalities. Nevertheless, the tiny amounts of sample emerging from the capillary often give rise to detection limits that are usually higher than those of conventional chromatographic methods. An exhaustive discussion on CE–ICP–MS was already given by Olesik (2000), who covers theory, application, and instrumentation of the procedure. One requirement for the interface is being able to close the electrical circuit from CE at the end of the capillary. The flow rate of CE generally does not match the flow rate needed for an efficient nebulization, but one possible way to circumvent this problem is to close the electrical circuit of CE during nebulization by applying a coaxial electrolyte flow around the CE capillary. The grounded outlet electrode is, in all cases, in contact with this electrolyte flow. The sheath flow has also been used to adapt the flow rate to a suitable nebulization efficiency. Optimization of nebulization efficiency was achieved by adjusting the flow rate when using MCN- or DIN-based systems or by exactly positioning the CE capillary to the point of nebulization (*e.g.*, by employing a micrometer screw).

27.2.3.5 Interfacing CE to ESI–MS

Electrospray ionization (ESI) interfaces for CE are commercially available. The closing of the electrical circuit from CE during ion evaporation is provided by an electrolyte sheath flow. Effective ion production is made possible by the use of a suitable spray voltage easily controlled by the instrument software.

27.2.3.6 Gas Chromatography

The choice of an adequate separation technique is determined by the physicochemical properties of the analyte (volatility, charge, polarity), whereas that of the detection technique is determined by the analyte level in the sample. The combination of gas chromatography (GC) with ICP–MS has become an effective method for the speciation of organometallic compounds in complex environmental samples. GC separates volatile and gaseous element species employing primarily capillary columns with bonded phases. The big advantage is seen in the fact that species no longer have to be transferred into an aerosol or gaseous phase. Sample input into, for example, ICP–MS is about 100%, thus improving detection limits. Very often, however, species are not sufficiently volatile, and in such cases time-consuming and tedious sample preparation, extraction, and derivatization procedures are necessary, especially when carrying out a Grignard derivatization (Quevauviller et al. 1996). This approach, on the other hand, removes the matrix components and makes separation easier.

Major problems are often encountered with derivatization, which is sometimes less selective than expected (Quevauviller et al. 1996). The detector response is species-derivative specific, an important aspect to consider in quantification; however, some derivatives have nearly no detector response. The separation typically is carried out at elevated temperature; thus, only thermally stable species can be handled by GC. Another problem arises when coupling to element-selective detectors such as ICP–AES or ICP–MS. This coupling is not as straightforward as with LC and is affected by several limitations; for example, analytes have to be maintained in the gas phase during their transport from the GC to the ICP–MS to avoid condensation effects. The transfer line to the plasma should also be heated, either by a pre-heated sheath gas or by electrical heating. The proximity of metal parts (heating wire) to the generator coil is problematic. Also, the effluent from the gas chromatograph (a few milliliters per minute) requires an additional carrier gas to achieve sufficient flow in the central channel of the plasma. For these reasons, most of the published papers devoted to the use of GC–ICP–MS address the construction and development of adequate interfaces. Typical applications of GC in speciation analysis are quantification of alkylated species of arsenic, mercury, lead, selenium, and tin.

27.2.3.7 Element-Selective Detection

Spectrometry-based techniques for the quantification of elements have long since taken root. Spectrochemical methods are frequently used because of their speed, detection power, sensitivity, and specificity. The impact that atomic spectrometric techniques in general have had on governmental institutions and international organizations has been carefully discussed by Minderhoud (1983), who made particular reference to the analysis of chemical wastes, sewage sludges, surface waters, and airborne particulates. The role of spectrochemistry in bioinorganic chemistry is still growing and will be even greater when chemical speciation can be fully accomplished.

Atomic absorption spectrometry (AAS) systems were introduced early to speciation analysis. AAS systems are comparatively inexpensive element-selective detectors which, however, are not that simply coupled on-line to separation devices: The detection power of flame-AAS is insufficient for normal environmental or physiological

concentrations (Dunemann and Begerow 1995). The sample intake is high (4–5 mL min^{-1}), which complicates online hyphenations with HPLC (optimized flow rates at about 0.5–2 mL min^{-1}); therefore, an auxiliary flow becomes necessary, which results in analyte dilution. Cold vapour (CV)-AAS and hydride generating (HG)-AAS use selective derivatization for matrix separation and detectability of relevant species, but the detector response is strongly species dependent and often easily interfered. Graphite furnace (GF)-AAS needs samples of only a few microliters and provides low detection limits of 0.1–5 μg L^{-1}. Matrix interferences are widely eliminated via Zeeman correction and matrix modifiers. An optimized temperature program is part of the determination, which is the reason for a discontinuous measurement; therefore, GF–AAS is unsuitable for online hyphenation to HPLC, as the chromatographic data points are gained only at intervals of a few minutes. This is too slow for peak description in a chromatogram. Moreover, during the temperature increase for sample drying, volatile species may be lost before being atomized.

Commonly used alternatives to AAS-centered approaches include ICP–AES and mainly ICP–MS detectors. The special diagnostic advantage of plasma-based techniques is their rapid screening ability, which has often confirmed suspected heavy metal poisoning (Ure et al. 1993), mineral deficiencies, or storage diseases. Equally as often, however, these techniques have identified completely unsuspected etiologic factors. Barnes (1991) surveyed the potential of ICP–AES combined with flow injection analysis, direct sample introduction and vaporization systems, electrothermal vaporization, hydride, metal vapor and gas generation, and chromatographic techniques. The big advantage of ICP–AES is offered by its multiple-element capability and sensitivity. Online hyphenation is easily set up. The ionization source is an inductively coupled plasma, and the temperature is around 5,000–9,000 K. The plasma (mostly argon) is formed within a quartz torch made up of three concentric quartz tubes, with the gas flowing at different rates through each. The outer flow is the highest and is known as the "plasma," "coolant," or "support" gas flow. It is tangentially introduced into the plasma torch to provide a helical flow that sustains the plasma itself. The central gas flow, known as the "auxiliary" gas flow, keeps the plasma away from the edge of the quartz torch. The inner gas flow, commonly called the "nebulizer" gas flow, transports the nebulized samples to the plasma. An RF field provides the energy to sustain the argon plasma, and the plasma transfers energy to the analyte(s) for excitation and ionization. Excellent discussions of plasma theory, mechanism, and applications are given by Montaser and Golightly (1992). Chemical interferences such as molecular emissions cause no major problems, but background correction should be applied in any case. Sample introduction is performed via a nebulizer and spray chamber. Nebulizer types, related problems, and some solutions have already been discussed earlier. Today, modern ICP–AES systems are equipped with a charge-coupled device (CCD) which provides on time multiple elements monitoring with maintaining full chromatographic resolution for any element line selected.

Inductively coupled plasma–mass spectrometry is now the technique of choice for a wide range of samples with element concentrations in the μg g^{-1} to sub-ng g^{-1} range. It has become a highly versatile technique with low detection limits and high sensitivity. Also, thanks to its element specificity, it is a technique of choice for chromatographic detection, including GC, LC, SFC, and CE. The striking advantages of ICP–MS techniques are their detection selectivity, multiple-element capability, and high sensitivity. Isotopic and elemental information of species is obtained, and species not totally resolved but pertaining to different elements are distinguished by the selective detector.

Sample introduction is performed through an interface connecting the LC system to the ICP–MS. The availability and pros and cons of such interfaces (various types of nebulizers) have been discussed previously. In hyphenated systems, detection limits for element concentrations in elemental species were reported to be in the range of 10–100 ng L^{-1} (Olesik 2000; Michalke et al. 2001).

ICP–MS can be of the quadrupole type (Q-ICP–MS) or the highly resolving sector field type (SF-ICP–MS). The latter is reported to improve detection ability by a factor of 10–100, or even up to 1,000 when equipped with a guard electrode (Prange and Schaumlöffel 1999). The quadrupole mass filter provides a resolution of only $m/\Delta m = 300$, resulting in a $\Delta m < 0.8$ for ions with a single positive charge, which is sufficient to distinguish between neighbouring isotopes ($\Delta(m/z) = 1$; m/z = mass/charge number). However, isobaric ions including polyatomic interferences cannot be resolved from the analyte ions: Therefore, polyatomic interferences are likely to occur, especially in the mass range of 40–80 amu. Well-known interferences are those of the $^{40}Ar^{35}Cl^{+}$ double ion on monoisotopic ^{75}As or of $^{40}Ar^{12}C^{+}$ on ^{52}Cr (Tittes et al. 1994). They are produced in the argon plasma when chlorine and carbon are introduced into the plasma (sample and buffer components, respectively). The highly resolving SF-ICP–MS can distinguish between the interference and the element isotope, as the mass resolution is 7,500–10,000 amu (as compared with 300 amu for Q-ICP–MS). An example for such separation is given in Fig. 27.2.

When using SF-ICP–MS at a high mass resolution, the detection power is reduced, and it falls into the same range of Q-ICP–MS. However, high resolution instruments are far more expensive than qICP-MS. Therefore, as an alternative modern quadrupole-based instruments use dynamic reaction cell (DRC) or collision cell (CC) technology to reduce

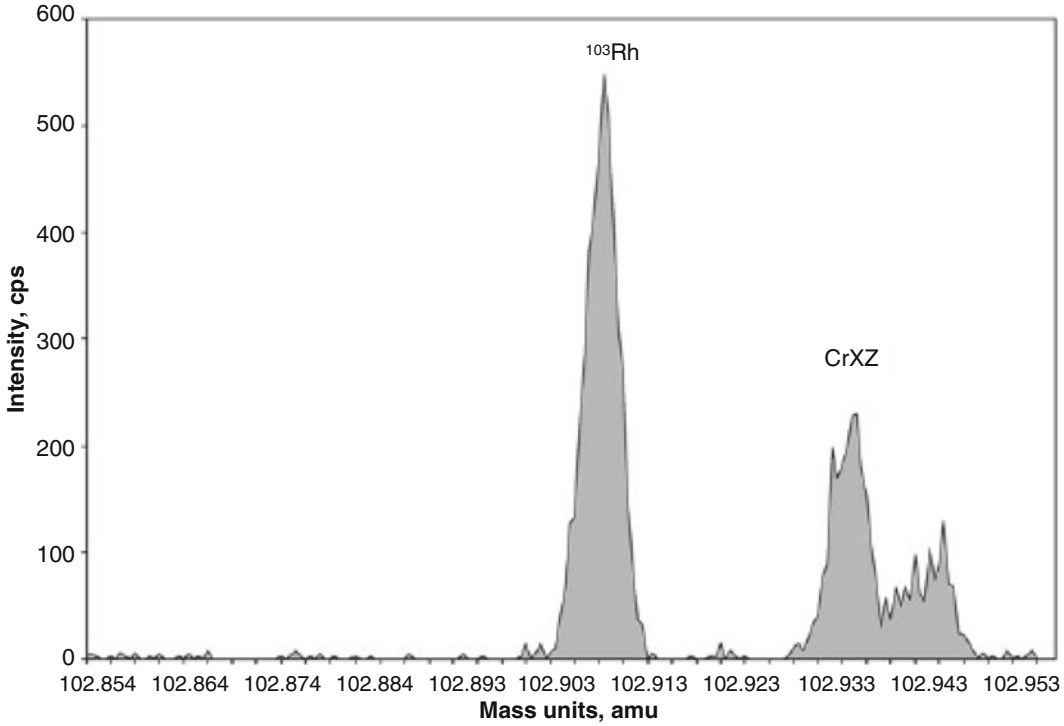


Fig. 27.2 ICP–MS detection may be affected by, for example, polyatomic interferences. Here, the resolution of a not clearly Identified Interference ^{52}Cr-XZ^{+} and ^{103}Rh by sf-ICP–MS is shown (From Krushevska et al. 2006, with permission)

spectral interferences by cluster ions. DRC uses a reaction gas (NH_3, O_2, CH_4) in the reaction cell, which is located between the ICP plasma and the mass analyzer, at very low flow rate of 0.2–0.7 mL min^{-1}. For example, $^{75}As^{+}$ reacts with oxygen to a clusterion of $[^{75}As^{16}O]^{+}$ which after passing the DRC can be monitored at m/z = 91 whilst the interference $[^{40}Ar^{35}Cl]^{+}$ remains at m/z = 75. In the case of $^{56}Fe^{+}$ determination the interfering $[^{40}Ar^{16}O]^{+}$ can be removed when using NH_3 as collision gas. A band pass filter between DRC and mass analyzer ensures that the cluster ions, which have lost more kinetic energy in the collision cell than the smaller analyte ions, are not entering the mass analyzer. Contrary, CC technology uses He or partially H_2 as a collision gas, also at very low flow rates. The $[^{40}Ar^{35}Cl]^{+}$ ions get disrupted and a properly adjusted band pass at the end of the CC rejects further movement of the collision product ions towards the mass analyzer. Both DRC and CC have been proven on results of high accuracy even at concentrations of few ng L^{-1} in diluted urine matrix (DFG 2012).

With element-selective detectors such as ICP–MS, one has to realize that only the element in the species is detected, not the entire molecule, which is advantageous because the separation of molecules bound to different elements does not need to be complete. They may be screened by the detector responding to different isotopes; however, it must be kept in mind that the molecule itself is not seen, and the species identification is only possible by comparing retention times. In natural samples, this is not always achieved beyond doubt; therefore, very often speciation analysis with only, for example, LC–ICP–MS is not enough for unequivocal speciation results. Multidimensional analytical concepts are strongly indicated in such cases (Michalke 1999a).

For the identification of polyatomic interferences, the monitoring of several isotopes of a given element can be helpful. Only when the natural isotope ratio is measured can interferences be ruled out. Unsatisfactory sensitivity is still a problem for very low concentration samples; hence, monitoring the most abundant isotopes of an element is recommended, except when these isotopes are interfered (e.g. poly-atomic interferences). On the other hand, ICP–MS is a sequential detector that monitors the programmed isotopes for several milliseconds. If too many isotopes are programmed for subsequent determinations, the detector operates too slowly to allow for highly resolved and fast-appearing peaks on one specific mass. This causes a loss in chromatographic resolution of the hyphenated system. The recently available time-of-flight (TOF) ICP–MS can overcome this drawback, as the mass filter does not jump in a time-consuming manner from one mass to another.

Instead, the different isotopes are distinguished as a function of their individual (*m*/*z*-dependent) time to reach the detector.

27.2.3.8 Species-Selective Detectors: Electrospray Ionization Mass Spectrometry Detection

Electrospray ionization (ESI) is a process that may preserve the whole species intact under optimal circumstances. ESI is suitable for extremely low flow rates. It is based on the so-called ion evaporation principle, where charged droplets of the analytes are transferred into the gas phase. A volatile buffer consisting of considerable amounts of, for example, methanol supports the ion evaporation process. In fact, the high volatilization capability of CE electrolytes is mandatory. The success of this detection method is based on the ability to produce multicharged ions from high molecular element species, such as metalloproteins, thus making the analysis of these compounds feasible up to molecular weights of 150,000–200,000 amu. The possibility of coupling this detector to LC or CE systems makes it extremely valuable. The soft ionization of element species finally allows preservation of the entire molecule (element species) when it is transferred into the gas phase and subsequent analysis by mass spectrometry (Cole 1997). Structural changes normally do not appear as long as covalent bonds are present. In special cases, stable element–organic molecules can be analyzed.

Selected Reaction Monitoring

When applying Collision Induced Dissociation (CID) together with a MS/MS system further structural information can be gained. The precursor ions are selected in the first quadrupole, and fragmented into characteristic product ions by collision with a gas in the second quadrupole. Those product ions, which have the same charge as the precursor ion, are then monitored by the third quadrupole (Q3). Such fragmentation processes are often very specific for the initially selected precursor ion, especially at optimized collision energy. When performing a full product ion scan on Q3 the sensitivity is in most cases low. However, when monitoring only few characteristic product ions for each precursor high sensitivity can be combined with high selectivity for monitoring of a large number of analytes. The sensitivity of this so-called Selected Reaction Monitoring (SRM) depends on the efficiency of the product ion formation in the collision cell and on the background intensity (noise) of the chosen reaction. In most cases the noise is low which allows high signal-to-noise ratio. Few other detection techniques are able to provide such detailed information about molecular weight and structure of the analysed compounds. Therefore, selected reaction monitoring is gaining increasing importance in speciation.

On the other hand, significant problems have been described for ESI. One problem arises due to the ion–solvent clusters. During the transfer of gas-phase ions into the high-vacuum (10^{-9} bar) zone, condensation of solvent molecules (*e.g.*, methanol, water) from the gas-phase ions is likely to happen. This is what is referred to as an ion–solvent cluster and is caused by the cooling-down phenomena that occur when a gas expands into a vacuum (*i.e.*, free jet expansion, adiabatic expansion). Ion–solvent cluster production results in the splitting up of one species into multiple signals, worsening detection limits, and increasing spectral complexity. Electrolytic processes at the metallic ESI tip needle that can generate new species or transformations of species (*e.g.*, by metal exchange) are also observed. When analyzing free metal ions such as Cu(II), multiple ion–solvent cluster signals are again detected. Most important, however, is the fact that native counter-ions of the metal ion are replaced by H_2O and/or methanol, independent of the counter-ion initially present (*e.g.*, $[Cu(MeOH)]^+$) instead of ($Cu^{2+}2Cl^-$). This implies a total loss of the original species information. In contrast to ICP-MS the ES ionisation is much stronger influenced by the sample matrix. For example, high salt contents or high amounts of coeluting organic compounds may suppress the signal substantially. Consequently, quantification of element species is often not possible by external calibration, but requires standard addition or use of isotopically labelled analytes for isotope dilution analysis.

27.3 Quality Control in Speciation

Causes of disagreement may be traced back to poor methodology, improper instrument calibration, faulty experimental techniques, impure reagents, or a combination thereof. Because of the lack of reliable data on trace elements in biological fluids, the reported diagnostic significance of some elements is controversial. The same holds for inaccurate measurements of pollutants in environmental matrices such as sediments, water, and particulate matter. Extremely sensitive instrumentation is readily available in laboratories not equipped to control contamination, and many users of a particular technique do not fully understand the limitations of a methodology. All of these factors introduce a great deal of questionable data. Much work on trace elements in human body fluids and tissues has suffered from methodology deficiencies, but accuracy is needed to reach rational conclusions upon which basic healthcare decisions are built. A rigorous program of quality control/quality assurance is needed to confirm the reliability of results for trace elements in biological materials.

27.3.1 General Aspects of Quality Control

Quality control (QC) plays a crucial role in element speciation. General hypotheses and analytical models depend on the reliability of data (Quevauviller et al. 1996). The key to successful speciation is the preservation of species information during the whole analytical procedure from sampling to final analysis. Actually, this is rarely guaranteed, and the range of errors is extremely high; thus, quality control is needed in the planning stages of an experiment. Pertinent literature must be sought and general criteria must be adapted to the actual problem relevant to the elements and matrices of interest.

All trace and ultratrace analyses require clean workplaces and the continuous use of appropriate control materials. These prerequisites are absolutely necessary for reliable results, and determination of the blank level will be that much more accurate. If these rules are not strictly followed, the money and time spent are completely wasted. For work with trace elements (and their chemical species) at very low concentration in biological materials, clean-room techniques may become mandatory. Other prerequisites are listed below:

1. *Identification*—A necessary prerequisite of reliable analytical work is a correct identification of the detected species. In particular when using separation techniques with element selective detection, the unambiguous identification based on the available standards is often critical because of matrix dependent changes of the retention times and potentially present (unknown) species in the sample which may co-elute with the available standards. Typically, spiking experiments are performed to ensure the matching identity of the species in the sample with the standard. Additionally multidimensional concepts and molecule-selective detection are valuable approaches to confirm the species identification.
2. *Calibration*—another indispensable prerequisite is correct calibration using calibrants of each investigated species with known stoichiometry. Preparation of two parallel calibrant solutions should be done according to weight and not volume. High purity of chemicals is compulsory. New lots of calibrants must be verified. Calibration graphs of the single species must be generated.
3. *Quantitative speciation*—species quantification is the basis for obtaining mass balances and gaining information about losses and contaminations. Quantification must be done using the relevant calibration curve (or standard for standard addition). If unknown compounds are monitored, quantification is not possible. Species can only be estimated by relating peak area calibration graphs to those of closely eluting known species.
4. *Certified reference materials*—one of the favored approaches for quality control is the use of reference materials (RMs) or certified reference materials (CRMs), although in environmental analysis there is a dismal lack of RMs to match real-world situations. The somewhat restricted market is certainly partially responsible for this. The introduction of more stringent regulations will enhance the demand for RMs. Operational complexity associated with the production of environmental RMs and the amounts required (5,000–1,000 kg) are also paramount factors.

Blending is an alternative procedure to prepare RMs of intermediate concentrations. Strategies to ensure accuracy include: (1) building accuracy into RMs under the guidance of a few centrally operating agencies, (2) transfer of accuracy from RMs to the measurement system, and (3) safeguarding the accuracy levels by continuous measurement quality control. Political and economical decisions concerning the environment are based on the correctness of analytical data. The last 30 years are testament to a growing awareness of the mistakes and pitfalls that accurate RMs should help eliminate. Suitable RMs and CRMs should be included at the earliest possible stage in any speciation analysis process. Confirmation of the certified value can prove the reliability of the results for unknown samples (Quevauviller et al. 1996); however, doing so never gives full evidence of the correctness of results, as sample matrices are rarely identical to the CRM matrix. The concentration range of the species or the species pattern may be different, which results in variation in their behavior during the analysis (Quevauviller et al. 1996). Many CRMs are available—for example, from the National Institute of Standards and Technology (NIST) in the United States, Institute for Reference Materials and Measurements (IRMM) in the EU, and National Research Council (NRC) in Canada. In practice, however, many other precautions are necessary, as summarized in Table 27.4.

Preservation of samples for future controls and investigations into their content with regard to unsuspected chemical species are well-recognized needs. Sample banking is a new concept that involves the preservation of important samples under unequivocal conditions that ensure their integrity over extended periods of time, but sample banking is expensive.

27.3.2 Quality Control Solutions for Clear Identification: Orthogonal Analytical Concepts

Analytical strategies that employ combinations of various separation and/or detection methods are referred to as *orthogonal analytical concepts*. They are an indispensable means for quality control in speciation and offer the best

Table 27.4 General criteria for speciation analysis

Sampling	Extent to which the samples are representative should be verified (Quevauviller et al. 1996)
	Sampling time should be kept short
	Contamination phenomena should be minimized
	Volume-to-surface ratio should be high to reduce wall effects
	Use of stainless steel tools is not advisable when sampling biological materials (Dunemann and Begerow 1995)
Sample preparation	Short storage at °C or freeze-drying (possible loss of volatile species) is recommended
	Extractions typically result in operationally defined speciation
	Mass balances and recovery rates (spiking of species) should be determined; species spiked can exhibit different extraction behavior
	Species spikes endanger native equilibria in the sample and could lead to a changed species pattern; comparison of different extractions is the best way to get reliable information (Quevauviller et al. 1996)
Derivatization	Derivatization should be avoided; it increases detection power but can affect the species
	Selectivity is much lower than expected unless other options are available
	Different efficiencies of derivatization and different detector responses of the derivatives are observed to some extent (Quevauviller et al. 1996)
Separation	Stationary phases may cause contamination and retain undesirable reactive groups
	Contamination or stability problems of species occur; these undesirable effects should be monitored by mass balances or checked by reinjection experiments (Michalke and Schramel 1997); potential species transfers should be investigated and possibly excluded
	Identification problems arise when the identity of a species is only attributed by retention times in a low resolving separation system
	Multidimensional analysis using various independent methods is the best alternative (Michalke 1999a; Szpunar 2000b)
Detection	Nonspecific detection should be avoided
	Calibration should be done properly
	Suitability of the detector for the problem at hand should be checked
	The discontinuous measuring nature of ETA–AAS results in too low chromatographic resolution; quadrupole ICP–MS exhibits maximum mass interference in the m/z range of 40–80
	Several isotopes of the same element should be monitored to check polyatomic interferences
Multidimensional analytical concepts	Combining various separation techniques and detection systems is recommended
	Two schemes are generally employed:
	1. Separation methods based on different separation principles are combined. After the first separation, fraction collection provides aliquots to be used for element determination and as samples for the second separation, which is again followed by online ICP–MS monitoring. This procedure provides an orthogonal characterization of molecules
	2. Separation is monitored in parallel by ICP–MS and ESI–MS. In this case, ICP–MS provides the element information and ESI–MS (or MS/MS) gives molecular or structural information; hence, obtaining maximum species information and sometimes identification becomes possible. These schemes can also be combined in different ways (Michalke 1999a; Szpunar 2000b)

opportunity for obtaining accurate speciation results and even identification of unknown species. In analytical systems with (only) one separation and one detection system, the risk of coelution, impossibility of species identification, or misidentification is high (McSheehy et al. 2002). This problem can be solved by employing different systems in various ways.

In short, these multidimensional analytical concepts rely upon combinations of various separation technologies and detection systems. Two schemes are primarily employed:

1. Two separation methods based on different separation principles are combined in a series. After the primary separation, fraction collection provides aliquots to be used for element determination and as samples for the secondary separation, which is again followed by an (online) ICP–MS monitoring. This results in an orthogonal characterization of molecules.
2. The separation is monitored in parallel by ICP–MS and ESI–MS. Here, ICP–MS provides the element information while ESI–MS (or ESI–MS/MS) delivers the molecular or structural information; hence, maximized species information and sometimes identification are possible.

Often, it is necessary to combine the two schemes in different ways (Michalke and Schramel 1997; Chassaigne and Lobinski 1998; Michalke 1999a; Szpunar 2000a). Figure 27.3 provides an orthogonal flow chart, and a current example has been published by McSheehy et al. (2002). Similarly, Nischwitz et al. (2003) checked species preserving extraction techniques using an orthogonal scheme, first using SEC with ICP–MS, fraction collection with

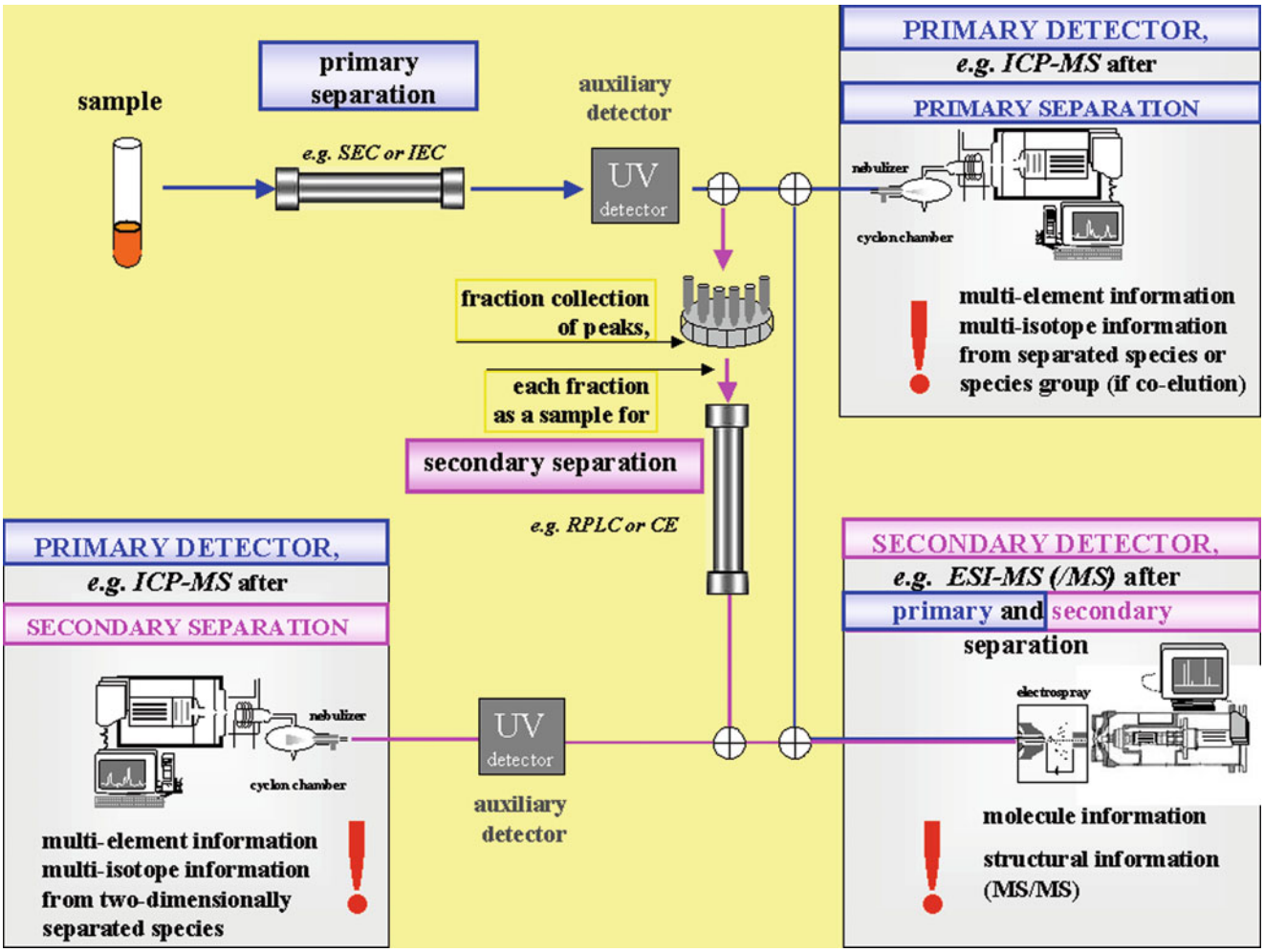


Fig. 27.3 The principles of an orthogonal (analytical) speciation concept are shown. General advantages include the ability to obtain multi-element/multi-isotope information using element-selective detection after the first species separation combined with molecular (and probably structural) information provided by a second molecule-specific detector. In case of coelution after the primary separation step, further purification is achieved by a secondary separation step, again providing multi-element/multi-isotope and molecule-selective information. Clear species identification is usually possible

subsequent RPLC–ICP–MS, and, finally, after collecting cleaned fractions from the latter, ESI–MS detection. But even more difficult is the identification of unknown species when no matching standard is available. Using elemental detection only, there is the option to monitor additional elements, for example S, P, which may provide more information about the type of compound or ligand. This can help in the case of thio-arsenic species or Pt-nucleotide adducts. Advances in collision cell ICP-MS technology improved the sensitivity for S and P, but still detection limits are at least 100 times worse than for elements like As and Pt which means that this approach is only possible for species present at sufficiently high concentration. In any case molecular mass spectrometry, for example electrospray mass spectrometry (ESI-MS/MS), should be applied in parallel to elemental detection for verifying the identity of unknowns independent on the availability of retention time matched standards.

27.3.3 Quality Control Solutions for Quantitative Species Determination

Further sources of error and uncertainty are the purity, the stability and the quantitative characterisation of the used standards. Even when using standard addition for quantification the column recovery should be determined to check if a significant percentage of the injected amount of the element of interest is remaining on the HPLC column. Low recoveries may indicate that some species of the element are not eluted at all under the applied conditions and thus are escaping detection. For example the thio-arsenosugars were found as new group of arsenic species due to further investigations of the reasons for low column recoveries (Fricke et al. 2004). In case there is low recovery of the species of interest improved mobile phase conditions and/or regular column cleaning is required to avoid artifacts from

random elution of accumulated analyte species from the column. Finally, low column recovery can be due to species transformation on the column. For example metal ions bound to proteins or other organic ligands via coordination bonds may be released due to stronger interaction of the metal ion with the stationary phase than with the original ligand. Another aspect for artifacts in quantification is interspecies conversion due to redox reactions, as has been found for trivalent and hexavalent Cr species.

Species specific isotope dilution techniques can eliminate most of the discussed sources of errors in quantification provided that the isotopically labeled species has the same identity as the analyte, is available in high purity and stability and that the equilibration between the natural species (analyte) and the isotopically labeled species is achieved at the earliest possible stage of sample preparation. Quantification via the isotope ratios can then compensate for analyte losses during sample preparation, for matrix dependent changes in sensitivity, for low column recovery and also for species transformation provided that there is still a clearly detectable peak of the analyte for accurate measurement of the isotope ratio(s). However, also this quantification strategy requires careful quality control and method development for accurate results. In particular the monitored isotopes need to be free from interferences and corrections for mass bias and detector dead time are required. In addition when using single IDMS the concentration and isotopic abundance of the spike solution of the isotopically labeled analyte needs to be exactly known or when using exact matching double spiking IDMS the isotopic abundance of the labeled spike and the exact concentration of a natural standard of the target species needs to be known (Nischwitz and Pergantis 2005). Wherever possible certified reference materials for the target species in the same or at least similar matrix should be included in the method validation.

27.4 Selected Examples of Speciation Studies on the Life Sciences

27.4.1 Environmental Applications

Environmental analysis deals with the detection of a variety of substances naturally or artificially present in our total environment and known or thought to exert adverse effects on human health. Environmental measurements are a special class of determinations with common problems. Ordinarily, data have large uncertainties, and the resulting decisions are controversial. Accuracy in environmental analysis is still a problem, notwithstanding the growing awareness of its importance. Trace element measurement quality is still far from satisfactory. Environmental data are obtained for the purpose of information and/or action. Since the industrial revolution, mankind has continued to pollute the Earth's biosphere with toxic substances such as heavy metals and halogenated organic compounds. Many of these substances are not amenable to processes that would lead to their removal; as a result, many industrial chemicals build up in the environment. During the past 100 years, the environment has been severely polluted by arsenic, cadmium, mercury, lead, and thallium. The increasing interest of governmental agencies, particularly the Environmental Protection Agency (EPA), in ICP–AES as a novel and useful alternative to established procedures for monitoring elements in environmental media was stressed 20 years ago in a report on the potential future of this technique (Barnes 1983).

Large amounts of impurities (both natural and industrial) can be found in the atmosphere. Some of them form flying ashes and other gaseous emissions with numerous elements, the toxicity of which depends on their chemical species. Limit values in regulations must be checked regularly, and levels of pollutants must be known in order to identify systematic changes and pollution sources.

The well-documented risks posed by concentrations of mercury in the environment are caused by human activity. Although the annual world consumption of mercury by industry is estimated to be around 10,000 tonnes, the total worldwide release of mercury as a result of human activities has been estimated to be 20,000–35,000 tonnes per year. It is unlikely that such releases can significantly increase the average mercury concentrations in the oceans, although increased levels of mercury can occur regionally and locally because of the release of mercury compounds into the environment. Water courses passing through areas where mercury-rich minerals are found (the main mercury minerals are cinnabar and metacinnabar, the two polymorphs of HgS) are known to result in elevated mercury levels. In the aquatic food chain, mercury plays an important role. Because of bioaccumulation processes, the mercury content in marine organisms is normally quite high compared to the content in water, and concentration factors of 10^4 and higher have been reported. The health hazards associated with mercury pollution of environmental waters were first brought to light in the early 1960s due to the Minamata mass poisoning catastrophe of several hundred people (mainly fishermen and their families) by methylmercury. The poisoning was caused by the consumption of fish and shellfish caught in Minamata Bay, which had been contaminated by industrial mercury discharge (see also Chap. 25, this volume).

Three categories of methodologies for environmental analysis can be defined as: (1) definitive techniques of high precision and accuracy—and high cost; (2) more routine techniques to be used for continuous monitoring of exposure to pollutants; and (3) field methods for preliminary semi-quantitative assessment under emergency circumstances requiring urgent countermeasures.

Marine- and freshwater have been significantly contaminated by organotin compounds, such as mono-, di-, and trisubstituted butyltin and phenyltin, because of their use as agrochemicals, biocides, and domestic products. The performance of four specific devices—namely, flame photometric detector, pulsed flame photometric detector, microwave-induced plasma atomic emission spectrometer, and inductively coupled plasma mass spectrometer—was compared in terms of ability to quantify the species of interest after separation by means of solid-phase microextraction and GC (Aguerre et al. 2002). The different approaches showed that the determination of ultratrace tin species is possible in natural waters even for routine purposes. Detection limits as good as 0.5 ng L^{-1} tin could be reached.

A detailed description of the complexation chemistry of copper in natural waters, with particular reference to water quality criteria and to possible effects on biota, was reported by Allen and Hansen (1986). It was concluded that the bioavailability and toxicity of Cu depend primarily on pH and alkalinity. When these two parameters are constant in a given system, then bioavailability and toxicity are proportional to the concentration of free copper ions and inorganic Cu complexes.

27.4.2 Nutritional Applications

It is a generally valid principle that each nutrient at excessive concentrations can be toxic; conversely, any of the trace elements now known for their toxicity might be shown in the future to have an essential function at low concentrations. Among the trace elements of nutritional interest are arsenic, chromium, manganese, molybdenum, nickel, selenium, silica, tin, vanadium, and perhaps cadmium. These elements present serious problems of analysis in the concentration ranges that are of interest to the nutritionist, either as toxic species even at low concentrations (*e.g.*, As (III)) or as essential element species.

The initial recognition that a micronutrient is essential for animal species raises questions about its nutritional and public health importance for humans. Nutritionists approach these questions by working through some specific tasks, such as identification of the metabolic parameters in humans that might be affected by the micronutrient, detection of disease states that can be prevented or cured by supplementation, determination of the human requirement for the micronutrient, assessment of the risk for dietary imbalances in population groups, and, finally, development of methods for assessing the nutritional status in individuals. All of these tasks, except for the last one, have been more or less accomplished.

Zinc, for example, plays a key role in erythrocyte carbonic anhydrase, an enzyme catalytically involved in the transport of CO_2 in blood. From a nutritional viewpoint, it is interesting that zinc deficiency is typically seen in population groups living in poverty. Consumption of vegetarian phytate-rich food is common and provides zinc–phytate complexes (species) of reduced bioavailability (Brätter et al. 1992). An analogous situation is known from formulas for newborns as compared to human milk. Formulas based on cows' milk contain higher amounts of casein, calcium, and phosphorus. These compounds together associate to casein–phosphorus–zinc micelles of low availability for zinc. Contrary to this, human milk shows zinc–citrate complexes that are easily accessed by the newborn's gut.

The ability of surfactants to differentiate between methylmercury and inorganic Hg(I) in fish-egg oil was ascertained by using an online, time-based injection system in conjunction with CVAAS (Burguera et al. 1999). An advantageous flow could be obtained from the highly viscous sample by injecting it into a three-phase surfactant (Tween 20®)–oil–water emulsion. Quantities of mercury as low as 0.1 μg L^{-1} could still be measured. The range of organic mercury concentrations in the catfish-egg oil sample was found to be 2.0–3.3 μg L^{-1}.

27.4.3 Biomedical Applications

The overall importance of analytical atomic spectrometry in biology and medicine was reviewed by Dawson (1986). While this author recognized the fundamental role played by ICP–AES and other spectroscopic techniques in generating accurate and reliable data, he also emphasized that the next step in the analysis of elements of well-established biological importance would be a growing emphasis on the identification and quantification of their biologically active species through coupling with adequate online separation procedures.

The importance of trace elements and the identification and quantification of their chemical forms cannot be overestimated. Some of the elements essential to humans (cobalt, chromium, copper, iron, iodine, manganese, molybdenum, nickel, selenium, silica, and zinc; see Brätter et al. 1992) have a vitamin character. Others are present in pharmaceuticals as active agents (aluminum, arsenic, gold, bismuth, copper, iron, mercury, lithium, platinum, and zinc). Uncontrolled use of pharmaceuticals can lead to undesired effects and intoxications, thus requiring immediate control therapy.

The human requirement for chromium is lower than that for any of the essential trace elements, except for cobalt as a part of vitamin B_{12}. Chromium is not only an essential ultratrace element but also a potent carcinogen. Many other essential trace elements are also toxic at excessive exposures, but chromium is unique in that its essentiality is

limited to one valence state and toxicity to another, and transformation from the essential to the toxic valence state does not occur in the living organism. Those and other properties of chromium that influence its interaction with biological systems reside in its chemical behavior. Chromium in the trivalent state is an essential element for animal species and humans; in the form of certain hexavalent compounds it is a potent carcinogen. Hexavalent compounds are manmade and do not occur naturally in living organisms. They penetrate biological membranes and are reduced by organic matter, which leads to oxidative damage of cell structures. While there is little evidence for a role of Cr (III) in enzyme systems, its interactions with nucleic acids and with the functions of the thyroid gland and of insulin have been demonstrated. The lack of a simple diagnostic procedure, which chromium shares with many other essential trace elements, is the main impediment to the wide application of chromium in medicine and nutrition. Without diagnosis of chromium status, the response of an individual to supplementation is unpredictable.

Among the essential trace elements, selenium is receiving increasing attention as a natural cancer-preventing agent. The anticarcinogenic effects of selenium have been demonstrated in numerous animal tumor model systems as well as under conditions simulating human dietary conditions (Patching and Gardiner 1999). Because the positive effects of selenium on Kashin-Beck disease are known, its protective mechanisms on several heart diseases, predominantly cardiovascular damage or congestive cardiomyopathy, have been widely investigated (see also Chap. 16, this volume). Detoxification effects of selenium have been proven and described for various metals such as arsenic, thallium, silver, cadmium, and mercury. Selenium deficiency is most critical for the brain and growth of infants. Furthermore, the thyroid metabolism may be impaired, because many deiodinases are Se proteins. These positive effects of selenium species led to several studies on selenium speciation in supplements, food (Quijano et al. 2000), and body fluids such as human milk (Michalke and Schramel 1997) and serum. The selenium speciation in serum is expected to mirror the available concentrations of relevant protective selenium species. As an example, Fig. 27.4 shows various chromatograms monitoring ^{82}Se after SAX separation from children's sera. Although at present not fully explained, these chromatograms from the sera of children with absorption abnormalities in the gastrointestinal tract show different patterns of selenium species as well as different total selenium levels. Gaining a full understanding of possible interrelations between alterations in the gut and the selenium species pattern will be a demanding task in the future.

Iodine, long known to act beneficially in human health as an essential micronutrient, is utilized by the thyroid gland for the biosynthesis of the thyroid hormones thyroxin (T_4) and triiodothyronine (T_3) (Keller 1991) (see also Chap. 6, this volume). These hormones strongly influence an extended range of biochemical reactions. Immune defense and antibody production are dependent on reliable thyroid function and the availability of T_4 and T_3 hormones. The speciation of various iodine species in serum or urine provides information about malfunction of the thyroid gland and can also explain other T_4/T_3-influenced metabolic abnormalities. The superiority of hyphenated techniques with ICP–MS (iodine-) detection over monoclonal antibody systems has recently been demonstrated, as the monoclonal antibody systems were generally unable to distinguish between iodinated (active) and non-iodinated hormones (Michalke et al. 2000). Recently, iodine speciation was reported during investigations into the disruption of normal thyroid function by xenobiotic chemicals.

The increased rate at which zinc-containing metalloenzymes have been identified in the past is largely due to the development of highly precise, rapid, and convenient methods for determining this element. Definitive knowledge that zinc is indispensable to living matter has emerged only in the last few decades. In succession, the biological effects of zinc have been viewed as mostly harmful, then questionable, then essential; however, it is now well established that zinc is essential for the growth and development of all living forms.

There is a lack of reliability in the techniques for aluminum determination that can affect a large number of samples encountered in clinical practice. Better measures for the quality control of such determinations have progressively led to improved detection limits for aluminum in biological material, thus paving the way to more reliable figures for the concentrations of their species of clinical relevance. Human serum was incubated *in vitro* with the radiotracers ^{51}Cr(III), ^{191}Pt(I), or carrier-free ^{48}V(V) (Lustig et al. 1999), and the protein-bound metals were measured by flatbed electrophoresis followed by autoradiography with laser densitometry, followed by subsequent detection of the proteins through silver staining. At this stage of its development, however, this proposed technique, although highly promising, does not possess adequate detection power.

27.5 Summary

The future of speciation analysis depends to a large extent on three major factors: (1) development of instrumentation actually designed for this purpose (and not simply assembled from apparatuses originally conceived for another purpose), (2) production of a substantial number of CRMs especially prepared for chemical species (and not only the total content of a given element), and (3) ability to transfer the relevant

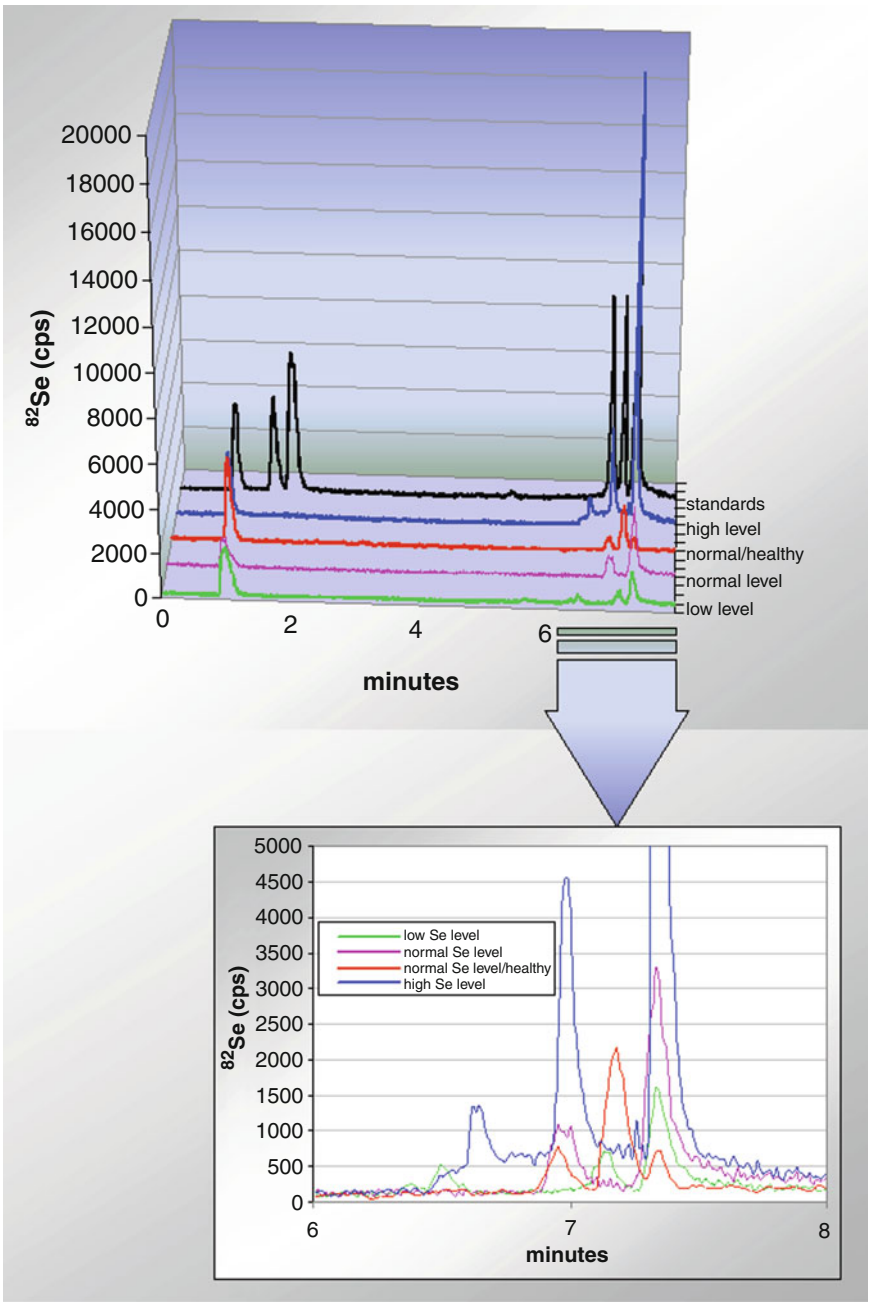


Fig. 27.4 Chromatograms of children's sera ^{82}Se after SAX separation are shown (with the region between 6 and 8 min retention time additionally magnified). These sera from children having different absorption abnormalities in the gastrointestinal tract show different patterns of Se species as well as different total Se levels. For comparison, a 20-µg Se L^{-1} standard chromatogram is also plotted

know-how from the expert laboratory to the routine laboratory. Currently, the most suitable speciation analysis combines systems primarily based on the hyphenation of separation technologies online with element- or molecule-specific detectors. The variety of separation principles available today allows the separation of most species present in liquid samples. The selective detection ability of ICP–MS provides only signals of interest (not all compounds can be resolved) with very high detection sensitivity; however, there are still several pitfalls, so strict quality control and quality assurance schemes must be implemented. In speciation analysis, one of the most promising approaches is based on the orthogonal scheme. The achievement of the above three goals will allow decision makers to develop regulations incorporating more effective, current knowledge on chemical elements and their role in the life sciences to the full benefit of human health and the environment.

References

Aguerre S, Lespes G, Desauziers V, Potin-Gautier M (2002) Speciation of organotin in environmental samples by SPME–GC: comparison of four specific detectors: FPD, PFPD, MIP-AES and ICP–MS. J Anal At Spectrom 16:263–269

Allen HE, Hansen DJ (1986) The importance of trace metal speciation to water quality criteria. Water Environ Res 68:42–54

Arnaud J, Andre D, Bouillet MC, Kia D, Favier A (1992) Problems associated with the use of exclusion diffusion chromatography for identification of zinc ligands in human milk. J Trace Elem Electrolyte Health Dis 6:81–90

Barnes RM (1983) Frontiers in inductively coupled plasma spectroscopy. Chem Anal Wars 28:179–198

Barnes RM (1991) Inductively coupled and other plasma sources: determination and speciation of trace elements in biomedical applications. In: Subramanian KS, Iyengar GV, Okamoto K (eds) Biological trace element research—multidisciplinary perspectives. American Chemical Society, Washington, DC, pp 158–180

Brätter P, Forth W, Fresenius W, Holtmeier HJ, Hoyer S, Kruse-Jarres J, Liesen H, Mohn L, Negretti de Brätter V, Reichlmayr-Lais AM, Sitzer G, Tölg G (1992) Mineralstoffe und Spurenelemente, Leitfaden für die ärztliche Praxis. Verlag Bertelsmann Stiftung, Gütersloh

Burguera JL, Quintana IA, Salager JL, Burguera M, Rondón C, Carrero P, Anton de Salager R, Petit de Peña Y (1999) The use of emulsions for the determination of methylmercury and inorganic mercury in fish-eggs oil by cold vapor generation in a flow injection system with atomic absorption spectrometric detection. Analyst 124:593–599

Caroli S (1996) Chemical speciation: a decade of progress. In: Caroli S (ed) Element speciation in bioinorganic chemistry. Wiley, New York, pp 1–18

Caroli S, Alimonti A, Coni E, Petrucci F, Senofonte O, Violante N (1994) The assessment of reference values for elements in human biological tissues and fluids: a systematic review. CRC Crit Rev Anal Chem 24:363–398

Caroli S, Cescon P, Walton D (eds) (2000) Environmental contamination in Antarctica: a challenge to analytical chemistry. Elsevier Science, Amsterdam

Chassaigne H, Lobinski R (1998) Speciation of metal complexes with biomolecules by reversed-phase HPLC with ion-spray and inductively coupled plasma mass spectrometric detection. Fresenius J Anal Chem 361:267–273

Cole RB (ed) (1997) Electrospray ionization mass spectrometry—fundamentals, instrumentation and applications. Wiley, New York

Crews HM, Clarke PA, Lewis DJ, Owen LM, Srutt PR, Izquierdo A (1996) Gastrointestinal extracts of cooked cod by high performance liquid chromatography–inductively coupled plasma mass spectrometry and electrospray mass spectrometry. J Anal At Spectrom 11:1177–1182

Dawson JB (1986) Analytical atomic spectroscopy in biology and medicine. Z Anal Chem 324:463–471

Diederich J, Michalke B (2010) Enhanced extract preparation for manganese and iron speciation in brain and liver tissue. Anal Bioanal Chem 399:1799–1806

Dunemann L, Begerow J (eds) (1995) Kopplungstechniken zur Elementspeziesanalytik. VCH, Weinheim

Florence TM (1989) Electrochemical techniques for trace element speciation in waters. In: Batley GE (ed) Trace element speciation: analytical methods and problems. CRC Press, Boca Raton, pp 77–116

Fricke MW, Creed PA, Parks AN, Shoemaker JA, Schwegel CA, Creed JT (2004) Extraction and detection of a new arsine sulfide containing arsenosugar in molluscs by IC-ICP-MS and IC-ESI-MS/MS. J Anal At Spectrom 19:1454–1459

Goessler W, Schlagenhaufen C, Kuehnelt D, Greschonig H, Irgolic K (1997) Can humans metabolize arsenic compounds to arsenobetaine? Appl Organomet Chem 11(4):327–335

Harms J, Schwedt G (1994) Applications of capillary electrophoresis in element speciation analysis of plant and food extracts. Fresenius J Anal Chem 350:93–100

Harrington CF, Le Pla RC, Jones GDD, Thomas AL, Farmer PB (2010) Determination of cisplatin 1,2-intrastrand guanine-guanine DNA adducts in human leukocytes by high-performance liquid chromatography coupled to inductively coupled plasma mass spectrometry. Chem Res Toxiocol 33:1313–1321

Keller H (1991) Klinische-chemische Labordiagnostik für die Praxis. Analyse, Befund, Interpretation. Georg Thieme Verlag, Stuttgart

Kersten M, Förstner U (1989) Speciation of trace elements in sediments. In: Batley GE (ed) Trace element speciation: analytical methods and problems. CRC Press, Boca Raton, pp 245–318

Krushevska AP, Zhou Y, Ravikumar V, Kim Y-J, Hinrichs J (2006) Chromium based polyatomic interferences on rhodium in ICP-MS. J Anal At Spectrom 21:847–855

Kuhn R, Hofstetter-Kuhn S (1993) Capillary electrophoresis: principles and practice. Springer, Berlin

Letsiou S, Nischwitz V, Traar P, Francesconi KA, Pergantis SA (2007) Determination of selenosugars in crude human urine using high-performance liquid chromatography/ atmospheric pressure chemical ionization tandem mass spectrometry. Rapid Commun Mass Spectrom 21:343–351

Lintschinger J, Schramel O, Kettrup A (1998) The analysis of antimony species by using ESI–MS and HPLC–ICP–MS. Fresenius J Anal Chem 361:96–102

Lustig S, Michalke B, Beck W, Schramel P (1998) Platinum speciation with hyphenated techniques: application of RP–HPLC–ICP–MS and CE–ICP–MS to aqueous extracts from a platinum treated soil. Fresenius J Anal Chem 360:18–25

Lustig S, Lampaert D, De Cremer K, De Kimpe J, Cornelis R, Schramel P (1999) Capability of flatbed electrophoresis (IEF and native PAGE) combined with sector field ICP–MS and autoradiography for the speciation of Cr, Ca, Ga, In, Pt and V in incubated serum samples. J Anal At Spectrom 14:1357–1362

McSheehy S, Szpunar J, Lobinski R et al (2002) Characterization of arsenic species in kidney of the clam *Tridacna derasa* by multidimensional liquid chromatography–ICPMS and electrospray time-of-flight tandem mass spectrometry. Anal Chem 74 (10):2370–2378

Michalke B (1999a) Quality control and reference materials in speciation. Fresenius J Anal Chem 363:439–445

Michalke B (1999b) Potential and limitations of capillary electrophoresis–ICP–mass spectrometry. J Anal At Spectrosc 14:1297–1302

Michalke B, Berthele A (2011) Contribution to selenium speciation in cerebrospinal fluid samples. J Anal At Spectrom 26:165–170

Michalke B, Nischwitz N (2010) Review on metal speciation analysis in cerebrospinal fluid—current methods and results. Anal Chim Acta 682(1–2):23–36

Michalke B, Schramel P (1997) Selenium speciation in human milk with special respect to quality control. Biol Trace Elem Res 59:45–56

Michalke B, Witte H, Schramel P (2000) Iodine speciation in human serum by reversed phase liquid chromatography–ICP–mass spectrometry. Biol Trace Elem Res 78:81–92

Michalke B, Witte H, Schramel P (2001) Developments of a rugged method for selenium speciation. J Anal At Spectrom 16:593–597

Michalke B, Witte H, Schramel P (2002) Effect of different extraction procedures on the yield and pattern of Se species in bacterial samples. Anal Bioanal Chem 372:444–447

Mikes O (1988) High-performance liquid chromatography of biopolymers and biooligomers. Elsevier, Amsterdam

Minderhoud A (1983) Atomic spectrometry for trace determinations in the environment with a view to government regulations. Spectrochim Acta 38B:1525–1532

Montaser A, Golightly DW (1992) Inductively coupled plasmas in analytical atomic spectrometry, 2nd edn. VCH Publishers, New York

Morrison GMP (1989) Trace element speciation and its relationship to bioavailability and toxicity in natural waters. In: Batley GE (ed) Trace element speciation: analytical methods and problems. CRC Press, Boca Raton, pp 25–42

Mota AM, Simaes Gonçalves ML (1996) Direct methods of speciation of heavy metals in natural waters. In: Caroli S (ed) Element speciation in bioinorganic chemistry. Wiley, New York, pp 21–96

Nischwitz V, Pergantis SA (2005) Liquid chromatography online with selected reaction monitoring electrospray mass spectrometry for the determination of organoarsenic species in crude extracts of marine reference materials. Anal Chem 77:5551–5563

Nischwitz V, Michalke B, Kettrup A (2003) Investigations on species-preserving extraction from liver samples. Anal Bioanal Chem 375:145–156

Nischwitz V, Kanaki K, Pergantis SA (2006) Mass spectrometric identification of novel arsinothioyl-sugars in marine bivalves and algae. J Anal At Spectrom 21:33–40

Olesik JW (2000) Capillary electrophoresis for elemental speciation studies. In: Caruso JA, Sutton KL, Ackley KL (eds) Element speciation: new approaches for trace element analysis. Elsevier, Amsterdam, pp 151–211

Patching SG, Gardiner PHE (1999) Recent developments in selenium metabolism and chemical speciation: a review. J Trace Elem Med Biol 13(4):193–214

Pellegrino C, Massanisso P, Morabito R (2000) Comparison of twelve selected extraction methods for the determination of butyl- and phenyltin compounds in mussel samples. Trends Anal Chem 19(2-3):97–106

Pickering WF (1981) Selective chemical extraction of soil components and bound metal species. CRC Crit Rev Anal Chem 12:233–238

Potin-Gautier M, Gilon N, Astruc M, de Gregori I, Pinochet H (1997) Comparison of selenium extraction procedures for its speciation in biological materials. Int J Environ Anal Chem 67:15–25

Prange A, Schaumlöffel D (1999) Determination of element species at trace levels using capillary electrophoresis–inductively coupled plasma sector–field mass spectrometry. J Anal At Spectrom 14:1329–1332

Quevauviller P, Maier EA, Griepink B (1996) Quality control of results of speciation analysis. In: Caroli S (ed) Element speciation in bioinorganic chemistry. Wiley, New York, pp 195–222

Quijano MA, Moreno P, Gutierrez AM, Perez-Conde C, Camara C (2000) Selenium speciation in animal tissues after enzymatic digestion by HPLC coupled to ICP–MS. J Mass Spectrom 35:878–884

Rodriguez-Pereiro I, Wasik A, Lobinski R (1997) Trace environmental speciation analysis for organometallic compounds by isothermal multicapillary gas chromatography—microwave induced plasma atomic emission spectrometry (MC GC MIP AES). Anal Chem 42:799–808

Szpunar J (2000a) Trace element speciation analysis of biomaterials by high performance liquid chromatography with inductively coupled plasma mass spectrometric detection (HPLC–ICP MS). Trends Anal Chem 19(2–3):127–137

Szpunar J (2000b) Bio-inorganic speciation analysis by hyphenated techniques. Analyst 125:963–988

Templeton DM, Ariese F, Cornelis R, Danielsson L-G, Muntau H, van Leeuwen HP, Lobinski R (2000) Guidelines for terms related to chemical speciation and fractionation of elements. Definitions, structural aspects and methodological approaches. Pure Appl Chem 72(8):1453–1470

Th.Göen, DFG-Lose Blätter Sammlung Validierte Methoden zum Huambiomonitoring: Arsen gesamt im Urin (2012), Wiley-VCH

Thiers RE, Vallee BL (1957) Distribution of metals in subcellular fractions of rat liver. J Biol Chem 226:911–920

Tittes W, Jakubowski H, Stüwer D (1994) Reduction of some spectral interferences in ICP–MS, finnigan MAT elemental mass spectrometry technical and application note 3. Finnigan MAT GmbH, Bremen

Urasa IT (1996) Developments of new methods of speciation analysis. In: Caroli S (ed) Element speciation in bioinorganic chemistry. Wiley, New York, pp 121–154

Ure AM, Quevauviller P, Muntau H, Griepink B (1993) Speciation of heavy metals in soils and sediments: an account of the improvement and harmonization of extraction techniques undertaken under the auspices of the BCR of the commission of the European communities. Int J Environ Anal Chem 51:135–151

Wolf C, Rösick U, Brätter P (2002) Sampling and processing of biopsy samples for speciation studies of cytosolic metalloproteins. Anal Bioanal Chem 372:491–494

Wolf RE, Morrison JM, Goldhaber MB (2007) Simultaneous determination of Cr(III) and Cr(VI) using reversed-phase ion-pairing liquid chromatography with dynamic reaction cell inductively coupled plasma mass spectrometry. J Anal At Spectrom 22:1051–1060

Xu M, Yang LM, Wang QQ (2008) Quantification of selenium-tagged proteins in human plasma using species-unspecific isotope dilution ICP-DRC-qMS coupled on-line with anion exchange chromatography. J Anal At Spectrom 23:1545–1549

Zeien H, Bruemmer GW (1989) Chemische Extraktionen zur Bestimmung von Schwermetallbimdungsformen in Böden. Mitt Dtsch Bodenk Gesellsch 59:505–510

Part IV

Techniques and Tools

Robert B. Finkelman

University of Texas at Dallas, Richardson, TX, USA
e-mail: bobf@utdallas.edu

Introduction

Geoscientists and medical researchers bring to medical geology an arsenal of valuable techniques and tools that can be applied to health problems caused by geologic materials and processes. Although some of these tools may be common to both disciplines, practitioners of these disciplines commonly apply them in novel ways or with unique perspectives. In this section, some of these tools and techniques are examined.

The geographic information system (GIS) is one of the more powerful techniques that is used by both the geoscience and public health/biomedical fields. GIS can help researchers to visualize the relationships between health problems and the physical environment. The value of GIS is dependent on having appropriate data to input, store, manipulate, and analyze.

In Chap. 28, an introduction is provided to a wide range of GIS databases that can be used for human health studies. The authors describe the types and features of databases that currently exist (with the caveat that this is a dynamic field with new databases appearing almost daily). They then provide several case studies that illustrate how geospatial data and GIS have been used in medical geology research. The examples include using GIS to determine the types of soil that Lyme disease-bearing ticks prefer and demonstrating the relationship of dental fluorosis to domestic coal combustion in southern China.

One of the most important types of data used in GIS is the information transmitted from Earth-orbiting satellites. Chapter 29 reviews fundamental aspects of remote sensing and describes the types of information that are available from many of the current remote-sensing satellites. It also describes how these geospatial data are used to gain insights into the occurrence and distribution of vector-borne diseases.

Classical mineralogical techniques, such as optical microscopy, X-ray diffraction, and microbeam analytical tools, have been used to characterize the complex biominerals formed in the human body that affect human health and well-being in many ways. Chapter 30 focuses on the largest biomineral in the human body, the bioapatites, which constitute the bulk of bones and teeth. It shows how these tools and techniques can be used to better understand these materials that are vital to our health. It also shows how classical geoscience tools and techniques have been used to gain and improve understanding of osteoporosis, a bone disorder afflicting millions of elderly people throughout the world.

The views expressed by the author are his own and do not represent the views of the United States GeologicalSurvey or the United States.

Accurate, comprehensive determination of the composition of geologic materials with which we come in contact and of natural materials that gain entry to the human body through ingestion, inhalation, or physical contact is essential for the protection of public health. In Chaps. 31 and 32, we can see the similarities and differences in the approaches that geoscientists and medical researchers adapt in characterizing these materials. In Chap. 31, a detailed overview is provided of several of the more important analytical techniques used for inorganic geochemical characterization of solids and liquids as well as several techniques for the characterization of organic matter. The authors describe the principles of each method, how each is applied, and the strengths and weaknesses of each.

Chapter 32 describes how biomedical researchers treat human tissue and body fluid samples to determine the impact of exposure to inorganic materials. In addition to histochemical techniques, a range of analytical tools used in identifying minute inclusions in human tissues are described. Some of these tools are similar to those described in Chap. 31.

Water is one of the most important media for transmission of disease. Hydrogeologists have developed powerful computer-based tools for modeling groundwater flow and quality. Chapter 33 provides a review of computer models of groundwater flow, solute transport, and geochemical reaction processes and describes how they can be applied to medical geology issues.

From images of Earth taken from thousands of miles above our heads to images of submicroscopic inclusions within human tissues, from the most sophisticated analytical equipment and computer programs to methods that have been used for 100 years, scientists are using their full arsenal of tools and techniques to solve a wide range of medical geology problems.

GIS in Human Health Studies

28

Joseph E. Bunnell, Alexander W. Karlsen, Robert B. Finkelman, and Timothy M. Shields

Contents

28.1 Introduction to Databases and Geographic Information Systems

Databases used in the field of medical geology are generally comprised of geospatial and/or temporal elements. Although these are not requirements for all medical geology research projects, much of the discussion in this chapter will be focused on databases incorporated into geographic information systems (GIS). GIS are computer-based (or manual) methods that allow a user to input, store, retrieve, manipulate, analyze, and output spatial data (Aronoff 1989). There are four major systems of GIS: engineering mapping systems (computer-aided design/computer-assisted mapping; CAD/CAM), geographic base file systems, image processing systems, and generalized thematic mapping systems. Various software packages are available that perform one or more of these systems, and the relative ability to move data back and forth between them can be critical to the needs and success of a particular GIS. Relational databases are the most commonly used types of databases in GIS (Cromley and McLafferty 2002). Relational database management models are convenient for linking formerly disparate databases together in a GIS. The databases to be joined must share one common attribute, usually an identifier such as coded patient number, sample site, or latitude/longitude. Other database management structures, such as hierarchical and network systems, are not as well suited to health GIS applications, although they may be useful for extremely large databases.

The capability to quickly and easily link large medical or public health databases with equally large geospatial databases represents an important technological breakthrough. Due to advances in computational power and speed, studies may be conducted today that could not have been done in reasonable time frames even just a few years ago. By linking disparate databases in a visually accessible manner (i.e., with maps), researchers are able to recognize

The views expressed by the author are his own and do not represent the views of the United States Geological Survey or the United States.

J.E. Bunnell (✉) • A.W. Karlsen
United States Department of the Interior, United States Geological Survey, Reston, VA, USA
e-mail: jbunnell@usgs.gov

R.B. Finkelman
Department of Geosciences, University of Texas, Dallas, Richardson, TX, USA

T.M. Shields
Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD, USA

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_28,

relationships or discern patterns of disease that can lead to an understanding of causality that was previously not apparent. The value in mapping disease occurrence is appreciated when doing so can illuminate the underlying cause of an outbreak, which may then enable mitigation measures to be taken to prevent further spread of a disease.

The earliest example of using such a spatioanalytical approach to solving an epidemiological riddle is generally credited to a physician named John Snow and the Broad Street Pump of London in 1854 (Cameron and Jones 1983). Dr. Snow mapped a major outbreak of cholera, in a time before the germ theory was well accepted, and hypothesized that there was a causal association between the putative source of the contagion and locations of cases. He convinced city officials to remove the handle of the pump dispensing contaminated water—an intervention that promptly quelled the outbreak. Although modern tools are much more sophisticated than those at Dr. Snow's disposal, our goal remains the same in applying GIS to public health issues.

Currently, databases used in GIS applications are often developed by the user or are available on the Internet or from other sources. In the first edition of this book, databases were typically stored in electronic media formats such as a hard drive, floppy disk, and compact disc-read only memory (CD-ROM). These media have since been replaced by hard drives that are measured in terabytes, external drives, web servers and the cloud (See Sect. 28.3).

GIS databases contain fields in columns and records in rows. A field, or item, is an element of a database record in which one piece of information is stored and represented as a column in a geodatabase table or spreadsheet (Kennedy 2001). *Records* represent different entities with different values for the attributes represented by the fields (Kennedy 2001). *Attributes* are information about geographic features, and they are contained within GIS data layers, or themes (Fig. 28.1). For example, a climate data layer (the feature) may contain the attributes of temperature, rainfall, and relative humidity for a specific geographic point location or region. Attribute data and spatial data comprise the two critical types of data in a GIS. A more in-depth look at how to assemble databases into a project is presented in Other Sources, part C.

Along with the individual databases, *metadata* are also included. Metadata, also called data dictionaries, are simply data about data. Metadata contain information such as the time/place the database was created, field and record identifier information (attributes), data development process (lineage), and individual(s) to contact regarding the data. If the data are displayed in a geographic environment, the metadata must also include additional information such as map scale and projection. Guidelines for what should be included in metadata are provided by the National Spatial Data Infrastructure, which is maintained by the Federal Geographic Data Committee (FGDC) (see Other Sources, part A). An interagency U.S. government organization, the FGDC sets guidelines for all aspects of spatial data, and works with offshore/international partners to develop the global spatial data infrastructure.

28.2 Types of Databases and Their Features

There are enormous numbers of databases available in digital format, but the types of data likely to be used by medical geologists fall into two broad categories: Earth science/geospatial databases and biomedical/health databases (see Other Sources, part A). What makes the field of medical geology innovative and unique is that it, by definition, brings together in a coherent manner databases from these two general areas in specific applications. This approach leads to fresh perspectives enabling recognition of connections between environmental factors and human health outcomes that may have previously gone unnoticed. Medical geological research can identify mechanistic connections that in turn can lead to new practices or policies. This may result in novel solutions to public health problems, which ultimately benefit large numbers of people. Two case studies are presented in Sects. 28.4 and 28.5 that illustrate the utility of such an approach.

Spatial data are represented in two models, vector and raster. In the vector models attribute data are attached or linked to one of three features: point, line, or polygon. Simply defined, a point is an x, y coordinate such as a mountain peak or soil sample location. The point feature does not have any length or area. A line is defined by the connection of two or more vertices (x, y coordinate pairs). A polyline is made up of numerous lines that represent the same feature, such as a road or river. The line or polyline feature has a length associated that is considered too narrow at the given scale to have an area. The polygon feature is defined by a series of lines that start and end at the same place, such as a state or country. Perimeter and area can be calculated for these features. In summary a line is a set of connected points whereas a polygon is a set of connected lines that have the same beginning and end points.

Features in a vector-based GIS can be linked or joined with attribute data from one or more databases provided a common identifier exists. The National Climatic Data Center (NCDC) generates climate data from ground-based observations. They supply a table of weather stations along with the stations' identifiers and x, y coordinates. This table can be imported into a GIS and point features can easily be made which correspond with weather station locations. Once these features exist in a GIS, they can be linked to other NCDC tables that contain station identifiers and

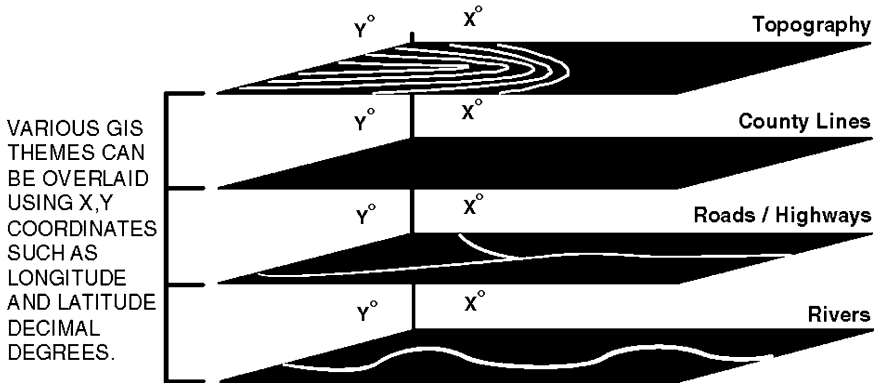


Fig. 28.1 GIS conceptual diagram. Note that different data layers (covers or themes) are overlaid such that queries about individual point locations reveal numerous attributes from a variety of source databases (Figure courtesy of Eric Morrissey, U.S. Geological Survey, Reston, VA)

hourly, daily, weekly, monthly, and/or annual climate data. Climatic conditions can be analyzed from the temporal frame of a single point in time or a complete historical compilation.

A key function of a vector-based GIS is topology. Topology is the spatial linkage between vector features. It stores the spatial relationship of the features with respect to each other. Topology enables the user to determine where a feature is in relation to other features, which parts of different features are shared, and how features are connected. Functions such as which sample points are located within a specific watershed or which counties does a river intersect can be performed (see Fig. 28.1). Topology also reduces the amount of information that must be stored. If two polygons are adjacent, that is they share a common line as a border, the common line needs to be stored only once. It will be saved as the right side of one polygon and as the left side of the other.

Raster systems present data in a regular grid of squares or cells. These cells are often called "pixels," which is an abbreviation of the words picture elements. Each pixel is defined by a row and column number and in GIS these can be converted to x, y coordinates. Pixels contain a single attribute value relating to the feature they represent. In an image that represents soil, a unique value would be given for each type of soil.

The differences in the manner in which vector and raster systems present and store geospatial data often lead people to contrast the two in order to determine which is better. Each model has its advantages and disadvantages. Vector systems, with points, lines, and polygons produce maps that are more like those drawn by hand and therefore are more aesthetically pleasing. Raster systems tend to represent continuous surfaces such as elevation or vegetation well. At the same time they demand large storage capacity because they require a value for every pixel in the image. The answer to the question of which system is better is dependent upon the application. Both systems are effective, but the nature of the application or task generally determines which one should be utilized.

28.3 Software, Computational Technology, and Technical Issues

Developments in software are facilitating the integration of these two forms of data. Vector-based systems now incorporate basic raster-based system functionality and vice versa. The crossover of functionality has helped the user immensely. No longer do entire layers of data need to be converted from one data structure to the other. The integration of the two forms has made data management faster and less expensive while providing better quality of data generated. While this integration has improved greatly over the last several years, there are limitations and certain functions need to be performed in a specific software processing environment (vector or raster) based upon the data model.

Improved computer power with regard to processing speed and storage capabilities has led to increasingly larger and/or higher resolution datasets being processed. Laptops now are more powerful than the high-end systems of just a few years ago. We have experienced the growth of web mapping servers to distribute geospatial data and now cloud computing. Cloud computing takes the storage and processing off site to a set of servers and data centers. One of the main advantages of the cloud is its scalable and elastic nature where it is able to scale up or down to meet the consumers' needs.

Getting data to be used in GIS can sometimes be challenging. Manual efforts such as digitizing (using a flat digitizing table to draw or copy a map into an electronic format), geocoding (using an electronic basemap to match an address to), or scanning (using a scanner to convert a paper map to electronic format) are time intensive.

Using existing digital sources is often preferable, but these sources can also have their share of issues pertaining to quality. These are outlined below.

Scale specifies the level at which real-world features have been reduced to be represented. It is usually stated as a ratio or

fraction such as 1:1,000 or 1/1,000, where 1 unit on the map represents 1,000 units in the area represented. An often confusing term for maps is small- or large-scale. Small-scale maps actually represent large areas but the ratio or fraction is a small number. Conversely, large-scale maps represent small areas. The map of a neighborhood would be considered large-scale whereas a map of Asia would be small-scale.

Resolution refers to the amount of detail in the features of the map. The scale of the map determines its resolution. It is the level at which features can be distinguished. On a small-scale map local features, such as ponds and small lakes, will not be represented. The terms fine (high) resolution and large-scale are synonymous; they contrast with coarse (low) resolution, and small-scale.

In GIS, issues of scale and resolution can determine which functions are appropriate as well as the level at which results can be stated (see Fig. 28.2). Ecological fallacy occurs when statements or predictions are made at one level based upon observations made at another. A soil survey conducted in 1 of the 24 counties of Maryland would be insufficient to predict soil characteristics throughout the state. Likewise, a small-scale digital elevation model would be inappropriate to use as part of a local water drainage study.

Accuracy is an important concept when working with GIS. Spatial accuracy, or how well the mapped features are located, must be stated and understood. If the map you are using states a spatial accuracy of ±1,000 ft, this will not be accurate enough to select a point to dig a well. Likewise, temporal accuracy must be stated. Population data from 1980 will not be accurate enough for current demographic studies. These are just a couple of ways error can be introduced in GIS.

Projections were developed to represent the curved surface of the Earth on a flat map. Projections are utilized to preserve local angles and shapes or relative size of areas. Every projection distorts the map in some manner and should be carefully chosen with respect to the study. Most GIS can convert data from one projection to another, which enables data layers from different sources to be compiled.

Metadata are information about data. All data sets utilized in GIS should have metadata accompanying them. This information should include the origin and characteristics of the data set, the purpose of the data set, and any problems the data set may have. This information is critical for the proper utilization of GIS data. The increase in the use and sharing of geospatial data has necessitated quality metadata. In 2007, the Federal Geographic Data Committee (FGDC) developed and released the EPA (US Environmental Protection Agency) Metadata Editor (EME). This free software simplifies the creation of metadata with the use of defaults and drop-down menus while meeting FGDC standards.

Remote sensing technology provides satellite imagery and high-resolution, high-altitude aerial photography. Such image data are also becoming more common in the context of GIS. NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) provides free multispectral imagery of the globe every 1–2 days at 250, 500, and 1,000 m resolution. The Landsat program provides free multispectral imagery of the earth's surface every 16 days with a resolution of 15–60 m. There are numerous other low to medium

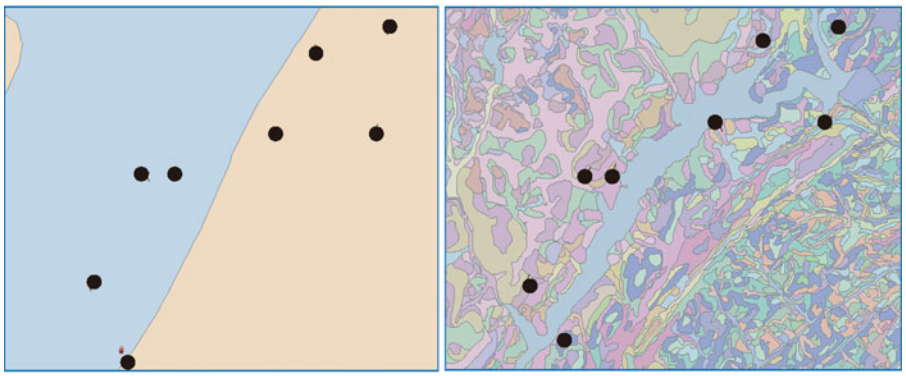


Fig. 28.2 Example of the effect of using databases of different scale. *Both panels* represent soil types displayed from digital databases for the exact same field sites (locations of individual transects are indicated with *black circles*, and latitudes and longitudes have been determined with a hand-held global positioning system [GPS] device). *Left panel* shows level of resolution available with STATSGO data (from the U.S. Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE), 1:250,000. *Right panel* reveals much greater detail available with SSURGO data (from the U. S. Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE), 1:24,000 (Data from Bunnell et al. 2003)

resolution imagery products available as well as a growing number of high resolution imagery. If purchasing or processing this imagery is impractical for consumers, GoogleEarth provides free visual access to an incredible amount of this imagery.

Raster data can exist either simply or with multiple values, for example, representing spectral bands. Geodatabase features with the same type of geometry make up simple or topological feature classes (Zeiler 1999). Object classes retain descriptive information related to geographic features, but they are not elements found on a map (Zeiler 1999). Depths of wells could make up an object class, for example, in a medical geology GIS examining proximity of drinking water wells to sources of arsenic in Bangladesh (see also Chaps. 12 and 29, this volume). Due to advances in computer technology, large raster data sets can now be manipulated and spatial data rigorously analyzed statistically with relative ease, thus making incorporation into epidemiological frameworks feasible (Robinson 2000).

New sources of geospatial data have developed as well. Historically these data would reside on a hard drive on a desktop computer to be utilized by trained GIS personnel. This platform has evolved into the laptop and PDA's (personal digital assistant) and smartphones with internet linkups. These offer additional capabilities in processing or viewing data. In addition these devices often can serve as a global positioning system (GPS) which determines coordinates which serve as input for features in a GIS. This has led to the growth of a multifaceted geospatial community with members of varying degrees of expertise ranging from novice to expert.

Medical geologists can now move beyond simply noting spatial coincidences of environmental features and disease patterns. By taking advantage of increasing computational speed and capabilities, sophisticated spatial statistics can be used in conjunction with GIS to reduce bias and correct for such potentially confounding effects as non-constant variance and autocorrelation (Haining 1998). Moreover, spatial statistical models can rigorously test for clustering versus random distributions, and can incorporate a fourth, temporal dimension to better assess correlation and offer clues into disease etiology (Kulldorff 1998).

28.4 Case Study 1: Lyme Disease

The first is a study designed to identify environmental determinants of tick abundance in the Mid-Atlantic region of the United States (Bunnell et al. 2003). Lyme disease is the most commonly reported vector-borne disease in the United States and it is still rapidly growing with over 10,000 new cases annually (Centers for Disease Control and Prevention 2001). In this part of the country, the black-legged, or deer, tick (*Ixodes scapularis*) transmits the microbial agents, that cause Lyme disease, ehrlichiosis and babesiosis, and other human and veterinary ailments. Tick abundance is likely a more reliable measure of the effect of landscape features on Lyme disease risk than human case data for several reasons, most notably because the location of Lyme disease cases will be, at best, the patients' home address, while many if not most of the cases are actually acquired elsewhere. Furthermore, by using tick distribution patterns, one avoids potentially misleading interpretations resulting from over- or underreporting due to the challenges in accurately diagnosing Lyme disease in humans. By better understanding the effects of environmental parameters on tick distribution, public health intervention strategies will likely be improved. Note that the word "vector" has a specialized meaning in the context of GIS (a mathematical definition, as above); in the biomedical community, a vector is an insect or other arthropod that actively transmits a pathogen from an infected reservoir host animal to another individual. Chapter 29 offers a thorough discussion of GIS technology applied to vector-borne diseases.

Environmental factors (covariates) that were previously known or suspected to correlate to tick abundance patterns included elevation, land cover, forest distribution, watersheds, and soil type (Glass et al. 1995; Ostfeld et al. 1996; Kitron and Kazmierczak 1997; Jensen et al. 2000). Digital elevation model (DEM) databases were obtained in a raster format (U.S. Geological Survey, Reston, VA). Land cover attributes were obtained from the remotely sensed multiresolution landscape characterization (MRLC) Landsat thematic mapper (TM) source data, managed by the Earth Resources Observation Systems (EROS), an entity within the U.S. Geological Survey (USGS). Initially, soils data were obtained from the state soil geographic database (STATSGO) maintained by the Natural Resources Conservation Service (U.S. Department of Agriculture, National Soil Survey Center, Lincoln, NE).

From a five-state region, 320 field sites were randomly selected and ticks were collected along transects at each site. Latitude and longitude were recorded at the beginning and end of each transect. At first, locations of the transects were noted on a paper map, 7.5-min topographical quadrangles were digitized by hand, and the sites were matched up. Digitizing a paper map requires a digitizing table and specialized hardware and computer software. The digitizing process was not required when latitudes and longitudes were recorded directly to a global positioning system (GPS) device, and then entered into a desktop personal computer. Newer GPS devices can further streamline the process by coupling directly with a laptop or desktop computer. This

obviates the need for data entry by hand. Each field site, or transect, became a unique identifier, and the latitude and longitude were the link to the environmental data downloaded. DEM data were used as is so that elevation in meters was obtained at each transect. Land cover data were of limited usefulness, as they only indicated whether a given field site was located in forest, low-intensity residential, open water, etc. Because tick population densities are positively associated with forest edges (ecotones), and because ticks are not found in open water, the GIS was used to calculate the distance from the midpoint of each transect to the nearest specific type of forest and body of water. The databases were thus manipulated from their native state, which resulted in new databases of particular utility to this specific research project.

Newly developed spatial statistics that incorporated spatial autocorrelation were applied to the data, and multiple regression analysis was performed (Das et al. 2002). These techniques revealed significant associations between tick abundance and certain environmental covariates that included soil type. This latter finding was particularly intriguing, and since the inception of the project, some soil data at much higher resolution were made available digitally. The soil survey geographic database (SSURGO) was released county by county as it became available (U.S. Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE). Only 75 of the 320 field sites happened to be located in counties with SSURGO data available at the time of analysis, and a second analysis was conducted on that subset of field sites. Because of the much finer resolution (1:24,000 vs. 1:250,000 for SSURGO and STATSGO, respectively), interpretations were made with greater precision (Fig. 28.2).

The enhanced resolution available with SSURGO data, combined with the more extensive set of attributes of this database, revealed some surprising results. For example, with STATSGO data, well-drained soils were found to be positively associated with tick abundance, in keeping with previous reports in the literature. However, upon analysis using SSURGO data, it was found that poorly drained soils, too, could be positively associated with tick abundance. This seeming contradiction was apparently resolved by considering precipitation factors and water-holding capacity of the soil. This example demonstrates the power of a GIS approach to examining environmental influences on factors controlling human disease risk. Observation of previously obscured patterns has enabled the generation of hypotheses now being tested to explain factors responsible for spatioanalytical trends in biological terms. Important advances in our understanding of basic Lyme disease ecology are likely to follow from this application of newly acquired and improved computational technology.

28.5 Case Study 2: Fluorosis in China

Elucidating the causes of fluorosis in the People's Republic of China offers another example of how GIS can be used to address the relationship between human health problems and geologic materials. Fluorosis, an abnormal condition of bones and teeth caused by exposure to excessive amounts of fluorine, affects millions of people throughout China. There are three principal pathways of exposure: drinking high-fluorine water, drinking tea made from tea leaves rich in fluorine, and exposure to fumes from residential combustion of high-fluorine coal or briquettes made with fluorine-rich clays as a binder (Zhang and Cao 1996; Ando et al. 1998) (see also Chap. 13).

Until recently, only general information existed on the epidemiology of fluorosis in China. For example, it was known that Kazakhs in the Xinjiang Autonomous Region in northwestern China were exposed to high levels of fluorine due to their preference for "brick tea" made from tea leaves rich in fluorine (Ben et al. 2000).

To determine where fluorosis was likely to be caused by exposure to fluorine-rich coal or coal briquettes, two GIS layers were required: the distribution of fluorosis and the distribution of coal deposits in China. No digital versions of either layer could be located. A map of the distribution and prevalence rates of dental fluorosis by county in China was located (Jianan 1989) and a map of the coal deposits of China was obtained (Ruiling et al. 1996). Both paper maps were digitized into electronic format at a computer workstation. Once in electronic form, the individual features of the maps (i.e., areas with the same prevalence rate of dental fluorosis) were assigned unique identifiers (attributes, in this case different colors). Parameters that control the way the map is displayed (projection) were adjusted so that the digital maps, representing the same geographic areas, would perfectly overlap each other (Karlsen et al. 2001).

In a GIS environment, the digital map of prevalence rates for dental fluorosis was overlain with the digital map showing coal distribution (Fig. 28.3). The combined maps confirm the association of fluorosis and coal in Guizhou Province, where more than ten million people are known to suffer from fluorosis (Zheng and Huang 1989). Figure 28.3 indicates that the high incidence of fluorosis in north central China (Shaanxi Province, which is the largest coal-producing province in China) may not be related to coal use.

In Guizhou Province the fluorosis is caused primarily by combining moderately high fluorine coals (average of about 200 ppm) with clays having very high fluorine contents (average of about 800 ppm) to form briquettes (Belkin et al. 1999). The high-fluorine clays are the residual products from intensive leaching of the limestone substrate that

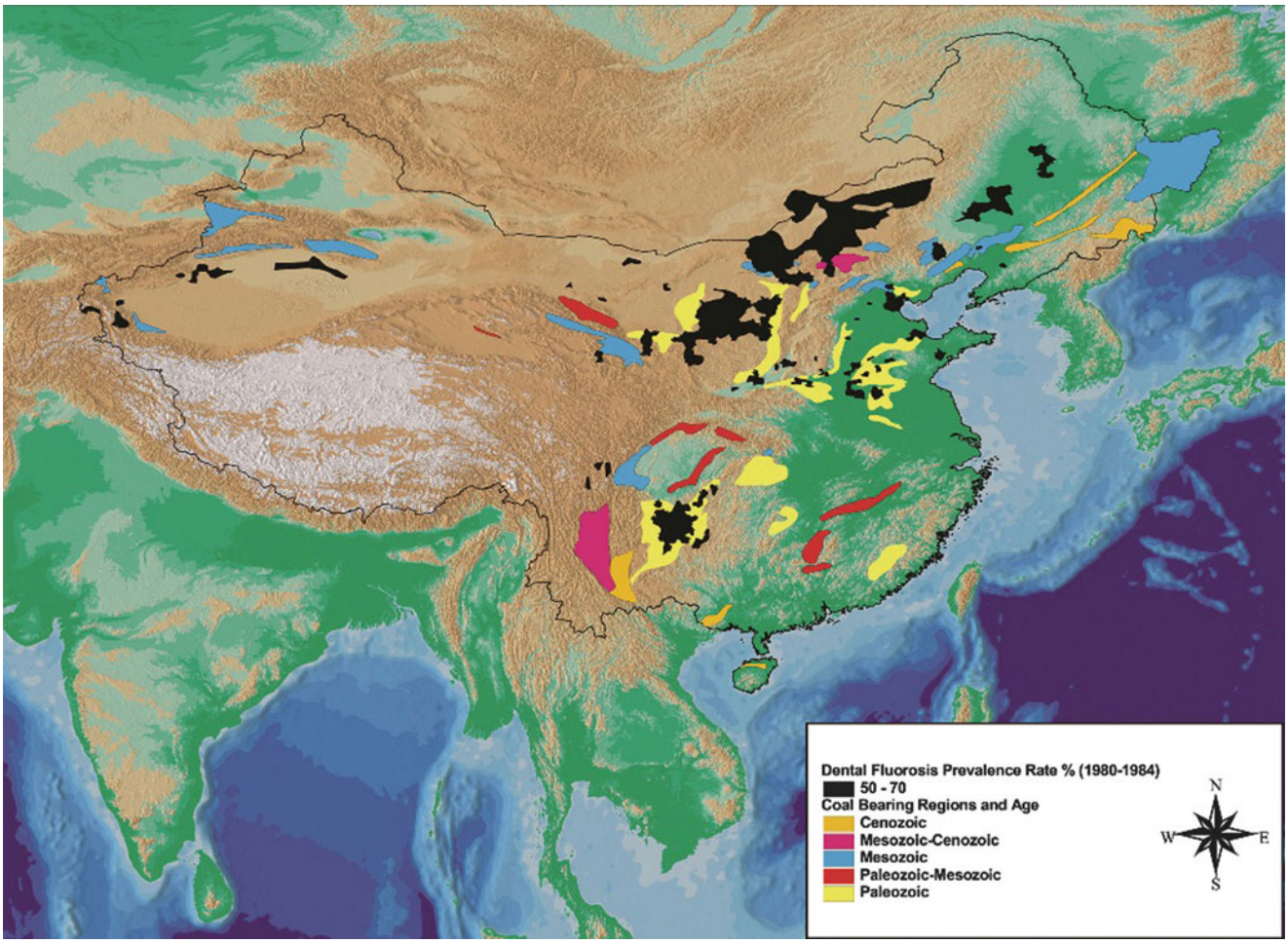


Fig. 28.3 Relationship of high prevalence rates of dental fluorosis and coal deposits in the People's Republic of China (From Karlsen et al. 2001)

formed the beautiful karst landforms for which the region is noted. In Shaanxi Province the substrate is primarily loess, a silicate-rich, wind-deposited sediment that is unlikely to have high-fluorine contents. Therefore, unless the coals in Shaanxi Province have exceptionally high fluorine contents, it is unlikely that the high incidence of dental fluorosis in that region is due to residential coal use.

Surprisingly, in the Xinjiang Autonomous Region the incidence of fluorosis also parallels the coal deposits. Perhaps this indicates that the distribution of fluorosis may be controlled by sedimentary rocks that favor the growth of the trees from which the fluorine-bearing tea leaves are obtained.

28.6 Other Case Studies

A number of other examples of GIS applications to medical geology problems may be found. These include the examples below.

African trypanosomiasis—Environmental factors that influence temporal and spatial distributions of African trypanosomiasis (sleeping sickness) were analyzed in a GIS that incorporated temporal Fourier analysis and discriminating analytical techniques such as Mahalanobis distance metric to aid in data interpretation (Rogers 2000). A relationship between the climate covariate included in the analysis was found to exist with vegetation patterns, which in turn influenced suitability of grazing for cattle, the main reservoir hosts for the tsetse fly (*Glossina* spp.) vector of the trypanosome parasites. Sequential statistical modeling of the tsetse fly populations and trypanosome disease transmission, when linked to biological modeling based on known differences due to the different species of tsetse, helped explain differences in the patterns of disease observed in different regions in Africa.

Hurricane Mitch—In 1998, one of the most powerful and deadly hurricanes in recorded history struck Central America. At least 6,500 people died and over 11,000 went missing in Honduras alone as a result of Mitch's fury. Many cases of human disease and death were caused by flooding,

even in areas not directly hit by the hurricane itself. Some of these floods, in turn, triggered lethal outbreaks of waterborne infectious diseases such as cholera, leptospirosis, Dengue fever, and malaria. GIS was used to predict high-risk areas for flood potential based on themes including river network configurations, elevation, and slope. This tool may have helped keep the casualty count low following this major disaster. Lessons learned from this experience may be helpful in using GIS to plan and execute preparedness and relief efforts before and after future catastrophic events (PAHO 2000).

Cadmium in The Netherlands—From 1892 until 1973, a zinc works in the Kempen area of The Netherlands discharged zinc and cadmium (Cd) into the environment in an uncontrolled fashion and seriously contaminated the soil with up to 8 ppm Cd (Stein et al. 1995). Cleanup efforts undertaken in the 1980s made use of a GIS and geostatistics to contour the cadmium distribution and improve sampling efficiency. Data from more than 1,700 soil samples were used as point data, and semi-variograms were created to compare stratified and ordinary kriging methods to interpolate the Cd concentrations. Neither mapping technique was uniformly superior; depending on the application (e.g., proximity to urban centers), one or the other map proved more useful. Had the GIS been used interactively, Stein et al. (1995) concluded that the number of soil samples necessary for testing could have been reduced approximately tenfold.

Malaria—A GIS analysis of malaria conducted in the Chiapas region, Mexico, and Peten, Guatemala, provides an example of how this analytical tool can be useful in active and iterative generation of testable research hypotheses. A priori hypotheses pertaining to environmental factors that influence malaria incidence by their impacts on the *Anopheles* spp. mosquito vectors invoked altitude, temperature, rainfall, land use, and vegetation type. In the course of developing the GIS, researchers observed that high-risk areas were often in close proximity to agricultural lands (SHA 2000). Further analysis led to the generation of a novel hypothesis that relates malaria risk to deforestation, a potential linkage that is presently being investigated by several groups.

Conclusions

In this chapter, the conceptual framework for GIS databases has been described, as have strategies and tools for conducting medical geology research projects. We have explained how recently developed GIS technology has truly revolutionized the study of human disease systems, which makes possible the simultaneous analysis of numerous interrelated factors that may exert unapparent and synergistic effects. Previously, such complex systems could only be addressed looking at one (or just a few) variable(s) at a time. With the case studies provided, the reader has seen examples of how to approach rigorous investigation into causes of human disease patterns that have strongly suspected environmental influences. Finally, we have supplied database resources that may provide a starting point for researchers wishing to conduct GIS studies of their own. In the time between the original publication of this chapter and this updated edition, computational technology has advanced dramatically, and geospatial data have become much more accessible.

See Also the Following Chapters. Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 13 (Fluoride in Natural Waters) • Chapter 29 (Investigating Vector-Borne and Zoonotic Diseases with Remote Sensing and GIS)

Further Reading

Ando M, Tadano M, Asanuma S, Matsushima S, Wanatabe T, Kondo T, Sakuai S, Ji R, Liang C, Cao S (1998) Health effects of indoor fluoride pollution from coal burning in China. Environ Health Perspect 106:239–244

Aronoff S (1989) Geographic information systems: a management perspective. WDL Publications, Ottawa

Belkin HE, Finkelman RB, Zheng BS (1999) Geochemistry of fluoride-rich coal related to endemic fluorosis in Guizhou Province, China, Pan-Asia Pacific conference on fluoride and arsenic research, Abstract 45, p 47

Ben K, Hua L, Hongchao H (2000) The current state of epidemic tea-induced fluorosis and its control countermeasures in Urumqi County, Xinjiang. In: Centeno JA, Collery P, Vernet G, Finkelman RB, Gibb H, Etienne J-C (eds) Metal ions in biology and medicine, vol 6. John Libby Eurotext, Paris, pp 303–305

Bunnell JE, Price SD, Lele SR, Das A, Shields TM, Glass GE (2003) Geographic information systems and spatial analysis of *Ixodes scapularis* (Acari: Ixodidae) in the middle Atlantic region of the U. S. A. J Med Entomol 40:570–576

Cameron D, Jones IG (1983) John snow, the broad street pump and modern epidemiology. Int J Epidemiol 12:393–396

Centers for Disease Control and Prevention (2001) Lyme disease–united states, 1999. MMWR 50(10):181–185

Cromley EK, McLafferty SL (2002) GIS and public health. The Guilford Press, New York, p 340

Das A, Lele SR, Glass GE, Shields TM, Patz JA (2002) Modeling a discrete spatial response using generalized linear mixed models: application to Lyme disease vectors. Int J Geog Inform Sci 16:151–166

Glass GE, Schwartz BS, Morgan JM III, Johnson DT, Noy PM, Israel E (1995) Environmental risk factors for Lyme disease identified with geographic information systems. Am J Public Health 85:944–948

Haining R (1998) Spatial statistics and the analysis of health data. In: Gatrell AC, Löytönen M (eds) GIS and health, GIS data VI. Taylor & Francis, London, pp 29–47

Hock R (2001) The extreme searcher's guide to web search engines, 2nd edn. CyberAge Books/Information Today, Inc., Medford, p 241

Jensen PM, Hansen H, Frandsen F (2000) Spatial risk assessment for Lyme borreliosis in Denmark. Scand J Infect Dis 32:545–550

Jianan T (ed) (1989) The atlas of endemic diseases and their environments in the People's Republic of China. Science Press, Beijing

Karlsen AW, Schultz AC, Warwick PD, Podwysocki SM, Lovern VS (2001) Coal geology, land use, and human health in the People's Republic of China, U. S. Geological Survey open file report 01–318 (CD-ROM)
Kennedy H (ed) (2001) Dictionary of GIS terminology. ESRI Press, Redlands
Kitron U, Kazmierczak JJ (1997) Spatial analysis of the distribution of Lyme disease in Wisconsin. Am J Epidemiol 145:558–566
Kulldorff M (1998) Statistical methods for spatial epidemiology: tests for randomness. In: Gatrell AC, Löytönen M (eds) GIS and health, GIS data VI. Taylor & Francis, London, pp 49–62
Ostfeld RS, Hazler KR, Cepeda OM (1996) Temporal and spatial dynamics of *Ixodes scapularis* (Acari: Ixodidae) in a rural landscape. J Med Entomol 33:90–95
PAHO (Panamerican Health Organization) (2000) Geographic information systems in health, special program for health analysis. PAHO, Washington, DC
Robinson TP (2000) Spatial statistics and geographical information systems in epidemiology and public health. In: Hay SI, Randolph SE, Rogers DJ (eds) Remote sensing and geographical information systems in epidemiology, advances in parasitology 47. Academic, San Diego, pp 81–128
Rogers DJ (2000) Satellites, space, time and the African trypanosomiases. In: Hay SI, Randolph SE, Rogers DJ (eds) Remote sensing and geographical information systems in epidemiology, advances in parasitology 47. Academic, San Diego, pp 129–171
Ruiling L, Tianyu H, Jianping W (Compilers) (1996) Coalfield prediction map of China. Surveying and mapping institute of Jilin Province, Publishing house of surveying and mapping, 9 map sheets, scale 1:2,500,000
SHA (Special Program for Health Analysis) (2000) Incidence of malaria and land use in Chiapas, Mexico and Peten, Guatemala, PAHO, Scientific paper No. 104
Sherman C, Price G (2001) The invisible web. CyberAge Books/Information Today, Inc., Medford
Stein A, Staritsky I, Bouma J, van Groenigen JW (1995) Interactive GIS for environmental risk assessment. Int J Geogr Inf Syst 9 (5):509–525
Zeiler M (1999) Modeling our world. ESRI Press, Redlands
Zhang Y, Cao SR (1996) Coal burning induced endemic fluorosis in China. Fluoride 29(4):207–211
Zheng B, Huang R (1989) Human fluorosis and environmental geochemistry in Southwest China. In: Developments in geoscience, contributions to 28th international geologic congress, Science Press, Washington, DC/Beijing, pp 171–176

Suggested Reading

Bernhardsen T (1999) Geographic information systems: an introduction. Wiley, New York
Beyer KMM, Comstock S, Seagren R (2010) Disease maps as context for community mapping: a methodological approach for linking confidential health information with local geographical knowledge for community health research. J Community Health 35(6):635–644
Briggs DJ, Elliott P (1995) The use of geographical information systems in studies on environment and health. World Health Stat Q 48:85–94
Burrough PA, McDonnell R (1998) Principles of geographic information systems. Oxford University Press, Oxford
Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, Oxford
Clarke KC (1998) Getting started with geographic information systems, 3rd edn. Prentice Hall, Upper Saddle River
DeMers MN (2000) Fundamentals of geographic information systems, 2nd edn. Wiley, New York
de Savigny D, Wijeyaratne P (eds) (1995) GIS for health and environment. International Development Research Centre, Ottawa
Glass GE (2000) Spatial aspects of epidemiology: the interface with medical geography. Epidemiol Rev 22(1):136–139
Goovaerts P (2008) Spatial uncertainty in medical geography: a geostatistical perspective. In: Shekhar S, Xiong H (eds) Encyclopedia of GIS. Springer, Berlin, pp 1106–1112
Green K (1992) Spatial imagery and GIS: integrated data for natural resource management. J For 90:32–36
Henshaw S, Curriero F, Shields T, Glass G, Strickland P, Breysse P (2004) Geostatistics and GIS: tools for characterizing environmental contamination. J Med Syst 28(4):335–348
Lang L (2000) GIS for health organizations, ESRI Press, Redlands, p 100 plus CD-ROM
Li S, Xiao T, Zheng B (2012) Medical geology of arsenic, selenium and thallium in China. Sci Total Environ 421–422:31–40
Longley PA, Goodchild MF, Maguire DJ, Rhind DW (eds) (1999) Geographical information systems, 2nd edn. Wiley, New York, p 1101, 2 volumes
Meade MS, Earickson RJ (2000) Medical geography, 2nd edn. The Guilford Press, New York
Melnick AL (2002) Introduction to geographic information systems in public health. Aspen Publishers, Gaithersburg
Moore GS (2002) Living with the earth: concepts in environmental health science, 2nd edn. Lewis Publishers, Boca Raton
Reisen WK (2010) Landscape epidemiology of vector-borne diseases. Annu Rev Entomol 55:461–483
Tabasi S, Abedi A (2012) A medical geology study of an arsenic-contaminated area in Kouhsorkh, NE Iran. Environ Geochem Health 34(2):171–179
Vine MF, Degnan D, Hanchette C (1997) Geographic information systems: their use in environmental epidemiological research. Environ Health Perspect 105:598–605

Other Sources

A. Sources of Earth Science/Geospatial Information

FGDC (http://www.fgdc.gov/)

The FGDC coordinates the sharing of geographic data, maps, and online services, that searches metadata held within the NSDI Clearinghouse Network. The National Geospatial Program (http://www.usgs.gov/ngpo/) organizes data through various portals which are listed here:

http://nationalmap.gov – base data, map products and geospatial web services

http://geo.data.gov – geopatial information discovery and access, communities, and partnership marketplace

http://nationalatlas.gov – maps of America for everyone's use

http://geonames.usgs.gov/domestic – geographic names authority and gazetteer for the nation

http://cegis.usgs.gov – leadership for geospatial information research

Seamless Data Warehouse (http://seamless.usgs.gov/)

The Seamless Data Warehouse is the ideal location to explore and retrieve data. U.S. Geological Survey (USGS)

and the Earth Resources Observation and Science (EROS) are committed to providing access to geospatial data through The National Map. An approach is to provide free downloads of national base layers, as well as other geospatial data layers. These layers are divided into framework categories:

Places, Structures, Transportation, Boundaries, Hydrography, Orthoimagery, Land Cover, and Elevation

Along with providing access to the data, this site contains:

Tutorial to help with downloads

Information about the downloadable products

Frequently Asked Questions (FAQs)

Links to product homepages and information pages

The Seamless Data Warehouse is always growing with new data, tools, features, and much more.

B. Libraries (for Further Research)

U.S. Geological Survey Library
950 National Center 12201 Sunrise Valley Drive
Reston, VA 20192
e-mail: library@usgs.gov

U.S. Library of Congress
101 Independence Avenue,
S. E. Washington, DC 20540
e-mail: lcweb@loc.gov

U.S. National Library of Medicine
National Institutes of Health
8600 Rockville Pike
Bethesda, MD 20894
e-mail: NIHInfo@OD.NIH.GOV

C. Where Does One Start a Search for Relevant Databases?

The FGDC Web site http://www.fgdc.gov provides access to over 38 simultaneously searchable data clearinghouses in the United States and internationally. These include databases related to Earth sciences, geography, landform information, ecosystem health, biological resources, and satellite imagery (see Other Sources, part A). The focus of this chapter has been on electronic sources, but don't forget to check your library's reference section, trade journals, or other specialized periodicals and books. To find databases on the Internet, you might use a search engine. Be aware that different search engines work in different ways, and that what may be overlooked by one search engine might be found by another one. There are also metasearch engines that use several search engines simultaneously (Hock 2001). Of course, another very efficient way to find out what databases are used by experts in a given field is to simply ask them. Contact information for university professors is often listed on their institution's Web site, which can be found on any search engine. Even if the principal investigator is hard to reach in person, his or her postdoctoral fellows, graduate students, and technicians may be willing to help.

An efficient strategy when starting out on medical geology research projects is to seek out relevant database clearinghouses. Using clearinghouses also offers some protection from rapidly changing unique or uniform resource locators (URLs), which are often referred to as Web site addresses. One example is the U.S. Geological Survey (USGS), which is a major clearinghouse for Earth science data. The URL for some particular databases contained therein may change, but the URL for a clearinghouse such as USGS generally remains stable over time. The primary clearinghouse organization will maintain proper internal links and keep access to all of their individual databases current.

The geographic and temporal range of the data needed must be ascertained at the outset of any medical geology GIS. It is better to err on the side of obtaining more information then deleting unnecessary elements, because it can be difficult to add data later if it is decided to examine additional parameters. But the initial cost of the data and the cost in resources to store data must also be taken into account. Because different databases will likely contain data archived in a variety of formats, it is advisable to store the initial downloaded data as is and make copies of it before any subsequent manipulation. The text-only ASCII file format is a "common denominator" useful for merging data from different sources into a single data set.

The use of Internet-derived databases can be made frustrating and difficult by two realities of this medium. One reality is that URLs change quickly, and so the Web site address that worked in the past may not take you to the same page today. This potential pitfall can be avoided by using the "gatekeeper" URL to a database clearinghouse as mentioned above, rather than by using direct URLs to individual databases. For example, one is advised to use a main clearinghouse Web site rather than a more specific URL for some individual database, such as for water table depths in India. The other major problem is the so-called Invisible Web (Sherman and Price 2001). There are a great many databases accessible via the Internet with no easy way to find them or to find out about them. Many databases can only be accessed after registering and entering a password. Search engines will miss these and other relevant sites, and they will often come up with totally irrelevant sites. Investigators must be mindful, too, of the reliability of database sources accessible via the Internet. *If associated metadata are not available, that database should not be used.*

Once a database of interest has been identified and the legitimacy of the organization that maintains it is verified, you are ready to download. Make sure you have the minimum requirements and sufficient memory space on your

computer before proceeding. As always when downloading any software or data to a personal computer, remember to have some tool in place for screening computer viruses. It is critical to ensure that once downloaded, the data were not corrupted in the process. You must examine the source data carefully and confirm that they match the data in the form that has been downloaded. Problems can arise, for example, if the source data are tab delimited and your default download is space delimited. As soon as you have downloaded the source data, you should make a backup copy before doing anything with the data. It is generally convenient to keep such files on a compact disc (CD). Now you are ready to open up your data with your spreadsheet software package and import it to your GIS application or to a statistical analysis package. In a GIS environment, you can easily query the data. That is, by clicking with a mouse on a location visibly displayed on a map, you can extract attributes of that point.

You will need to join data from different databases for use in a GIS project. Get to know the raw data well as you must always maintain quality assurance/quality control (QA/QC). It is easy to mix up or somehow corrupt data when manipulating it. For instance, if you sort the data for some reason, make sure you keep a copy of the original unsorted data. Also be careful not to sort only one field, but rather keep your unique identifiers tied to the data in the proper order. If your ultimate aim is to do some statistical analysis of the data, you should work closely with a statistician right from the start. The statistician will help you determine the appropriate data you need to answer the questions you are asking.

Investigating Vector-Borne and Zoonotic Diseases with Remote Sensing and GIS

29

Stephen C. Guptill and Chester G. Moore

Contents

29.1 Introduction

For centuries, people have been intuitively aware of the relationships between human health and the environment. Today, geographic information systems (GIS), remote sensing satellites, and other technologies are providing scientists with the tools and the data to make clear the geographic relationships between the habitats of disease agents, their vectors and vertebrate hosts, and the occurrence of disease in the human population. Although the utility of the foregoing tools as an aid to epidemiology was pointed out 30 years ago (Cline 1970), the medical community has been slow to put them to use.

In this discussion, we provide an introduction to remote sensing technology and then use vector-borne and zoonotic diseases to demonstrate some current and potential uses of remote sensing, GIS, and related technologies for the surveillance, prevention, and control of disease. Vector-borne is defined as those diseases that are transmitted from one vertebrate host to another by an invertebrate, usually an insect, a tick, or a snail. A zoonosis (pl. zoonoses) is a disease that normally exists in a non-human host, or reservoir. For example, passerine birds are the natural hosts of several viruses that can infect humans. Many vector-borne diseases are also zoonoses (for example, yellow fever, Lyme disease, and plague).

There are several reasons for choosing this group of diseases as our examples. First, these diseases account for a large portion of the annual global morbidity and mortality. It is estimated that 41% of the world's population (about 2.3 billion people) live in areas with malaria risk (Gratz 1999). Of approximately 300–500 million people who become infected each year, some 1.5–2.7 million die from this disease. Dengue, a virus related to yellow fever, attacks as many as 50–100 million people each year, and there are more than 250,000 cases of the more severe forms of dengue hemorrhagic fever and dengue shock syndrome. Rabies, a zoonosis, is responsible for at least 35,000–45,000 deaths

S.C. Guptill (✉)
United States Department of the Interior, United States Geological Survey, Reston, VA, USA
e-mail: sguptill@guptillgeoscience.com

C.G. Moore
Department of Microbiology, Immunology and Pathology, School of Veterinary Medicine and Biomedical Sciences, Colorado State University, Fort Collins, CO, USA

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_29,

each year. Second, these diseases—especially the zoonoses—are naturally occurring systems that are impacted by a wide variety of physical and biotic factors that may be susceptible to remote measurement (e.g., Ostfeld et al. 1996). Finally, there is considerable concern about the potential impact of global change on the dynamics and spread of these diseases (e.g., Shope 1992; Reeves et al. 1994; Martens et al. 1995; Jetten and Focks 1997).

29.2 Fundamentals of Remote Sensing

Remote sensing is a technology that involves the analysis and interpretation of images gathered through techniques that do not require direct contact with the object. Electromagnetic radiation sensors are used to record images of the environment. The sensors used in these devices can detect radiation from the ultraviolet through the visible and infrared spectra to microwave radar. In studying the Earth, the remote sensing devices are usually deployed in aircraft or Earth-orbiting satellites. For convenience, in the remainder of this paper, we will generally refer to remote sensing satellites, but the concepts and principles also apply to remote sensing instruments carried aboard aircraft.

Remote sensing satellites are designed to collect various types of information about the Earth's surface and atmosphere. The combination of spatial resolution of the sensor, the wavelengths detected by that sensor, and the frequency of data collection determine the types of applications for which the satellite will collect useful information. These design parameters must be weighed one against another, as they are somewhat mutually exclusive. Each of these parameters will be examined below along with descriptions of how they have been implemented in various operational systems.

29.2.1 Electromagnetic Spectrum

Electromagnetic radiation (EMR) extends over a wide range of energies and wavelengths (frequencies). A narrow range of EMR extending from 0.4 to 0.7 μm, the interval detected by the human eye, is known as the *visible region* (also referred to as *light*, but physicists often use that term to include radiation beyond the visible). White light contains a mix of all wavelengths in the visible region. The distribution of the continuum of all radiant energies can be plotted either as a function of wavelength or of frequency in a chart known as the electromagnetic spectrum.

Using spectroscopes and other radiation detection instruments, scientists have arbitrarily divided the electromagnetic spectrum into regions or intervals and applied descriptive names to them. At the very energetic (high frequency and short wavelength) end are gamma rays and x-rays, whose wavelengths are normally measured in angstroms (Å), which in the metric scale are in units of 10^{-8} cm. Radiation in the ultraviolet extends from about 300–4,000 Å. It is convenient to measure the mid-regions of the spectrum in one of two units: micrometers (μm), which are multiples of 10^{-6} m or nanometers (nm), based on 10^{-9} m. The visible region occupies the range between 0.4 and 0.7 μm, or its equivalents of 4,000–7,000 Å or 400–700 nm. The infrared region (IR), spanning between 0.7 and 100 μm, has four subintervals of special interest: (1) reflected IR (0.7–3.0 μm); (2) its film responsive subset, the photographic IR (0.7–0.9 μm); and (3) thermal bands at 3–5 μm and (4) 8–14 μm. Longer wavelength intervals are measured in units ranging from millimeters to centimeters to meters. The microwave region spreads across 0.1–100 cm, which includes the entire interval used by radar systems. These systems generate their own active radiation and direct it toward targets of interest. The lowest frequency–longest wavelength region beyond 100 cm is the radio bands, from VHF (very high frequency) to ELF (extremely low frequency). Within any region, a collection of continuous wavelengths can be partitioned into discrete intervals called *bands*.

Most remote sensing is conducted above the Earth either within or above the atmosphere. The gases in the atmosphere interact with solar irradiation and with radiation from the Earth's surface. The atmosphere itself is excited by EMR and thus becomes another source of released photons. Figure 29.1 (Short 2003) is a generalized diagram which shows relative atmospheric radiation transmission of different wavelengths.

Shaded zones mark minimal passage of incoming and/or outgoing radiation, whereas white areas denote atmospheric windows in which the radiation does not interact much with air molecules and hence, is not absorbed.

Most remote sensing instruments on air or space platforms operate in one or more of these windows by making their measurements with detectors tuned to specific frequencies (wavelengths) that pass through the atmosphere. However, some sensors, especially those on meteorological satellites, directly measure absorption phenomena, such as those associated with carbon dioxide (CO_2) and other gaseous molecules. Note that the atmosphere is nearly opaque to electromagnetic radiation in part of the mid-IR and all of the far-IR regions. In the microwave region, by contrast, most of this radiation moves through unimpeded, so radar waves reach the surface.

Remote sensing of the Earth traditionally has used reflected energy in the visible and infrared and emitted energy in the thermal infrared and microwave regions to gather radiation that can be analyzed numerically or used to generate images whose variations represent different intensities of photons associated with a range of wavelengths that are received at the sensor. This gathering of a (continuous or discontinuous) range(s) of wavelengths is the essence of what is usually termed multispectral remote sensing (Short 2003).

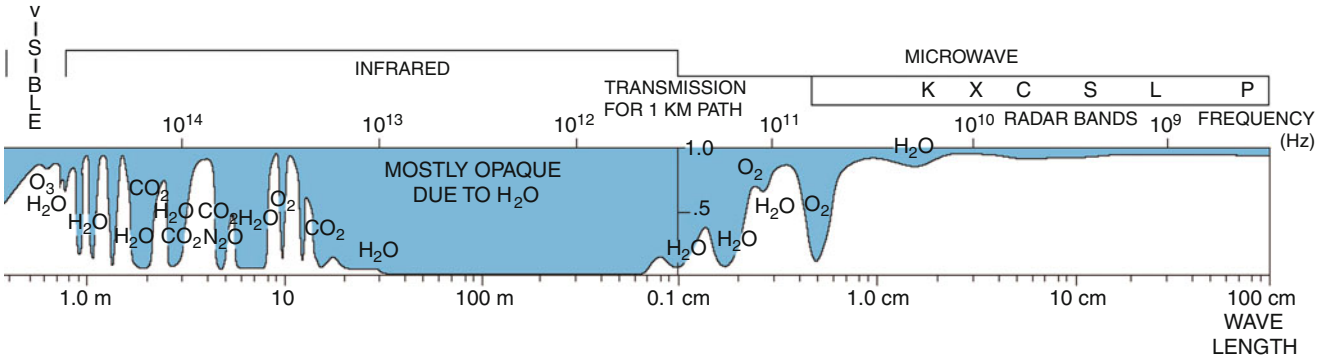


Fig. 29.1 Electromagnetic spectrum showing areas of transmission (*in white*) and areas of atmospheric absorption (*in blue*)

29.2.2 Spatial, Temporal, and Spectral Resolution

Spatial, temporal, and spectral resolution are three parameters that largely determine the characteristics of the data collected by a remote sensing instrument and thus to some degree the applications for which those data can be used.

The orbital characteristics of the satellite are the primary determinant of its temporal characteristics, that is, the time required for the sensor to re-image the same geographic location on the Earth. This can range from a daily revisit for the advanced very high resolution radiometer (AVHRR) sensor on the National Oceanic and Atmospheric Administration (NOAA) satellite to a 16-day revisit time for Landsat 7. Most satellites image the area directly below the sensor (nadir). However, some satellites are designed to point the sensor off-nadir, allowing the satellite to image the target area more frequently.

The imaging sensors are designed to collect energy from a certain portion or portions of the electromagnetic spectrum. For example a "panchromatic" sensor collects one set of data (i.e., a grayscale image) across a broad spectrum of visible and near infrared energy (450–900 nm). In contrast a hyperspectral sensor, like Hyperion, collects 250 bands of data, each 10 nm in width from 43–2,400 nm. Figure 29.2a–d shows the ways in which different features are highlighted or hidden depending on the band combinations used to make the image.

Spatial resolution refers to the smallest unit of area within which the sensor integrates EMR. More commonly it is referred to as "pixel size." This size is basically determined by the height of the orbit above the Earth and the magnification power of the optics on the sensor. As of the fall of 2003 the QuickBird satellite obtains the highest spatial resolution image with a 61-cm (approximately 2 ft) pixel size panchromatic (black and white) image collected in the 450- to 900-nm band. The detail available at this resolution can be seen in Fig. 29.3 which is an image of the Eiffel Tower taken on April 9, 2002. The coarsest image commonly used in our applications is from the AVHRR sensor, with a ground resolution of approximately 1 km.

Using digital image processing, satellite data of varying resolutions can be combined. In the example below, QuickBird 61-cm panchromatic imagery is combined with 2.4-m multispectral imagery (collected at the same time) to create a color image with 61-cm resolution. This image of Prague (Fig. 29.4) was taken on August 17, 2002, and shows the extent of flood damage to the city.

29.3 Remote Sensing Satellites

Remote sensing instruments fall naturally into three groups based on their principal applications: land observation, meteorology, or oceanography. However, many of the satellites provide useful information for more than one set of applications. Here, compiled by Nicholas Short (2003), are the principal remote sensing spacecraft flown by the United States and other nations (identified in parentheses) along with the launch date (if more than one in a series, this date refers to the first one put successfully into orbit):

1. *Land observation:* Government satellites—Landsat 1–7 (1973); Seasat (1978); HCMM (1978); RESURS (Russia) (1985); *IRS* 1A–1D (India) (1986); ERS 1–2 (1991); JERS 1–2 (Japan) (1992); Radarsat (Canada) (1995); ADEOS (Japan) (1996); Terra (1999)
   Commercial satellites—SPOT (France) (1986); Resurs-01 series (Russia) (1989; became commercial in the 1990s); Orbview-2 (U.S.) (1997); SPIN-2 (Russia) (1998); IKONOS (U.S.) (1999); QuickBird (U.S.) (2001); EROS A (ImageSat International; Israel) (2002)
2. *Meteorological observation*: TIROS 1–9 (1960); Nimbus 1–7 (1964); ESSA 1–9 (1966); ATS (g = geostationary) 1–3 (1966); DMSP series I (1966); the Russian Kosmos (1968) and Meteor series (1969); ITOS series (1970); SMS(g) (1975); GOES(g) series (1975); NOAA 1–5 (1976); DMSP series 2 (1976); GMS (Himawari) (g)

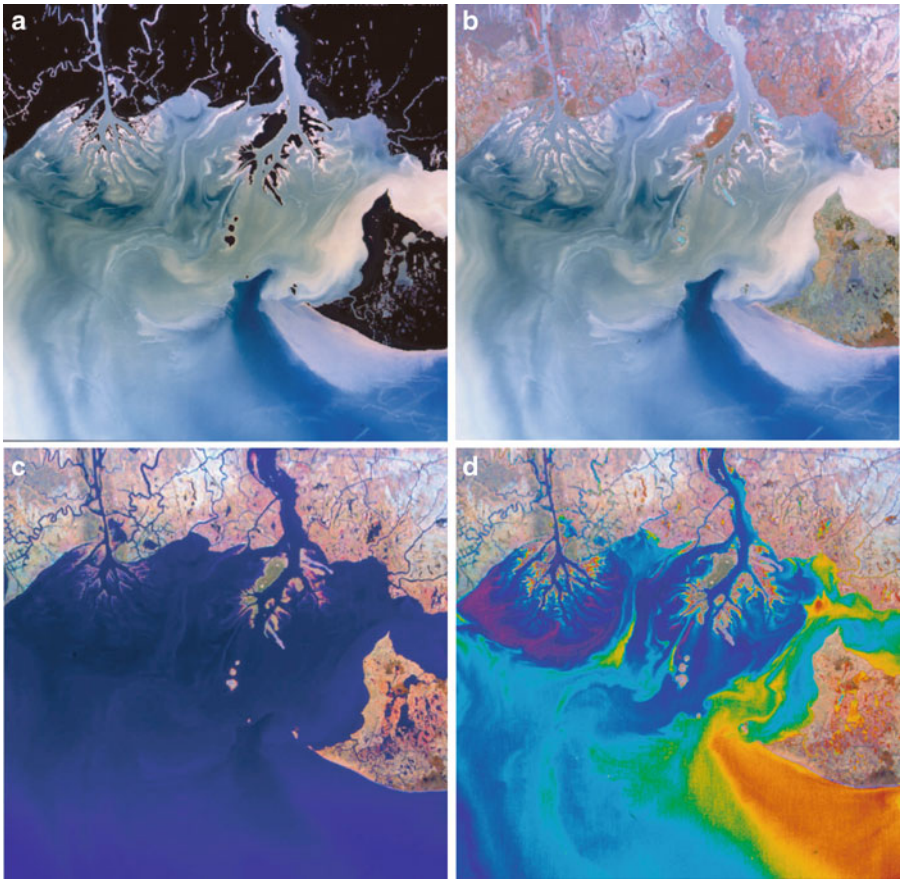


Fig. 29.2 Atchafalya Bay, Louisiana (Courtesy NASA). (**a**) *Bands* showing natural color, (**b**) *bands* showing color-infrared, (**c**) *bands* showing middle-infrared, and (**d**) *bands* showing thermal radiation

Fig. 29.3 Panchromatic image of the Eiffel Tower, Paris, France, with resolution of 61 cm (Courtesy DigitalGlobe)

Fig. 29.4 Panchromatic/multispectral merged image of Prague Czechoslovakia with effective resolution of 61 cm (Courtesy DigitalGlobe)

series (Japan) (1977); Meteosat(g) series (Europe) (1978); TIROS-N series (1978); Bhaskara(g) (India) (1979); NOAA (6–14) (1982); Insat (1983); ERBS (1984); MOS (Japan) (1987); UARS (1991); TRMM (U.S./Japan) (1997); Envisat (European Space Agency) (2002); Aqua (2002)

3. *Oceanographic observations*: Seasat (1978); Nimbus 7 (1978) included the *CZCS*, the Coastal Zone Color Scanner that measures chlorophyll concentration in seawater; Topex-Poseidon (1992); SeaWiFS (1997)

The Committee on Earth Observation Satellites (CEOS), supported by NASA, maintains a comprehensive database that shows the basic characteristics of current and future planned remote sensing satellite systems. A dynamic version of this table can be found at http://database.eohandbook.com/database/missiontable.aspx. As of January 2012, there were over 100 operational earth observation satellites.

So what applications are addressed by the data collected from these instruments? As part of NASA's Earth Observing System program, Michael King (2000) has created a table that shows a variety of physical measurements and characterizations that can be constructed from remote sensing data. The measurements/characteristics are shown on the left. The satellite/sensor that collects the raw information used in this process is on the right.

Atmosphere	
Cloud properties (amount, optical properties, height)	MODIS, GLAS, AMSR-E, MISR, AIRS, ASTER, SAGE III
Radiative energy fluxes (top of atmosphere, surface)	CERES, ACRIM III, MODIS, AMSR-E, GLAS, MISR, AIRS, ASTER, SAGE III
Precipitation	AMSR-E
Tropospheric chemistry (ozone, precursor gases)	TES, MOPITT, SAGE III, MLS, HIRDLS, LIS
Stratospheric chemistry (ozone, ClO, BrO, OH, trace gases)	MLS, HIRDLS, SAGE III, OMI, *TES*
Aerosol properties (stratospheric, tropospheric)	SAGE III, HIRDLS, MODIS, MISR, OMI, GLAS
Atmospheric temperature	AIRS/AMSU-A, MLS, HIRDLS, TES, MODIS
Atmospheric humidity	AIRS/AMSU-A/HSB, MLS, SAGE III, HIRDLS, Poseidon 2/JMR/DORIS, MODIS, TES
Lightning (events, area, flash structure)	LIS
Solar radiation	
Total solar irradiance	ACRIM III, TIM
Solar spectral irradiance	SIM, SOLSTICE
Land	
Land cover and land use change	ETM+, MODIS, ASTER, MISR
Vegetation dynamics	MODIS, MISR, ETM+, ASTER
Surface temperature	ASTER, MODIS, AIRS, AMSR-E, ETM+

(*continued*)

(continued)

Atmosphere	
Fire occurrence (extent, thermal anomalies)	MODIS, ASTER, ETM+
Volcanic effects (frequency of occurrence, thermal anomalies, impact)	MODIS, ASTER, ETM+, MISR
Surface wetness	AMSR-E
Ocean	
Surface temperature	MODIS, AIRS, AMSR-E
Phytoplankton and dissolved organic matter	MODIS
Surface wind fields	SeaWinds, AMSR-E, Poseidon 2/JMR/DORIS
Ocean surface topography (height, waves, sea level)	Poseidon 2/JMR/DORIS
Cryosphere	
Land ice (ice sheet topography, ice sheet volume change, glacier change)	GLAS, ASTER, ETM+
Sea ice (extent, concentration, motion, temperature)	AMSR-E, Poseidon 2/JMR/DORIS, MODIS, ETM+, ASTER
Snow cover (extent, water equivalent)	MODIS, AMSR-E, ASTER, ETM +

In the sections below, a few of the most widely available and useful systems for analyzing environmental influences on human health will be characterized.

29.3.1 Landsat

The Landsat program has been in operation since the early 1970s. Since then, many different satellites have been sent into orbit. Beginning with Landsat 1 in 1972 and most recently Landsat 7, these satellites have taken thousands of images of the Earth and documented the rapidly changing landscape of the planet.

Landsat 7 was launched on April 15, 1999, from the Western Test Range aboard a Delta II expendable launch vehicle. At launch, the satellite weighed approximately 4,800 lb (2,200 kg). The spacecraft is about 14 ft long (4.3 m) and 9 ft (2.8 m) in diameter. It consists of a spacecraft bus that is provided under a NASA contract with Lockheed Martin Missiles and Space in Valley Forge, Pennsylvania, and the Enhanced Thematic Mapper Plus (ETM+) instrument, pro-cured under a NASA contract with Raytheon (formerly Hughes) Santa Barbara Remote Sensing in California.

The ETM + instrument is an eight-band multispectral scanning radiometer capable of providing high-resolution imaging information of the Earth's surface. It detects spectrally filtered radiation at visible, near-infrared, short-wave, and thermal infrared frequency bands from the sun-lit Earth in a 115-mile (183 km) wide swath when orbiting at an altitude of 438 miles (705 km). Nominal ground sample distances or pixel sizes are 49 ft (15 m) in the panchromatic band; 98 ft (30 m) in the six visible, near- and short-wave infrared bands; and 197 ft (60 m) in the thermal infrared band. A Landsat WorldWide-Reference System has catalogued the world's landmass into 57,784 scenes, each 115 miles (183 km) wide by 106 miles (170 km) long. The ETM + produces approximately 3.8 Gb of data for each scene. The seven bands have the following characteristics:

Band no	Wavelength interval (μm)	Spectral response	Resolution (m)
1	0.45–0.52	Blue-green	30
2	0.52–0.60	Green	30
3	0.63–0.69	Red	30
4	0.76–0.90	Near-IR	30
5	1.55–1.75	Mid-IR	30
6	10.40–12.50	Thermal-IR	120
7	2.08–2.35	Mid-IR	30

The satellite orbits the Earth at an altitude of approximately 438 miles (705 km) with a sun-synchronous, 98-degree inclination and a descending equatorial crossing time of 10 a.m. The orbit will be adjusted upon reaching orbit so that its 16-day repeat cycle coincides with the Landsat Worldwide Reference System. This orbit will be maintained with periodic adjustments for the life of the mission. A three-axis attitude control subsystem will stabilize the satellite and keep the instrument pointed toward Earth to within 0.05degrees. A state-of-the-art solid-state recorder capable of storing 380 Giga bits of data (100 scenes) is used to store selected scenes from around the world for playback over a U.S. ground station. In addition to stored data, real-time data from ETM + can be transmitted to cooperating international ground stations and to the U.S. ground stations.

These spectral bands allow ETM + to detect subtle variations in surface characteristics. For example, on the border between Chile and the Catamarca province of Argentina lies a vast field of currently dormant volcanoes. Over time, these volcanoes have laid down a crust of magma roughly 2 miles (3.5 km) thick. It is tinged with a patina of various colors that can indicate both the age and mineral content of the original lava flows. This is shown in Fig. 29.5.

The U.S. Geological Survey operates the Landsat 7 satellite. For more information on Landsat 7 including how to order data, go to http://landsat7.usgs.gov/index.php.

29.3.2 AVHRR

AVHRR is a broad-band, four- or five-channel (depending on the model) scanner, sensing in the visible, near-infrared, and thermal infrared portions of the electromagnetic spectrum. This sensor is carried on NOAA's Polar Orbiting

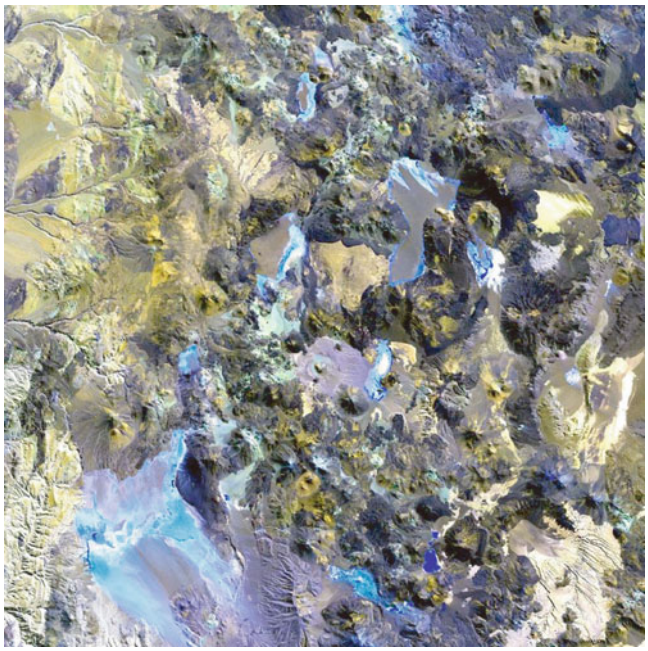


Fig. 29.5 Landsat ETM + image of Chile and Argentina (Courtesy USGS)

Environmental satellites (POES), beginning with TIROS-N in 1978 and most recently on the NOAA-15 (launched in 1998) and NOAA-16 (launched in 2000) satellites.

The AVHRR sensor provides for global (pole to pole) on-board collection of data from all spectral channels. Each pass of the satellite provides a 1,491-mile (2,399-km) wide swath. The satellite orbits the Earth 14 times each day from 517 miles (833 km) above its surface.

The average instantaneous field-of-view (IFOV) of 1.4 milliradians (mrad) yields a LAC/HRPT ground resolution of approximately 1.1 km at the satellite nadir from the nominal orbit altitude of 517 miles (833 km). The GAC data are derived from an on-board sample averaging of the full resolution AVHRR data. Four out of every five samples along the scan line are used to compute one average value and the data from only every third scan line are processed, yielding a 1.1 × 4 km resolution at nadir.

The current sensors cover five spectral bands as shown below:

Band	Wavelength (μm)	*IFOV* (mrad)
1	0.58–0.68	1.39
2	0.725–1.10	1.41
3	3.55–3.93	1.51
4	10.3–11.3	1.41
5	11.5–12.5	1.30

AVHRR data provide opportunities for studying and monitoring vegetation conditions in ecosystems including forests, tundra, and grasslands. Applications include agricultural assessment, land cover mapping, producing image maps of large areas such as countries or continents, and tracking regional and continental snow cover. AVHRR data are also used to retrieve various geophysical parameters such as sea surface temperatures and energy budget data.

Online requests for these data can be placed via the U.S. Geological Survey Global Land Information System (GLIS) interactive query system. The GLIS system contains metadata and online samples of Earth science data. With GLIS, you may review metadata, determine product availability, and place online requests for products. Additional data sets include the Alaska twice-monthly AVHRR and the U.S. Conterminous bi-weekly composites. These comprehensive time series data sets are calibrated, georegistered daily observations and twice-monthly maximum NDVI composites for each annual growing season. Global experimental bi-weekly normalized difference data, computed from Global Vegetation Index (GVI) data, are analyzed to monitor global vegetation and are a potential tool in global climatic studies.

Figure 29.6 shows an AVHRR image of the Mississippi River basin soon after the summer floods of 1993. This figure shows the extent of flooding and demonstrates the value of daily observations.

29.3.3 EOS Terra Spacecraft

On December 18, 1999, NASA launched the Earth Observing System (EOS) "flagship"—EOS Terra—to begin collecting a new 18-year global data set on which to base future scientific investigations about our complex home planet.

Physically, the EOS Terra spacecraft is roughly the size of a small school bus. It carries a payload of five state-of-the-art sensors that will study the interactions among the Earth's atmosphere, lands, oceans, and radiant energy (heat and light). Each sensor has unique design features that will enable EOS scientists to meet a wide range of science objectives.

EOS Terra orbits the Earth from pole to pole descending across the equator in the morning when cloud cover is minimal and its view of the surface is least obstructed. The satellite's orbit will be perpendicular to the direction of Earth's spin, so that the viewing swaths from each overpass can be compiled into whole global images. Over time, these global images will enable scientists to show and tell the stories of the causes and effects of global climate change.

The sensors on EOS Terra will not actively scan the surface (such as with laser beams or microwave pulses). Rather, the sensors work much like a camera. Sunlight that is reflected by Earth, and heat that is emitted from Earth, will pass through the apertures of Terra sensors. This radiant energy will then be focused onto specially designed detectors that are sensitive to selected regions of the electromagnetic spectrum that range from visible light to heat. The

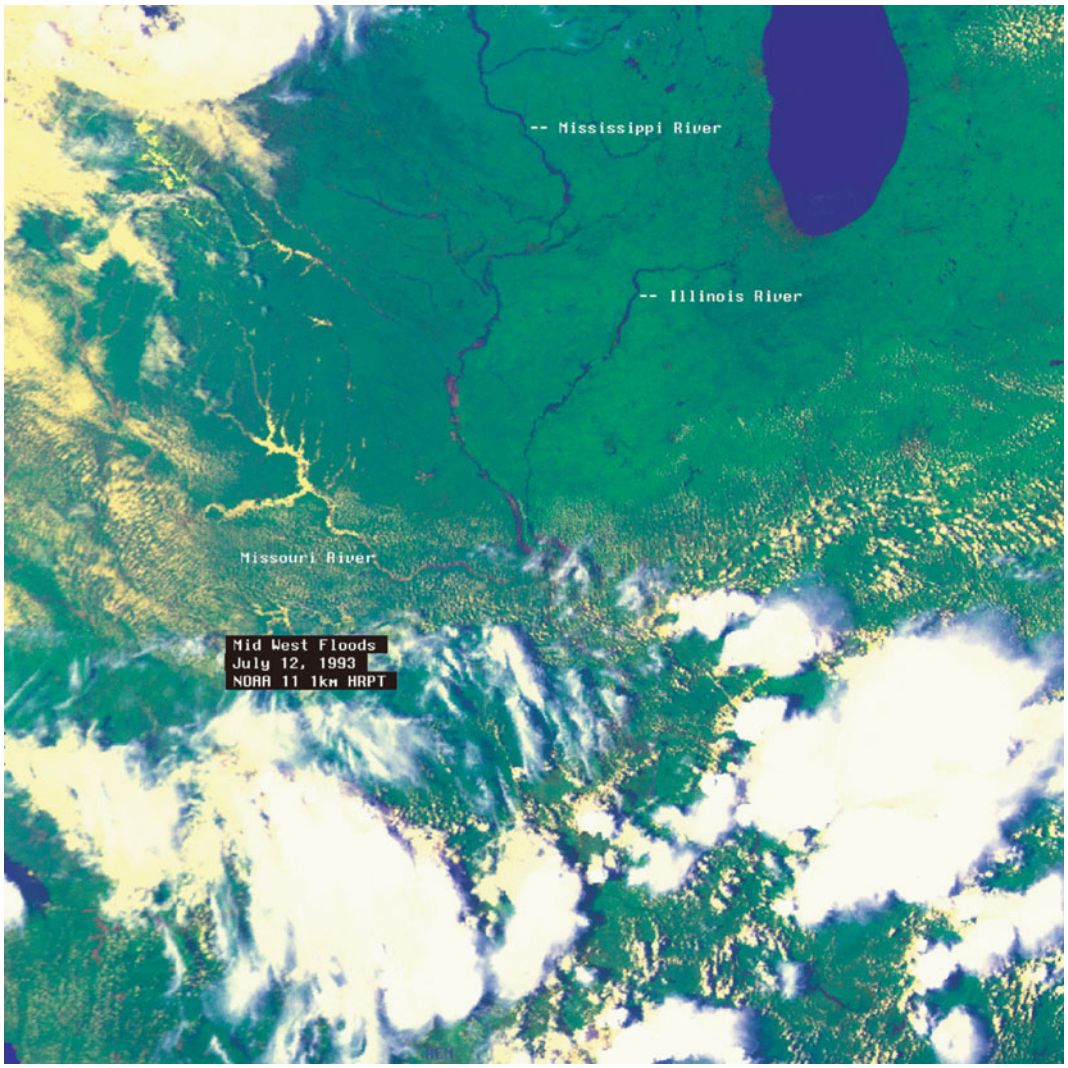


Fig. 29.6 AVHRR image of the central United States showing the Mississippi floods of 1993 (Courtesy NOAA)

information produced by these detectors will then be transmitted back to Earth and processed by computers into images that we can interpret.

The five Terra onboard sensors are

1. ASTER, or Advanced Spaceborne Thermal Emission and Reflection Radiometer
2. CERES, or clouds and earth's radiant energy system
3. MISR, or multi-angle imaging spectroradiometer
4. MODIS, or moderate-resolution imaging spectroradiometer
5. MOPITT, or measurements of pollution in the troposphere

MODIS provides continuous global coverage every 1–2 days and collects data from 36 spectral bands. Two bands (1–2) have a resolution of 250 m. Five bands (3–7) have a resolution of 500 m. The remaining bands (8–36) have a resolution of 1,000 m. The swath width for MODIS is 2,330 km. In its application, MODIS can be viewed in some ways as a higher resolution version of AVHRR. The MODIS image mosaic shown in Fig. 29.7a–c is a "greenness" map of the United States in January, April, and June 2001. The NDVI index (or greenness) is calculated from several of the MODIS bands and the seasonal variation in vegetation vigor can be seen in this image sequence.

ASTER provides 14 spectral bands with 15- to 90-m resolution, depending on bands. ASTER does not acquire data continuously, and its sensors are activated only to collect specific scenes upon request. The instrument consists of three separate telescopes, each of which provides different spectral range and resolution. The VNIR (visible and near-infrared) sensor provides 4 bands at 15-m resolution. The SWIR (short-wave infrared) sensor provides 6 bands at 30-m resolution. The TIR (thermal infrared) sensor provides 5 bands at 90-m resolution. The swath width for all sensors is 60 km.

ASTER data is generally available in the universal transverse mercator (UTM) projection, although some individual

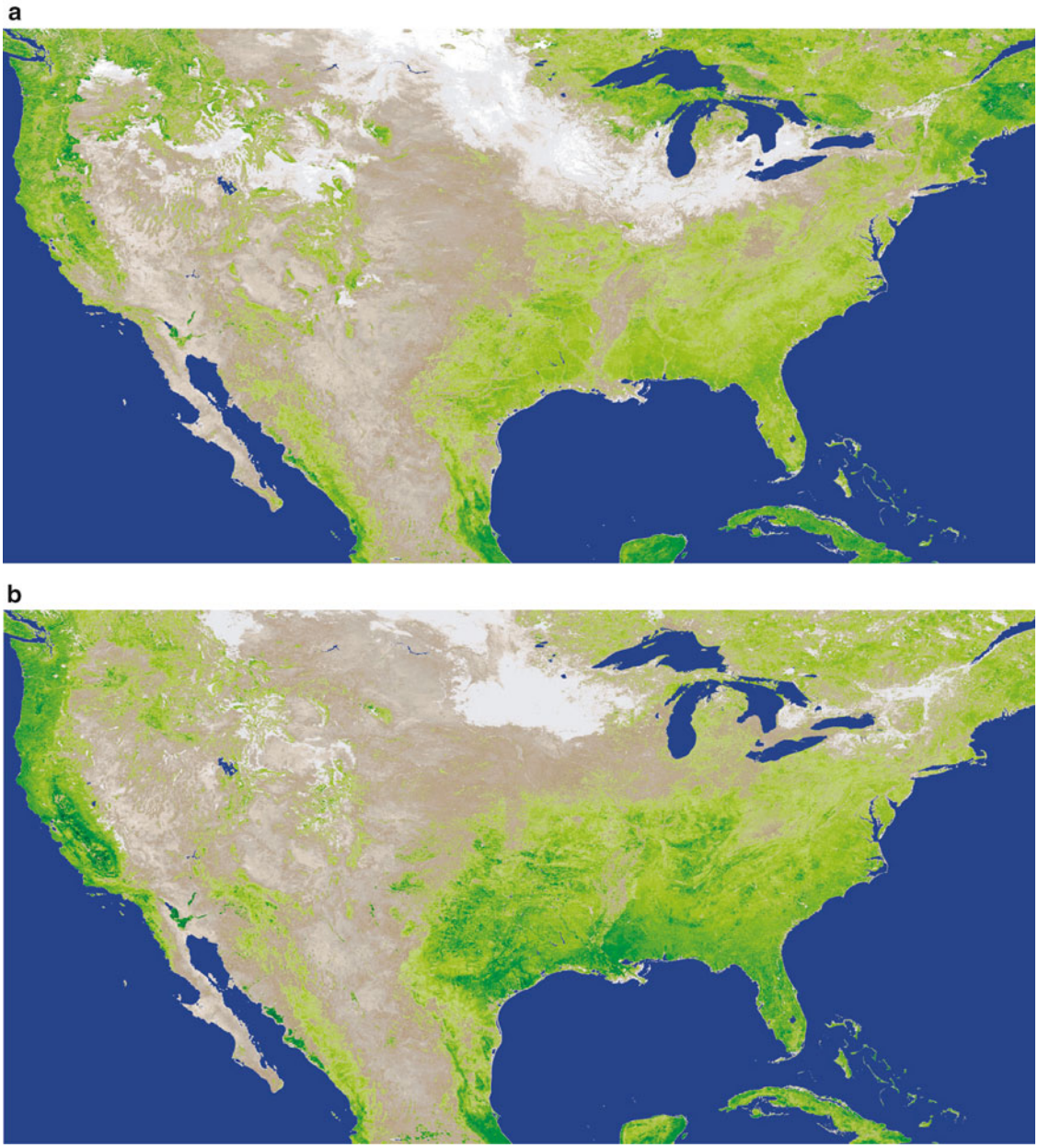


Fig. 29.7 MODIS NDVI image of the United States (Courtesy NASA). (**a**) January 2001, (**b**) April 2001, and (**c**) June 2001

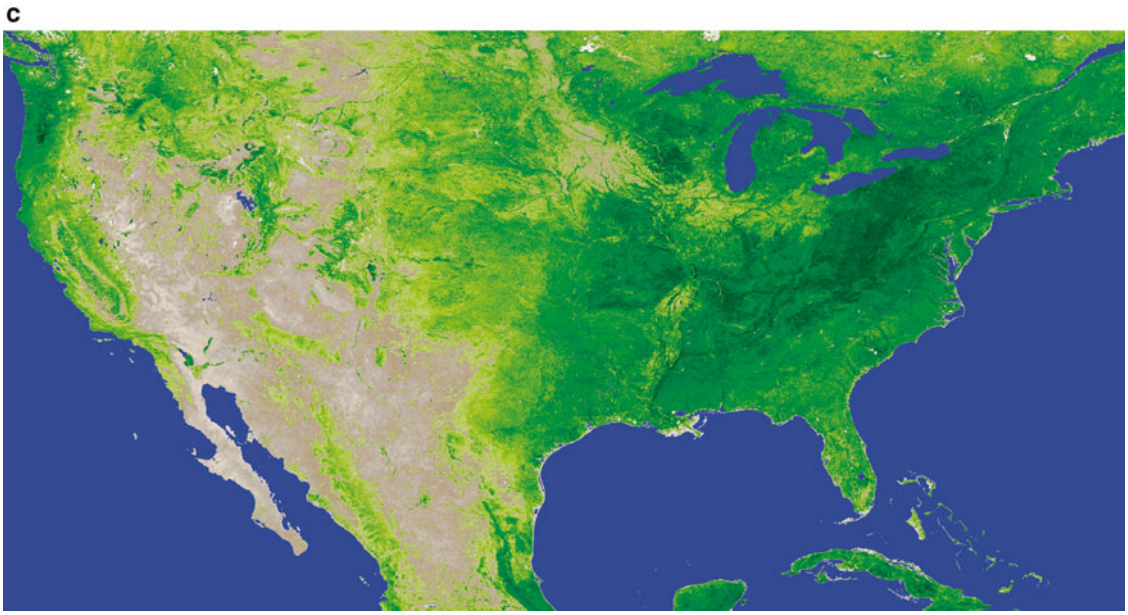


Fig. 29.7 (continued)

scenes may be cast to an alternative projection. The data is referenced to the World Geodetic Survey (WGS) System of 1984 (WGS84). Files are in the HDF-EOS format, and are distributed on CD-ROM, DVD, DLT, 8-mm tape, and file transfer protocol (FTP).

The ASTER data have about the same spatial resolution as Landsat, but they provide more spectral bands, which provide scientists with a greater ability to characterize various surface phenomena. Figure 29.8 shows an ASTER image of Washington, DC in color-infrared.

The life expectancy of the EOS Terra mission is 6 years. It will be followed in later years by other EOS spacecraft that take advantage of new developments in remote sensing technologies.

29.4 Image Processing and Geographic Information Systems

Since the early days of monitoring the Earth by orbiting spacecraft, the development of computer-aided techniques for reliably identifying many categories of surface features within a remotely sensed scene, either by photo interpretation of enhanced images or by classification, ranks in itself as an outstanding achievement. Numerous practical uses of such self-contained information are made without strong dependence on other sources of complementary or supporting data. Thus, automated data processing assists in recognizing and mapping, for example, major crop types, estimating their yields, and spotting early warning indicators of potential disease or loss of vigor.

However, many applications, particularly those involving control of dynamic growth or change systems, decision making in management of natural resources, or exploration for nonrenewable energy or mineral deposits, require a wide variety of input data (from multiple sources) not intrinsic to acquisition by space-borne sensors such as those on Landsat, the commercial satellite SPOT, and others of similar purpose. Data from remote sensing satellites combined in a geographic information system with geospatial data on themes such as soils, terrain, geology, and hydrology provide the means for characterizing the land surface over extended areas.

Some data are essentially fixed or time-independent—slope, aspect, rock types, drainage patterns, archaeological sites, etc.—in the normal span of human events. These data are usually collected and shown on maps. When digitized, they are included as layers in a GIS. Other data come from measurements or inventories conducted by people on the ground or in the air—such as weather information, population censuses, soil types and so forth. These too can be incorporated as GIS data layers. However, many vital data are transient or ephemeral—crop growth, flood water extent, insect infestation, limits of snow cover, etc.—and must be collected in a timely sense. Remote sensing data play a key role in this last instance, and in fact satellite monitoring is often the only practical and cost-effective way to acquire frequent data over large regions.

Using these tools we can relate GIS data layers and features shown on remote sensing imagery. For example, Fig. 29.9 (BGS and DPHE 2001) shows a map of arsenic concentrations in shallow water wells in Bangladesh. Comparing this map with the Landsat image of Bangladesh

Fig. 29.8 ASTER image of Washington, DC (Courtesy NASA)

(Fig. 29.10), one can see that the areas of high arsenic concentrations correspond with the coastal flood plains of the country (see also Chap. 12, this volume).

29.5 Landscape Ecology and Disease Systems

Vector-borne diseases usually have complex life cycles (Fig. 29.11). For any given system, the host(s), vector(s), and pathogen are each subjected to a variety of "pressures" within the ecosystem. Many environmental factors drive or constrain the system: weather and climate, food and space resources, predators, and parasites. For example, vertebrate hosts are affected by food quantity and quality, availability of nesting sites, and exposure to predators or parasites. The vector is affected by temperature, humidity, food resources (which may differ between adult and immature stages), and by predators and parasites. The pathogen is affected by host immune status and the frequency and timing of contact between vector and host. Temperature has a major impact on the development rate of the pathogen when it is developing in the vector. With the exception of vector-borne diseases that have humans as the primary or only vertebrate host (e.g., malaria, dengue, and bancroftian filariasis), humans often become involved in the transmission cycle by accident, and do not develop sufficient parasitemia or viremia to infect additional vectors; that is, they are "dead-end" hosts (see also Chap. 20, this volume).

Landscape ecology deals with the mosaic structure of landscapes and ecosystems, and considers the spatial heterogeneity of biotic and abiotic components as the underlying mechanism that determines the structure of ecosystems (Forman and Godron 1986; Kitron 1998). Vector-borne diseases are complex in their spatial and temporal distribution (Fig. 29.12). All components of the system, pathogen, vector, and host must occur together in time and space for epizootics or epidemics to occur. Variations in landscape structure create a patchwork of suitable and unsuitable habitats, which lead to focal disease activity. Barriers, such as water bodies, deserts, or mountain ranges, may prevent the occurrence of a pathogen in an otherwise suitable location. Remotely sensed images of the land, coupled with other geographic data that characterize the landscape, help us to understand the ecosystem structure and to identify areas where risk of disease is greater.

29.6 Relating the Disease Cycle to the Underlying Physical Environment

GIS, global positioning systems (GPS), remote sensing, and spatial statistics provide techniques and methods that we can use to analyze and integrate the spatial component into studies of the ecology and epidemiology of vector-borne disease (Kitron 1998). These spatial tools, together with the concepts of landscape ecology, can help us to better

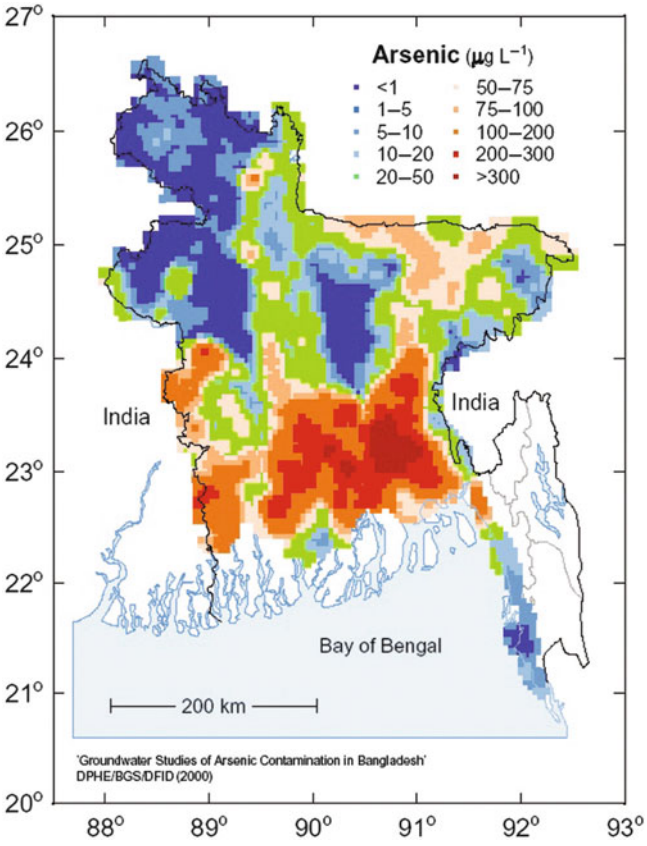


Fig. 29.9 Arsenic concentrations in groundwater in Bangladesh (Courtesy BGS)

understand emerging infectious diseases as well as the potential impact of global change on vector-borne diseases.

Several studies have utilized satellite-derived estimates of ground temperature and moisture to predict the distribution or risk of vector-borne disease. Malone et al. (1997) used an annual time series of diurnal temperature difference (dT, derived from day-night pairs of AVHRR data) to map high and low prevalence zones of the parasite, *Schistosoma mansoni* in Egypt. The dT values were associated with the depth of the water table, a major environmental determinant of the distribution of this water and snail-associated parasite.

Lindsay and Thomas (2000) characterized the climate at sites in Africa where surveys for lymphatic filariasis had taken place by using computerized climate surfaces. Logistic regression analysis of the climate variables predicted with 76% accuracy whether sites had microfilaremic patients or not. A map of the risk of lymphatic filariasis infection across Africa, built from the logistic equation in a GIS, compared favorably with expert opinion. A further validation, using a quasi-independent data set, showed that the model correctly predicted 88% of the infected sites. They then used a similar procedure to map the risk of microfilaremia in Egypt, where the dominant vector species differs from those in sub-Saharan Africa. By overlaying risk maps on a 1990 population grid, and adjusting for recent population increases, they estimated that around 420 million people would be exposed to this infection in Africa in the year 2000. The approach described by these authors could be used to produce a sampling frame for conducting filariasis surveys in countries that lack accurate distribution maps.

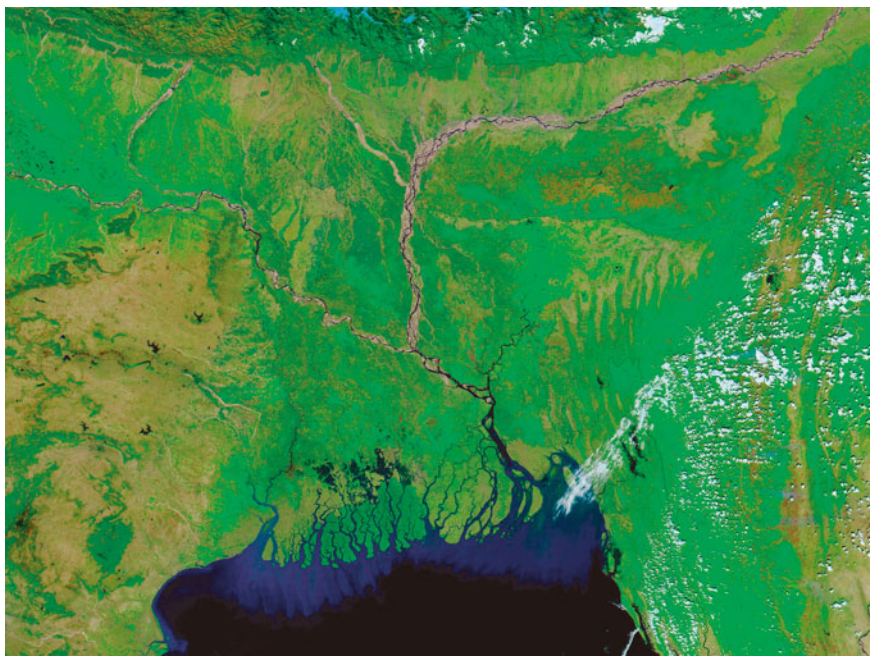


Fig. 29.10 Landsat image of Bangladesh (Courtesy USGS)

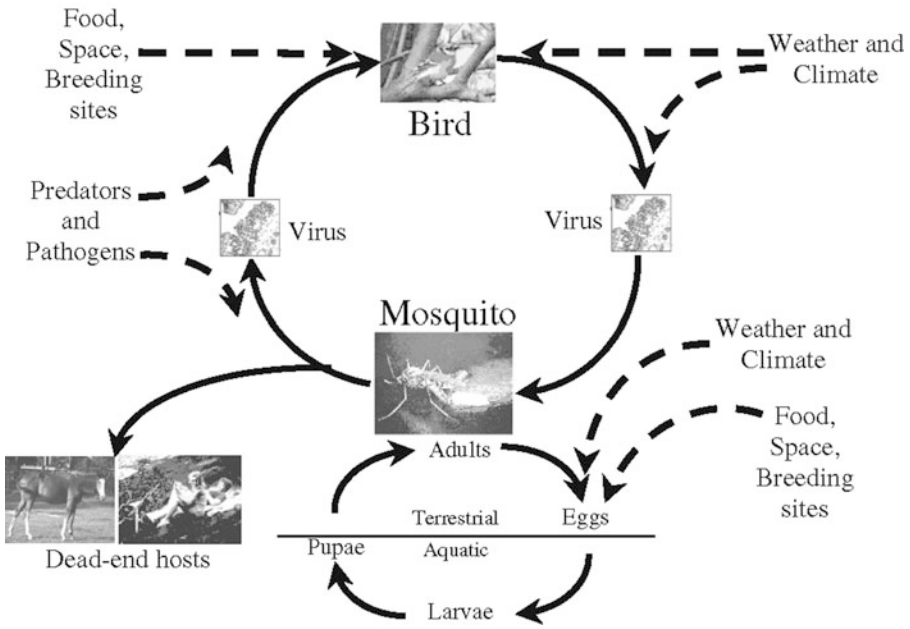


Fig. 29.11 A typical arbovirus cycle showing sources of environmental pressure on the system

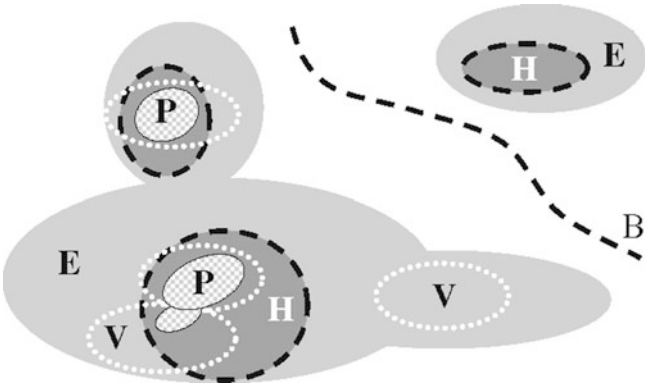


Fig. 29.12 Spatial relationships in the arbovirus-vector-host ecosystem. All components are limited by a suitable environment (*E*). Populations of the vertebrate host (*H*) do not necessarily overlap with the distribution of the vector (*V*). It is only in those regions of *E* where both *H* and *V* overlap that the pathogen (*P*) can survive. Barriers (*B*) to dispersal may limit the presence of *V* or *H*

American and Mexican researchers collaborated in a study to map habitats of the malaria vector, *Anopheles albimanus*, using Landsat Thematic Mapper (TM) imagery and extensive ecological and epidemiological data (Beck et al. 1994). Pixel categories were associated with landscape type both by ground surveys and by comparison to color-infrared photography of the study area. Since the relationship between landscape type and suitability as larval habitat for *Anopheles albimanus* was very strong, it was possible to correctly distinguish between villages with high and low vector abundance with an overall accuracy of 90%. These analyses indicated that the most important landscape elements in terms of explaining mosquito abundance were the proportions of transitional wetlands and unmanaged pasture. Using these two landscape elements as predictors, they were able to correctly distinguish villages with high and low mosquito abundance with an overall accuracy of 90%.

Robinson et al. (1997) mapped the distribution of tsetse (*Glossina* spp.) habitat by using climate and remotely sensed vegetation data. Coarse-resolution (7.6 km) AVHRR images were combined with smoothed climate surfaces derived from continent-wide, long-term weather station records. Predictions were improved by subdividing habitats prior to classification. This system might be improved by using satellite-derived weather data and by using finer grain imagery (e.g., 1.1 km AVHRR).

Several groups have used remotely sensed data to improve our understanding of Lyme disease. Glass et al. (1992), for example, conducted a case–control study of Lyme disease in Baltimore County, Maryland. Land use/land cover maps (derived from Landsat TM imagery) were combined with soils, elevation, geology, and watershed maps to evaluate risk of exposure to Lyme disease and its vector, *I. scapularis*. The risk of disease was significantly lower in highly developed areas, and risk decreased with increasing distance from forests. Dister et al. (1997) used Landsat TM images to evaluate Lyme disease exposure risk on 337 residential premises in two communities in Westchester County, New York. Premises were categorized as no, low, or high risk based on seasonally adjusted densities of *I. scapularis* nymphs (previously determined by sampling). Spectral indices from the TM scene provided relative measures of vegetation structure and moisture (wetness) as

well as vegetation abundance (greenness). They used GIS to spatially quantify and relate the landscape variables to risk category. A comparison of the two communities showed that the community with more high-risk premises was significantly greener and wetter than the community with fewer high-risk premises. Furthermore, high-risk premises were significantly greener and wetter than lower risk premises in the high-risk community. The high-risk sites appeared to contain a greater proportion of broadleaf trees, while lower risk sites were interpreted as having more non-vegetative cover or open lawn. The ability to distinguish these fine-scale differences among communities and individual properties illustrates the efficiency of a remote sensing/GIS-based approach for identifying peridomestic risk of Lyme disease over large geographic areas.

More recently, Moncayo et al. (2000) combined Landsat TM images with aerial videography to generate a map of landscape elements around 15 human and horse cases of eastern equine encephalomyelitis (EEE) in southeastern Massachusetts. EEE exists in enzootic foci transmitted between birds and the mosquito, *Culiseta melanura* (the enzootic vector). In Massachusetts and surrounding states, epidemic/epizootic transmission may involve as many as six additional mosquito species (*Aedes canadensis*, *Aedes vexans*, *Culex salinarius*, *Coquillettidia perturbans*, *Anopheles quadrimaculatus*, and *Anopheles punctipennis*), each with its characteristic larval and adult habitat requirements. Stepwise regression analysis showed that wetlands and more specifically, deciduous wetlands, were the most important major class element. Deciduous wetlands accounted for up to 72.5% of the observed variation in the host-seeking populations of *A. canadensis*, *A. vexans*, and *C. melanura*. The authors propose combining habitat mapping with street maps to identify and prioritize areas in need of vector mosquito control.

Even coarse-grain satellite imagery can be used to advantage. Daniel and Kolár (1990) analyzed a 25-year database on the distribution and abundance of *I. ricinus*, the vector of tick-borne encephalitis in Europe, in relation to land cover types derived from a 41 × 41 km section of a Landsat multispectral scanner (MSS) scene. They showed that *I. ricinus* is associated with specific land cover types, which allowed them to generate risk maps that could be used in public education and other prevention programs.

29.7 Technical Issues and Limitations

29.7.1 Analytical and Statistical Issues

Although a picture may be worth a thousand words, it is still necessary to know whether the patterns we think we see are, in fact, statistically significant rather than the result of random noise. Better integration of GIS and spatial statistics software would be advantageous, but it is probably not desirable or economically feasible for each GIS package to integrate a spatial statistics module. Rather, it would be helpful for developers from the two fields to agree upon common import/export formats to permit the rapid movement of data between the two types of software. Bailey (1994) provides a good review of progress in the integration of spatial statistics and GIS. As the focus of GIS and remote sensing in vector-borne disease moves from description to prediction, there will be a greater need for analytical techniques that simultaneously deal with space and time. Cressie (1996) identified several issues and approaches to dealing with spatiotemporal processes.

29.7.2 Modeling and Simulation

Perhaps the greatest potential for applying remotely sensed information to the surveillance, prediction, and control of vector-borne and zoonotic diseases is the ability to use these data in predictive models. Models come in a great variety, and it is crucial to use the right type of model to answer a specific question. Thus, models designed to elucidate the mechanism of a particular ecological process may not be good predictive models and vice versa. An area where considerable progress is being made is in landscape ecology, where issues of sustainable harvestable resources and preservation of endangered species habitats have become crucial. A good survey of theory and applications in this area can be found in Turner and Gardner (1991).

29.7.3 Temporal and Spatial Resolution

The spatial scale of satellite imagery should match the scale of the object of study at ground level. Early remote sensing studies (e.g., Hayes et al. 1985) were limited to Landsat MSS images, with a pixel resolution of about 80 m. This placed severe limitations on the size of habitats that could be detected. Today, with the advent of commercial imagery with 1-m resolution, the lack of sufficient high-resolution, ground-level data has become a limiting factor in the application of remote sensing to solving issues in vector-borne disease ecology.

29.7.4 The Issue of Scale

Ecological phenomena exhibit patterns at different scales (e.g., Turner et al. 1991; Quattrochi and Goodchild 1997). Vector-borne diseases are no exception to this phenomenon (Korenberg 1989). Thus, a virus-vector-vertebrate host

system, such as LaCrosse encephalitis, can exhibit distributional pattern at the continental scale, the regional scale, and at the local scale (Fig. 29.13). For a given question or hypothesis about any disease, there is probably an optimal scale at which to measure the system components. Thus, it is crucial to clearly understand the scale(s) at which different components of the disease system operate.

29.7.5 Imagery Issues

The resolution of remotely sensed imagery may not be satisfactory for studies at a particular scale. Resolution here can refer to space, time, or spectral composition. In particular, studies of some diseases, such as dengue and LaCrosse encephalitis, would benefit from the ability to identify water-holding containers on individual premises. This can be accomplished by using high-resolution aerial photography, but temporal coverage will be sparse due to the high cost of flying such missions. Also, the spectral coverage will be very narrow, which reduces the utility of the imagery.

Cloud cover and other disturbances are a perennial problem. One anticipated use of remote sensing is in evaluating the threat of disease outbreaks following floods, hurricanes, or other natural disasters. Because many of these disasters are weather related, it is not uncommon for the areas of interest to be completely covered by clouds for weeks following the event. Perhaps additional studies on the use of real-time radar in conjunction with historical imagery in other spectral regions or similar systems could improve prospects for using remote sensing in this setting (e.g., Imhoff and McCandless 1988).

Landscape feature classification is an important tool for understanding and modeling vector-borne and zoonotic diseases. One or more specific land cover types (e.g., scrub, piñon-juniper, oak-hickory forest, etc.) often define these diseases. Thus, it is important to be able to distinguish between cover types. All of the problems related to slope, aspect, and shading as they impact the remotely sensed image also need to be dealt with. Soil type, elevation, slope, and aspect often have a major impact on the distribution of vertebrate hosts of these disease agents, especially small mammals. It should be possible, at least in theory, to infer some of these qualities from the remotely sensed spectral signals, particularly from hyperspectral imagery.

29.7.6 Ground Data Issues

One of the first things that strikes the entomologist or epidemiologist when looking at remotely sensed imagery is the enormous amount of information in comparison to the available data on the diseases we would like to study. For many diseases, such as cancer and birth defects, there are well-developed registries that give detailed information about cases. Such information allows the epidemiologist to search for associations between environmental indicators (from the remotely sensed data) and the location of cases. Unfortunately, reporting systems for vector-borne diseases are generally poorly developed or entirely absent. For example, dengue which, as reported above, is responsible for as many as 50–100 million cases each year, is not even reported in many countries. Only the more severe cases are entered into the reporting system. Even in the United States, the vector-borne diseases are generally grossly underreported. To successfully apply remote sensing to the surveillance, prevention, and control of vector-borne and zoonotic diseases, we need well-designed surveillance systems to provide the "ground truth" data to validate the models that are being developed.

29.7.7 Privacy Issues

A separate but extremely important issue relating to the collection of disease surveillance data is the issue of privacy. There is concern on the part of some citizens that their personal privacy would be violated by having the precise location of their residence, school, workplace, etc., entered into a database, particularly if that information in also associated with health-related information. Thus, many reporting systems show only the city or county of residence. This severely limits the utility of the data for remote sensing studies.

One way to bypass this problem, at least with the vector-borne and zoonotic diseases, is to monitor disease activity in the vector or the wild vertebrate host (in some situations, susceptible domestic animals can also be monitored). Detailed protocols have been developed for collecting the appropriate data from these parts of the disease system (e.g., Moore et al. 1993; Moore and Gage 1996). Data from these field-oriented surveillance systems can be used to predict disease activity in the human (or domestic animal) population. Because these surveillance systems are generally removed in space (often rural) and time (activity precedes disease in humans) from the human population at risk, there is often a shortage of funding and staffing for these programs. Thus, the surveillance agency wants to collect as much useful information as possible in the shortest number of hours with the smallest possible work force. To do this, programs are increasingly moving to automated field data recording systems, GPS, and similar equipment to reduce paperwork, data entry error, and similar time-consuming problems. The Centers for Disease Control is currently funding the development of an integrated database, GIS,

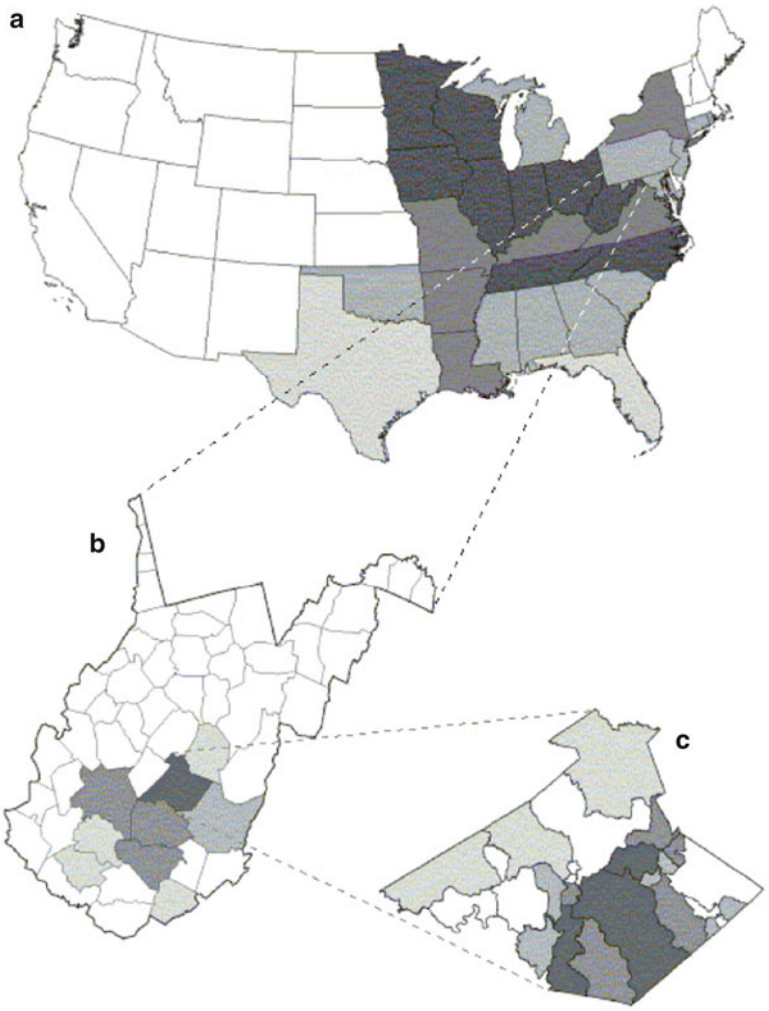


Fig. 29.13 Effect of scale on pattern and distribution of LaCrosse encephalitis in the United States; patterns are evident at several spatial levels. *Darker colors* indicate more viral activity

and decision support system that will speed the collection of vector-borne disease related data. Once completed, the system, called the National Electronic Arbovirus Reporting System (NEARS), will be accessible via the Worldwide Web. Data at differing spatial resolutions can be downloaded for use with remotely sensed data.

29.8 Prospects for the Future

The utility of remotely sensed data for improving our understanding, prevention, and control of vector-borne and zoonotic diseases has been abundantly demonstrated. Even greater opportunities exist for applying the imagery from current and future remote-sensing platforms. Much of the eventual progress in applying that imagery to the solution of human problems, including the diseases of interest to us, will depend on the priorities set by universities, governments, and funding agencies. There is a need for cross-disciplinary programs in universities to provide qualified and broadly trained researchers for this growing field. Collaborative research, within and between agencies and departments, should be encouraged. It is no longer possible for a single researcher or a small group within a single discipline to "do it all." By sharing valuable resources and insights (expertise), progress can be more rapid and costs can be reduced.

See Also the Following Chapters. Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 20 (The Ecology of Soil-Borne Human Pathogens) • Chapter 28 (GIS in Human Health Studies)

References and Further Reading

Bailey TC (1994) A review of statistical spatial analysis in geographical information systems. In: Fotheringham S, Rogerson P (eds) Spatial analysis and GIS. Taylor and Francis, London, pp 13–44

Beck LR, Rodriguez MH, Dister SW, Rodriguez AD, Rejmankova E, Ulloa A, Meza RA, Roberts DR, Paris JF, Spanner MA, Washino RK, Hacker C, Legters LJ (1994) Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. Am J Trop Med Hyg 51:271–280

Beck LR, Lobitz BM, Wood BL (2000) Remote sensing and human health: new sensors and new opportunities. Emerg Infect Dis 6 (3):217–227

BGS, DPHE (2001) Arsenic contamination of groundwater in Bangladesh. In: Kinniburgh DG, Smedley PL (eds) British Geological Survey technical report WC/00/19. British Geological Survey, Keyworth

Clarke KC, McLafferty SL, Tempalski BJ (1996) On epidemiology and geographic information systems: a review and discussion of future directions. Emerg Infect Dis 2:85–92

Cline BL (1970) New eyes for epidemiologists: aerial photography and other remote sensing techniques. Am J Epidemiol 92:85–89

Colwell RR (1996) Global climate and infectious disease: the cholera paradigm. Science 274:2025–2031

Cressie N (1996) Statistical modeling of environmental data in space and time. In: Mowrer HT, Czaplewski RL, Hamre RH (eds) Spatial accuracy assessment in natural resources and environmental sciences: second international symposium, U.S. Department of Agriculture, Forest Service, GTR RM-GTR-277, Fort Collins, CO, pp 1–3

Daniel M, Kolár J (1990) Using satellite data to forecast the occurrence of the common tick *Ixodes ricinus* (L.). J Hyg Epidemiol Microbiol Immunol 34:243–252

Dister SW, Fish D, Bros SM, Frank DH, Wood BL (1997) Landscape characterization of peridomestic risk for Lyme disease using satellite imagery. Am J Trop Med Hyg 57(6):687–692

Estes JE, Loveland TR (1999) Characteristics, sources, and management of remotely-sensed data. In: Longely PA, Goodchild MF, Maguire DJ, Rhind DW (eds) Geographical information systems, 2nd edn. Wiley, New York, pp 667–675

Forman RTT, Godron M (1986) Landscape ecology. Wiley, New York

Glass GE, Morgan JM III, Johnson DT, Noy PM, Israel E, Schwartz BS (1992) Infectious disease epidemiology and GIS: a case study of Lyme disease. GeoInfoSystems 2:65–69

Gratz NG (1999) Emerging and resurging vector-borne diseases. Annu Rev Entomol 44:51–75

Hayes RO, Maxwell EL, Mitchell CJ, Woodzick TL (1985) Detection, identification and classification of mosquito larval habitats using remote sensing scanners in earth-orbiting satellites. Bull WHO 63:361–374

Imhoff ML, McCandless SW (1988) Flood boundary delineation through clouds and vegetation using L-band space-borne radar: a potential new tool for disease vector control programs. Acta Astronaut 17:1003–1007

Jetten TH, Focks DA (1997) Potential changes in the distribution of dengue transmission under climate warming. Am J Trop Med Hyg 57:285–297

King M (2000) EOS measurements, National Aeronautics and Space Administration, Goddard Space Flight Center. http://eospso.gsfc.nasa.gov/ftp_docs/measurements.pdf

Kitron U (1998) Landscape ecology and epidemiology of vector-borne diseases: tools for spatial analysis. J Med Entomol 35(4):435–445

Korenberg EI (1989) Population principles in research into natural focality of zoonoses. Sov Sci Rev F Physiol Gen Biol 3:301–351

Lindsay SW, Thomas CJ (2000) Mapping and estimating the population at risk from lymphatic filariasis in Africa. Trans R Soc Trop Med Hyg 94(1):37–45

Linthicum KJ, Anyamba A, Tucker CJ, Kelley PW, Myers MF, Peters CJ (1999) Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya. Science 285:397–400

Malone JB, Abdel-Rahman MS, El Bahy MM, Huh OK, Shafik M, Bavia M (1997) Geographic information systems and the distribution of *Schistosoma mansoni* in the Nile delta. Parasitol Today 13:112–119

Martens WJM, Niessen LW, Rotmans J, Jetten TH, McMichael AJ (1995) Potential risk of global climate change on malaria risk. Environ Health Perspect 103:458–464

Moncayo AC, Edman JD, Finn JT (2000) Application of geographic information technology in determining risk of eastern equine encephalomyelitis virus transmission. J Am Mosq Control Assoc 16(1):28–35

Moore CG, Gage KL (1996) Collecting methods for vector surveillance. In: Beaty BJ, Marquardt WC (eds) The biology of disease vectors. University Press of Colorado, Niwot, pp 471–491

Moore CG, McLean RG, Mitchell CJ, Nasci RS, Tsai TF, Calisher CH, Marfin AA, Moore PS, Gubler DJ (1993) Guidelines for arbovirus surveillance in the United States. U. S. Dept. of Health and Human Services/Centers for Disease Control, Fort Collins

Ostfeld RS, Jones CG, Wolff JO (1996) Of mice and mast: ecological connections in eastern deciduous forests. Bioscience 46:323–330

Pope KO, Rejmankova E, Savage HM, Arredondo-Jimenez JI, Rodriguez MJ, Roberts DR (1994) Remote sensing of tropical wetlands for malaria control in Chiapas, Mexico. Ecol Appl 4:81–90

Quattrochi DA, Goodchild MF (eds) (1997) Scale in remote sensing and GIS. Lewis Publishers, Boca Raton

Reeves WC, Hardy JL, Reisen WK, Milby MM (1994) Potential effect of global warming on mosquito-borne arboviruses. J Med Entomol 31:324–332

Robinson T, Rogers D, Williams B (1997) Mapping tsetse habitat suitability in the common fly belt of Southern Africa using multivariate analysis of climate and remotely sensed vegetation data. Med Vet Entomol 11:235–245

Shope RE (1992) Impacts of global climate change on human health: spread of infectious diseases. In: Majumdar SK, Kalkstein LS, Yarnal B, Miller EW, Rosenfeld LM (eds) Global climate change: implications, challenges and mitigation measures. Pennsylvania Academy of Science, Easton, pp 361–370

Short NM (ed) (2003) The remote sensing tutorial. National Aeronautics and Space Administration/Goddard Space Flight Center, Greenbelt

Srivastava A, Nagpal BN, Saxena R, Sharma VP (1999) Geographic information system as a tool to study malaria receptivity in Nadiad Taluka, Kheda District, Gujarat, India. Southeast Asian J Trop Med Public Health 30(4):650–656

Star JL, Estes JE, Davis F (1991) Improved integration of remote sensing and geographic information systems: a background to NCGIA initiative 12. Photogramm Eng Remote Sens 57:643–645

Turner MG, Gardner RH (eds) (1991) Quantitative methods in landscape ecology. Springer, New York

Turner SJ, O'Neill RV, Conley W, Conley MR, Humphries HC (1991) Pattern and scale: statistics for landscape ecology. In: Turner MG, Gardner RH (eds) Quantitative methods in landscape ecology. Springer, New York, pp 17–49

Mineralogy of Bones

30

H. Catherine W. Skinner

Contents

30.1 Introduction

Medical geology encompasses many scientific endeavors with global activities and impact, but it also includes aspects that are very personal and individual. Local environment is sampled through what is ingested and inhaled whether or not it is salubrious, marginal, or downright unhealthy. Human bodies react to the remarkable range of natural and man-made chemicals that they are exposed to every day. It is common knowledge that a range of nutrients is required, but some of the chemicals can be hazardous to our health. This chapter focuses on bones, and specifically the mineral portion in these tissues, the component essential to the functions of these organs. These discussions illustrate several attributes of the emerging field of medical geology. The scientific information outlined herein is drawn from a diversity of disciplines and expertise, from biophysical and biochemical sciences, from physicians and dentists, and from geologists, mineralogists, and engineers. This knowledge enables us to address the many individual and collective roles that minerals play in the body. A disorder that affects people on every continent, osteoporosis, is presented as an example of how information on minerals can be applied. It is only one of the possible targets of opportunity where mineralogical/geological expertise has, and continues to have, the potential to ameliorate suffering and promote better global and personal health. Selected classic and recent references are included to whet the appetite of those who will make contributions to our knowledge in the future.

30.2 The Skeleton and Mineralized Tissues

Humans can be distinguished from all other mammalian species by their skeletons which are composed of over 200 bones and 32 teeth. Each of these skeletal components is a separate organ composed of mineralized tissues. Created by the actions of distinct cell systems, the tissues are true

H.C.W. Skinner (✉)
Department of Geology and Geophysics and Orthopaedics and Rehabilitation, Yale University, New Haven, CT, USA
e-mail: catherine.skinner@yale.edu

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_30,

composites, an intimate association of extracellular macro- and microbioorganic molecules, and inorganic mineral materials. The tissues are in dynamic equilibrium with a highly controlled fluid whose chemical composition resembles that of the ocean: highly oxidized, with a pH around 7 with sodium, Na^+, and chlorine, Cl^-, as the dominant ionic species Halstead (1974).

The mineral of bones and teeth is a calcium phosphate that closely resembles the naturally occurring mineral, hydroxylapatite (Gaines et al. 1997), and because of its intimate association with biological activity and molecules it is known as bioapatite. The mineral formed and maintained in the soft tissue or matrix of bones has peculiar attributes. Each mineralized tissue is a remarkable and unique chemical repository whose maintenance is only beginning to be comprehended as part of the dynamic skeletal system.

30.2.1 Normal Mineralized Tissues: Bones and Teeth

The relative amounts of mineral to bioorganic components and the distinct spatial aggregations or structures of the mineral-matrix combination have been objects of investigation for over 250 years. Four different types of normal human mineralized tissues have been described (Glimcher 1976; Mann 2000). With increased sensitivity and availability of analytical techniques, the four types have been shown to have discrete cell systems, chemical, and spatial expressions that change during growth, with age, or with disease. Studies, especially those from different life stages, have allowed us to identify essential participants and gain some understanding of the expected or "normal" state and the reactions required to maintain function appropriate for the skeleton.

Three of the four tissues are found in teeth: enamel, dentine, and cementum (Fig. 30.1) (Miles 1967), and the fourth is in bone. Subsets of all four types have been described and amplified with each new and more sophisticated analytical technique. Optical examination and specialized methodology on thin sections of mineralized tissues at high resolution has defined the anatomy of their components, the cells, extracellular products, or the typical textures (Fig. 30.1a). Collagen is the most common protein in the body; a normal biomolecular component of all connective tissues and the dominant protein (approximately 90 wt.%) in the organic fraction of bone tissues. Fibroblasts are the cells that can produce collagen and collagen is also normally produced a the fist organoolecule in bone by the osteoblasts. Many biochemical varieties of the collagen molecule with slightly different amino acid compositions and intramolecular cross links have been described (Miller 1973; Skinner 1987), but type 1 collagen is typical of bone, dentine, and cementum. Collagen is not found in enamel.

Tissues typical of bones (Fig. 30.1b) have woven, lamellar, and haversian textures with mineral distribution and content that varies with tissue age, with nutrition, or with disease (Albright and Skinner 1987). All three textures may occur in cortical bone, the heavily mineralized portions of bone, or in trabecular bone, the porous and spongy segments of the organ, and are expressions of their dynamic nature. The normal response of bioapatites is to our peculiar and individual life styles attempting to maintain the essential three dimensional framework and structure but there are other deposits many of similar composition to bioapatite in humans that are not normal nor expected (Skinner 2012).

30.2.2 Pathological Apatitic Deposition

In addition to the normal bones and teeth, bioapatite may deposit pathologically, that is, in the tissues of other organs that would normally not mineralize. Nodular apatitic spherules may be detected throughout the body. One site where bioapatites may occur is in tumors, both benign and cancerous, where rapid cell production may cause accumulations of dead cells. When a cell dies it releases elements and bioorganic molecules into the surrounding fluid. Any elevated phosphate and any calcium concentrations in the circulating serum then becomes much higher than within the cell and bioapatite may nucleate and precipitate. Another likely site for bioapatite mineralization is at scar tissues, the squeal to tissue trauma, where excessive amounts of the fibrous protein collagen accumulate as a result of the normal cellular repair systems throughout the body.

Pathological bioapatite may be found in the arteries, an expression of cardiovascular disease, and in kidney "stones" (Skinner 2000a). It may occur in association with other calcium phosphate species (Table 30.1). For example, calcium pyrophosphate, $Ca_2P_2O_7 \cdot H_2O$, occurs in joints (Skinner 2000a), and octacalcium phosphate, $Ca_8H_2(PO_4)_6 \cdot 5H_2O$, and/or whitlockite, $CaMg(PO_3OH)\ (PO_4)_6$, in dental calculus, which is a mineral deposit that forms just below the gum line in the soft tissue around teeth (see Driessens and Verbeeck 1990, for a complete discussion of dental plaque mineral constituents).

To cover all calcium phosphate mineral species that may be found in the human body is beyond the scope of this presentation. An introduction to the usual and predominant mineralizing system Sect. (30.3), the submicron characteristics of the mineral Sect. (30.4), and the methodology and techniques used to determine the composition, concentration, and distribution of the mineral in mineralized tissues are presented (Sect. 30.5). The final Sect. (30.6) on

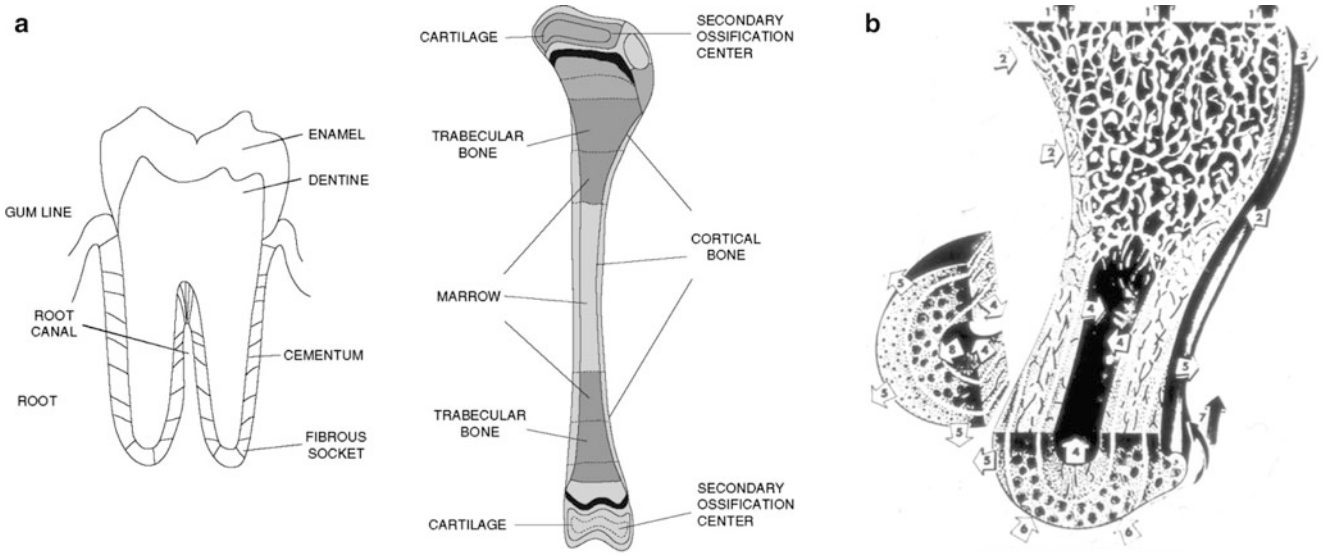


Fig. 30.1 (**a**) Sketch of a tooth and a long bone indicating the typical tissues found in these organs. (From Skinner 2000a, Figure 1, p. 356.) (**b**) Sketch of a longitudinal section and a cross section through a long bone illustrating the tissue types and textures that occur and the changes that take place during the development of a bone. *Arrows* indicate the directions of growth and remodeling. (From Albright and Skinner 1987, Figure 5–16, part A, p. 175.) Numbers at specific sites illustrate the following changes: *1:* length increase, typical growth direction of a long bone; *2:* thickness or diameter increase that also takes place as the organ changes shape and size. Some initial trabecular tissues become cortical bone tissues. The textures depicted in *2* and *5* are typical of heavily mineralized lamellar bone that results from the remodeling of trabecular tissues into cortical tissues. *3:* all bones require local remodeling to achieve final organ size and shape in order to function as part of the skeleton; *4:* remodeling is also needed to maintain the internal (marrow) cavity size and shape; *5:* extensive remodeling in the mid-shaft of long bones where previous trabeculae have been remodeled into heavily mineralized cortical tissue; *6:* the circular patterns of haversian bone express the sites of resorption and re-deposition of mineralized tissues which continue throughout life; *7:* an outer protuberance on the bone, probably the site of muscle attachment, which will change as the bone responds to growth and development; and *8:* a cross section showing distinct layers of haversian and lamellar bone tissues surrounding the marrow cavity

Table 30.1 Formulae of hydroxylapatite, bioapatite, and other calcium phosphate phases and the Ca/P ratios for their ideal compositions

Name	Formula	Ca/P (wt.7%)
Hydroxylapatite	$Ca_5(PO_4)_3(OH)$	2.15
Bioapatite[a]	$(Ca, Na, [\,])_{10}(PO_4, HPO_4, CO_3)_6(OH, F, Cl, H_2O, CO_3O, [\,])_2$	1.33–2.25
Monetite	$CaHPO_4$	1.25
Brushite	$CaHPO_4{\cdot}2\ H_2O$	1.25
Octacalcium phosphate	$Ca_8H_2(PO_4)_6{\cdot}5\ H_2O$	1.33
Tricalcium phosphate	$Ca_3(PO_4)_2$	1.875
Whitlockite	$Ca_9\ Mg(PO_3OH)\ (PO_4)_6$	1.11–1.456
Calcium pyrophosphate dihydrate	$Ca_2P_2O_7{\cdot}2\ H_2O$	1.25
Tetracalcium phosphate	$Ca_4P_2O_9$	2.50
Dahllite	$(Ca, X)_{10}(PO_4, CO_3)_6(O, OH)_{26}$[b] $CO_2 > 1\%$ $F < 1\%$	2.08
Francolite	$(Ca, X)_{10}(PO_4, CO_3)_6(O, F)_{26}$[b] $F > 1\%$ $CO_2 > 1\%$	2.08

[a]An approximate formula containing vacancies [ ]
[b]Formula and composition from McConnell (1973), where X indicates a range of cations

the disease osteoporosis illustrates how knowledge of the mineral, distinguishing normal from the abnormal, and mineral dynamics have become foci for medical research.

Although the techniques for studying the mineral are similar to those employed for all solid materials, mineralized tissues present special problems. Foremost are the difficulties of obtaining sufficient sample when the object of study is a living human. Initial diagnosis of osteoporosis usually employs noninvasive techniques such as transmission X-ray analyses (radiology) of the skeleton. Once a clinical diagnosis is made it may be followed, especially today, by extraction (biopsy) of mineralized tissue samples in spite of the difficulties of procurement and preparation (Sects. 30.5.1, 30.5.2, 30.5.3, and 30.5.4). The tiny amounts of tissues can provide the information that is essential in defining and adequately treating the disease in many patients. Whenever

mineralized tissue samples are studied, either normal or pathological, the physical and chemical characteristics of the tissues and the mineral fraction benefit future health of patients and populations.

30.3 Crystal Chemistry of the Mineral in Mineralized Tissues

The normal and much of the abnormal or pathological mineral deposits in humans is a calcium phosphate, a member of the apatite group of minerals (Gaines et al. 1997). The biodeposits conform crystal chemically most closely to the mineral species hydroxylapatite, ideal formula $Ca_5(PO_4)_3(OH)$, (hereafter abbreviated as HA), which is described in most introductory mineralogical texts (Klein and Hurlbut 1985). However, the precise composition and crystal structure of the biological mineral, bioapatite (Table 30.1) has proved difficult to pinpoint so the term "apatitic" is often used. These difficulties relate to the chemical variability reported from many analyses of the mineral from different mineralized tissues. Further, X-ray diffraction analyses, the criteria for accurate identification of any mineral species, applied to biomineral samples is only partially successful primarily because of the very small grain size (nanometers of individual particles normally a mineral aggregate) of the mineral materials. What has been shown from extensive investigations is that the biomineral, although apatitic, is not the ideal or stoichiometric chemical compound whose formula is presented above. The following formula is a more appropriate presentation:

$$(\mathrm{Ca, Na, Mg, [\,]})_{10}\ (\mathrm{PO_4, HPO_4, CO_3})_6$$
$$(\mathrm{OH, F, Cl, H_2O, CO_3, [\,]})_2$$

The brackets indicate vacancies in some lattice sites of the solid to achieve a charge-balanced solid phase.

This complicated chemical solid can be described as follows: bioapatite is predominantly a calcium phosphate mineral most closely resembling the species hydroxylapatite but usually contains many elements and molecular species other than calcium and phosphate that probably contribute to its physical attributes and reactivity, and should be part of any identification. The crystal structure of all apatite minerals is dominated by tetrahedral anions. In the case of hydroxylapatite and bioapatite it is phosphate, the phosphorous-oxygen tetrahedral anionic groups $(PO_4)^{3-}$, that forms the backbone of the structure and bonds predominantly to calcium, but other cation and anion species, vacancies, and complex molecular groups e.g. CO_3 are usually detected on analysis.

30.3.1 The Mineralizing System: $CaO–P_2O_5–H_2O$

The chemical system that defines bioapatitecan be simplified to $CaO–P_2O_5–H_2O$. Figure 30.2 is an experimentally determined phase diagram. A portion of the chemical system for the mineralization of bones and teeth, it is a summary of the solid and liquid phases that exist at equilibrium at a specific temperature and pressure in the simplified system (Skinner 1973a). The following solids—monetite, $CaHPO_4$, brushite, $CaHPO_4{\cdot}2H_2O$, and $Ca_2P_2O_7$, all with Ca/P weight ratios that equal 1.25 and mole ratios that equal 1.0—may occur in stable association with HA when the fluid phase is pure water. This phase diagram is the result of experiments at elevated temperatures and pressures to accelerate reaction times and provide crystalline products for accurate analyses. It shows the stability fields of the solid phases and fluid compositions for a range of bulk compositions. Brushite and octacalcium phosphate identified in some biomineral analyses are not found in this diagram as they occur under different temperature and pressure conditions. It should be noted that HA is the stable phosphate mineral phase throughout most of this diagram, but HA coexists with other minerals from pH 4 to 12. Interestingly, only at pH 7 is HA the only solid phase associated with a variable composition fluid. At lower pH, the solid phases that occur with HA have higher phosphate content. At a pH greater than 7, the second mineral phase in this simplified experimental system is a mineral known as portlandite, $Ca(OH)_2$. Although the mineral analyzed after low temperature extraction from tissues (discussed in Sect. 30.6) may show a Ca/P higher than that of hydroxylapatite, portlandite has not been identified in bop; pgoca; systems, but $CaCO_3$ has (Skinner 1973b, 2000a, b).

30.3.2 Composition of the Mineral in Mineralized Tissues

The bulk chemical composition of the tissues and of the included bioapatites found in them, the dominant human mineralized tissues, are listed in Table 30.2. Table 30.2A lists the bulk composition of the major components in these tissues, and Table 30.2B presents the range of major elements and the different Ca/P ratios of the bioapatites from three different sources. The measured Ca/P wt.% range, from 1.3 to 2.2, differs from the ideal or stoichiometric value for hydroxylapatite of 2.15. Because the values are both above and below the stoichiometric value, several hypotheses for such variations have been proposed.

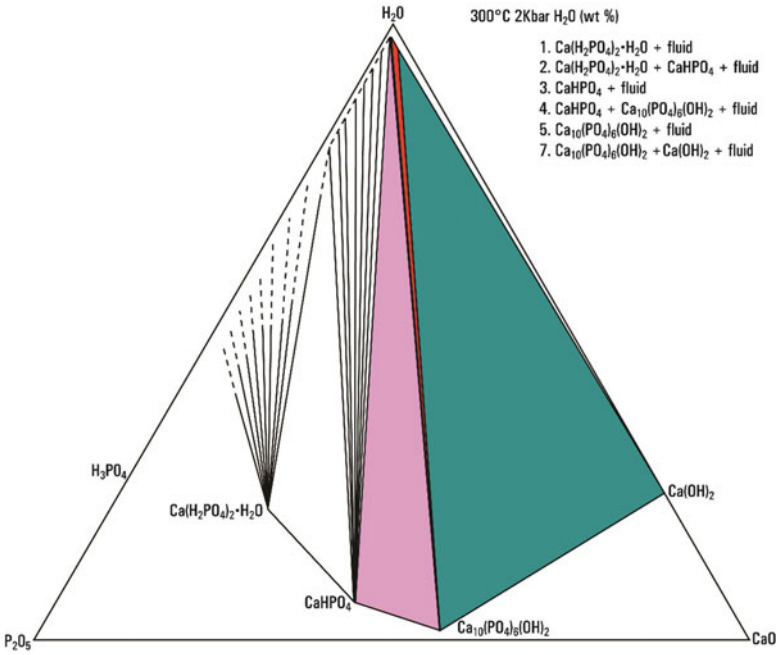


Fig. 30.2 The phase diagram of the system $CaO–P_2O_5–H_2O$ at 300¡C and 2 kbars (H_2O pressure) to illustrate the several calcium phosphate phases that may occur with hydroxylapatite under equilibrium conditions (From Skinner 1973a, b, Figure 1A)

Table 30.2 The composition of normal mineralized tissues

	Bone[a]		Dentine[a]		Enamel[a]	
	wt.%	vol.%	wt.%	vol.%	wt.%	vol.%
A. Bulk composition of bone, dentine, and enamel						
Inorganic	70	49	70	50	96	90
Water	6	13	10	20	3	8
Organic	24	38	20	30	1	2
Density (avg.) (g cm^{-3})	2.35		2.52		2.92	
B. Composition of major elements and the Ca/P ratio of the bioapatites in three tissues (wt.% on a dry, fat-free basis)[b]						
ASH	57.1		70.0		95.7	
Ca	22.5		25.9		35.9	
P	10.3		12.6		17.0	
Ca/P	2.18		2.06		2.11	
Mg	0.26		0.62		0.42	
Na	0.52		0.25		0.55	
K	0.089		0.09		0.17	
CO_2	3.5		3.19		2.35	
Cl	0.11		0.0		0.27	
F	0.054		0.02		0.01	

[a]From Driessens and Verbeeck (1990), Table 8.2, p. 107; Table 9.4, p. 165; and Table 10.5, p. 183
[b]From Zipkin (1970), Table 17, p. 72

30.3.2.1 Variations Due to Multiple Minerals

One group of investigators suggested that the biomineral matter was not a single mineral phase. Table 30.1 shows that there are a number of calcium phosphate mineral species with different, and in many cases lower, Ca/P ratios than HA. Most of these minerals may nucleate and are stable in the body fluid-tissue environment. The presence, and a variable amount, of a second mineral with lower Ca/P that might occur with HA in biodeposits has been cited as a possible reason for the compositional variability measured in tissues. The higher Ca/P ratio may reflect the substitution of another anion, e.g., CO_3^{2-}, for a portion of the PO_4^{3-} (discussed in Sect. 30.4.4).

30.3.2.2 Variations Due to Nucleation and Maturation

Another possible reason why the analyses of bioapatite may not show uniform and stoichiometric composition is that the mineral precipitate changes over time. It is well known that

phosphate ions ($H_2PO_4^{2-}$, HPO_4^{-}) are usually detected in the fluid phase at initial mineral nucleation. This led to the suggestion that nucleation was a discrete process from growth and maturation of the mineral phase and the composition of the fluid at least locally varied over time as mineralization proceeded (Roberts et al. 1992). Glimcher (1976) suggested that nucleation could be induced by charges from phosphate groups associated with matrix collagen molecules. Alternatively, because the initial solid was so poorly crystalline, Posner (1977) suggested a separate phase called "amorphous" calcium phosphate mineral, octacalcium phosphate was suggested by Brown et al. (1987), and a third suggestion of brushite was made by Neuman and Neuman (1958). All three mineral materials were considered intermediates in the mineralization process before a mature biomineral phase that more closely resembled hydroxylapatite was produced. It should be remembered that even during the investigations of the basic mineralizing system, the synthetic calcium-phosphate-water investigations that delineated the formation of HA, this stable and ubiquitous phase was most often associated with other solid phases and had incongruent solubility (Van Waser 1958; Skinner 1973a).

A variety of chemical techniques for determining the Ca/P ratio in the synthetic chemical system and early mineral deposition confirmed that initial precipitates produced by adding calcium to phosphate-rich solutions, or vice versa, was a calcium phosphate mineral with higher phosphorous (P) content than that of stoichiometric hydroxylapatite (and therefore a lower Ca/P). Although the presence of other minerals as initial phases has not been confirmed, this question remains under study (Brown et al. 1987; Roberts et al. 1992; Kim et al. 1995; Aoba et al. 1998).

30.3.2.3 Variations Due to Substitutions Within the Biomineral

Putting aside the nucleation and maturation processes and identification of the initial precipitates, biomineral has been mostly thought of unique calcium phosphate phase whose variations in Ca/P may be accommodated via substitutions and/or vacancies within the solid. The opportunity for elements and molecular species other than calcium (Ca), phosphorus (P), oxygen (O), and hydroxyl (OH) to occur in normal mineralized tissues is a most important consideration from the perspective of medical geology.

A multiplicity of ions exist in the natural environment, and it is probable that at least some of them will become part of the dynamic mineral systems found in bones, teeth or pathological deposits. Crystallographic investigations of the apatite crystal structure, presented in Sect. 30.4, enable us to discuss the ability of this species to incorporate a variety of chemical species. Each of these incorporations, known as solid solution, could alter the Ca/P ratio of the solid. For example, in geological environments, sodium (Na), lead (Pb), or strontium (Sr) may substitute for some of the calcium which means an accurate analysis for Ca/P that would be slightly lower than stoichiometric HA. Alternatively, if an anionic group, such as sulfate (SO_4), became incorporated in place of some of the phosphate (PO_4), the Ca/P would be above the stoichiometric HA value. Calculation of a Ca/P ratio, or a cation/anion ratio, from a chemical analysis will depend not only on which chemical elements and species were available when the mineral formed but also on the completeness of the analytic data used in the calculation. If only calcium and phosphate are measured, for example, no matter how accurately, calculation of a Ca/P for bioapatite may be misleading. The following section on crystal structural details of mineral apatites will allow investigators to more fully appreciate the possible uptake and incorporation of specific elements, or molecular species, into bioapatites.

30.4 The Crystal Structure of Calcium Apatites

The structure of apatites, a common group of naturally occurring minerals in many rock types throughout the world, is spoken as having hexagonal symmetry (Gaines et al. 1997). The most recent reviews of apatites, and bioapatites, can be found in a volume published by the Mineralogical Society of America and edited by Kohn, Rakovan and Hughes (2002). In this latest summary investigations utilizing very high resolution analytical techniques the crystal structure of HA shows the mineral is s in detail actually monoclinic although the older and traditional hexagonal presentation Fig. 30.3 is used to illustrate r the following unit cell specifics (Hughes et al. 1991a, b).

Figure 30.3a depicts one projection of the three-dimensional arrangement of the atoms that make up the apatite structure. This view is down the unique *c*-axis and depicts a plane perpendicular to *c*. The rhombohedral outlines are the disposition of two of the three a axes (directions at 120° apart) and the atoms are precisely placed conforming to the crystallographic characteristics of the repeating unit, or unit cell, of the apatite structure.

To further illustrate the distribution and importance of the tetrahedral orthophosphate groups (PO_4^{3-}) that form the backbone of the apatite structure, Fig. 30.3b presents another view of the structure. It is a projection 90° to the *c*-axis along one of the *a* axes. The yellow phosphorus atom is in tetrahedral coordination surrounded by two white oxygen atoms in the same plane as the phosphorus and two light purple oxygen atoms, one above the plane, the other below it.

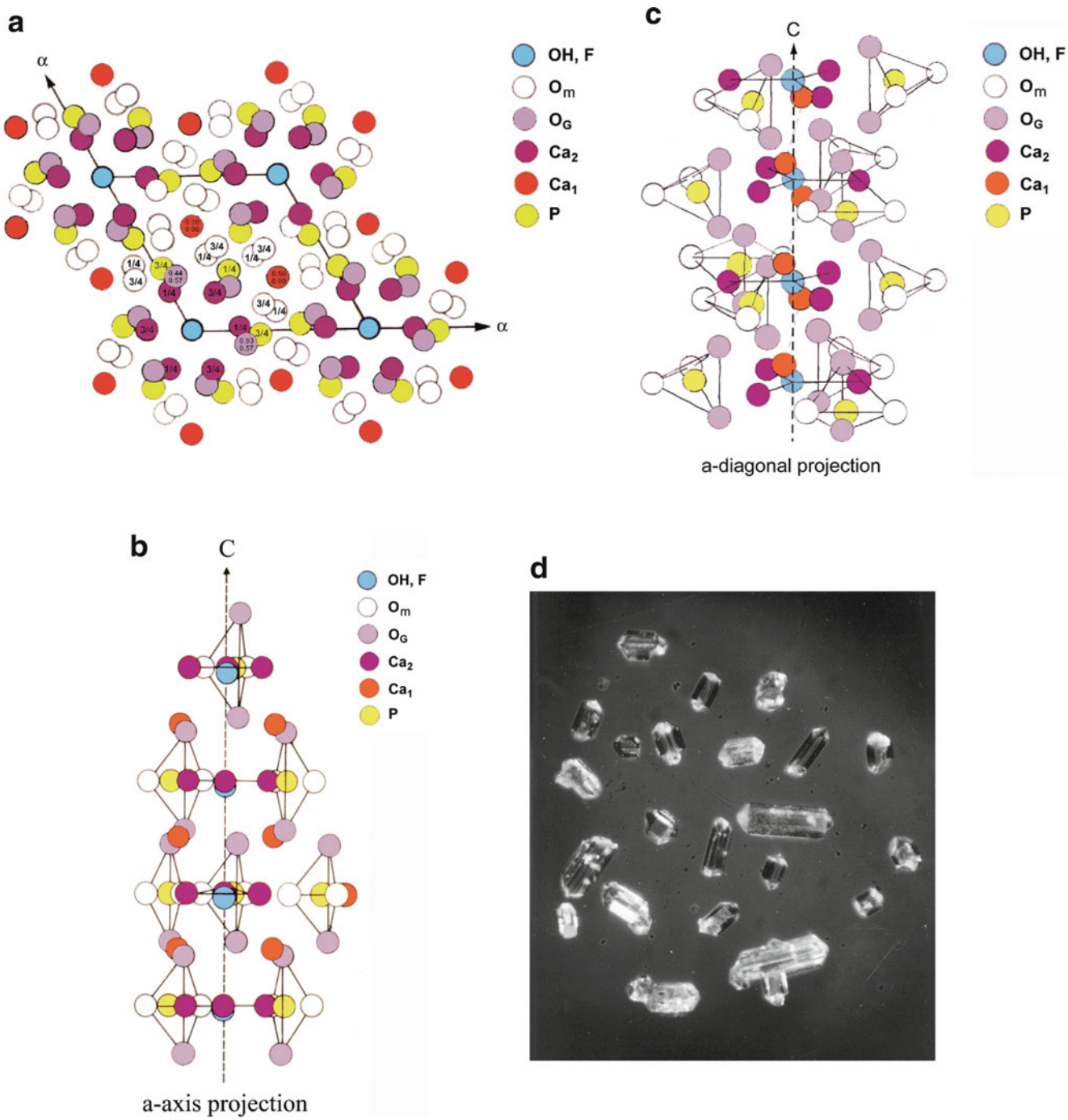


Fig. 30.3 The crystal structure of hydroxylapatite, ideal formula $Ca_5(PO_4)_3(OH)$. (**a**) Projection down the *c*-axis, showing the distribution of all the atoms in the unit cell. (**b**) Projection down the *a* axis of the unit cell, note the tetrahedral orthophosphate groups. (**c**) Projection down the *a* diagonal, a different view of the three-dimensional arrangement of atoms in the unit cell. (**d**) Crystals of hydroxylapatite (synthetic), note the hexagonal prismatic morphology

Phosphate anionic groups coordinate with calcium cations to produce a charge neutral solid when a singly charged species, OH^{1-} or F^{1-}, the blue atom site, is added. Figure 30.3c is the view down the diagonal between two *a* axes. Note the distribution of the calcium atoms and the channel ways where the OH^{1-} or F^{1-} occur parallel to the *c*-axis direction. The prismatic hexagon morphology typical of many apatite mineral samples shows up in Fig. 30.3d. It is a selection of single crystals of HA produced during investigations of the simplified chemical system (Skinner 1973a, b). The crystals

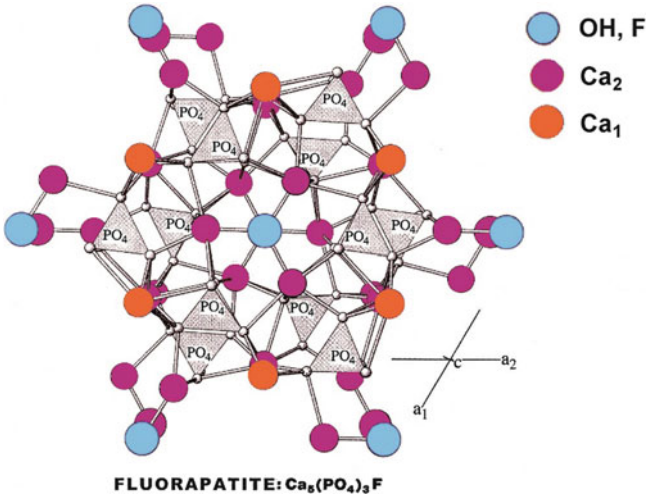


Fig. 30.4 Ffluorapatite crystal structure, note the cation distribution in two sites, Ca_1 and Ca_2,> View down unique *c*-axis, similar to Fig. 30.3a but with F in place of OH

are synthetic hydroxylapatites and are up to several millimeters in length.

There are two distinct sites for cations and they are discretely colored orange and raspberry red. One is connected to the tetrahedral phosphate (PO_4) backbone oxygens and the atoms in channelways (OH, F, or Cl) that parallel the *c*-axis, and the other is related to the trigonal a axes in the center of the cell. The two calcium sites are designated Ca1 and Ca2 in Fig. 30.4. When the channels are occupied by OH^{1-} the mineral is known as hydroxylapatite, and when occupied by F^{1-} the name is fluorapatite, which is another member of the calcium apatite mineral group.

An amazing number of different composition minerals, each a separately named mineral species, may form with the apatite structure type. These minerals show mixed chemistries with very minor variations in structural detail that are related to the amount and exact placement of the particular elements. Because of the importance of the chemical variations to our medical purview and the fact that many of these naturally occurring minerals have been studied in great detail, the sections below present examples to show what elements find their way into bioapatites and where they may be located. Bioapatite composition mirrors elemental bioavailability.

30.4.1 Substitutions in the Phosphate Backbone

Though the most common minerals are probably the phosphate apatites, there are arsenate or vanadate minerals that form with virtually identical apatite-like crystalline structures. The AsO_4^{3-4-}, and VO_4^{3-} anions have a charge and size very similar to the PO_4^{3-} group. Naturally occurring arsenate or vanadate apatites usually contain lead (Pb) rather than calcium. Lead phosphate apatite, the mineral pyromorphite with ideal formula $Pb_5(PO_4)_3Cl$, has all cation sites occupied by lead, but chlorine, Cl^{1-}, takes the place of the OH in the channelways (see Sect. 30.4.3).

Other tetrahedral anionic groups, sulfate SO_4^{2-} and silicate SiO_4^{4-}, although of different charges, may also form minerals with apatite structure type, and small amounts of these groups as well as arsenate and vanadate may be found in the phosphate apatite minerals. Britholite-Y, a naturally occurring mineral, is one of a series of phosphate apatites that contain silicon substituting for phosphorus, but there is a second cation in addition to calcium, yttrium (Y). Both OH and fluorine (F) are present in the channelways.

30.4.2 Substitutions in the Cation Sites

There are two different cation sites, or distinct lattice locations, known as Ca1 and Ca2 in hydroxylapatite (Fig. 30.4). The site designated Ca1 bonds to nine oxygen atoms of the tetrahedral phosphate groups while Ca2 bonds to six phosphate oxygens and one OH ion. Many other cations, especially those with double positive charge similar to Ca^{2+} such as Sr^{2+}, can be accommodated in these sites. Table 30.3 lists the elements by charge and by size that could substitute in apatites.

Lead apatite, the mineral pyromorphite, has an apatite structure but lead is a large ion relative to calcium. However, small amounts of lead occur in the predominantly calcium bioapatites. Lead isotope ^{207}Pb has been used to discriminate between the possible sources. Lead from leaded gasoline, or uranium, for example, might be taken up in human tissues. Lead may also occur in high concentrations indoors. It might be found as a component of dust distributed about the living areas with the possible source local soils tracked inside a residence or be from lead-based paint. Some from either source could be ingested and become sequestered in bones or teeth. These exposures make lead a "silent" hazard, a potential danger especially for young children who constantly put their fingers in their mouths. The term silent is used because Pb is not visible, and detection and amount determined in the environment or in the body is impossible without specialized analytical methods and tools. Therefore biological uptake and sequestration of lead in bones, which may be cumulative, is not monitored. Lead exposure goes unrecognized unless some distinctive disease or signal, as described in a study on schoolchildren (Mielke 2003), is recognized, evaluated, and related to lead levels.

Strontium apatite, ideal formula $(Sr, Ca)_5(PO_4)_3(OH)$, has been designated a separate mineral species and a

Table 30.3 Ions that can substitute in calcium phosphate apatites, their charge and ionic radii depending on coordination number, i.e. VII, IX. . .

Ion	Ionic radius
Ca^{2+}	1.06[a,b](VII) 1.18[a](IX)
Cd^{2+}	1.14[a](VII)
Mg^{2+}	0.79[a](VII) 0.89[b](VIII)
Sr^{2+}	1.21[a](VII) 1.31[b](IX)
Ba^{2+}	1.38[b](VII) 1.47[b](IX)
Mn^{2+}	0.90[b](VII) 0.96[b](VIII)
Na^{1+}	1.12[b](VII) 1.24[b](IX)
K^{1+}	1.46[b](VII) 1.55[b](IX)
Pb^{2+}	1.23[b](VII) 1.35[b](IX)
P^{5+}	0.17[a,b](IV)
As^{5+}	0.335[a](IV)
Si^{4+}	0.26[a,b](IV)
V^{5+}	0.54[a](VI)
S^{6+}	0.12[b](IV) 0.29[b](VI)
Sb^{5+}	0.61[b](VI)
Al^{3+}	0.39[a](IV)
U^{3+}	0.98[a](VII)
Ce^{3+}	1.07[a](VII) 1.146[b](IX)

[a]From Shannon and Prewitt (1969)
[b]From Shannon (1976)

member of one of the subsets within the apatite group, because the amount of Sr is greater than 50% of the total cations present. Studies on naturally occurring rare-earth-element-substituted apatites show coordinated substitution: when rare earth elements (REE) with a charge of 3^+ are incorporated in the structure, there may be a substitution of Si^{4+} for part of the phosphorus. With both substitutions, a charge balance comparable to the original association of Ca^{2+} and P^{5+} in the calcium apatite structure can be maintained. Another coordinated substitution is when an REE^{3+} is incorporated along with Na^{1+}. Charge balance may be achieved because the cations are distributed at both sites in calcium apatites (Hughes et al. 1991b).

Some elements prefer one cation site over the other. A study of REE-containing apatite minerals by Hughes et al. (1991b) demonstrated that some of the REE preferred the Ca^2 site while others preferred the Ca^1 site, provided charge balance was maintained by substitutions at the anionic sites or with additional and differently charged cations. Site designation for particular elements can be determined using high-resolution X-ray diffraction analyses, paramagnetic resonance, thermo luminescence, or infrared spectroscopy (Suich et al. 1985).

Apatite samples can show slightly different compositions in spite of coming from similar sources or sites. Although it is possible with modern techniques to show elements at specific sites in the crystal structure of geological samples, the tiny crystallites of bioapatite are too small for such detailed investigations. It is worthwhile to reiterate that bioapatite composition in one bone may not be identical to another bioapatite forming elsewhere at the same moment or at different times. The fluid-cell-matrix-mineral system composition is unlikely to be constant from one moment to another, much less from year to year as the human ages, resides in different geographic localities, and ingests water or food from different sources.

There is another set of concerns that relate directly to cation substitution in bioapatites: bone seeking α-emitting radionuclides and ionizing radiation exposures of humans. During atomic bomb tests a half century ago in the southwest United States there was a scare related to fallout of radioactive nuclides, especially ^{90}Sr. The anxiety was based on the similarities in behavior of calcium and strontium, the half-life of the nuclide (28 years), the prevailing wind direction toward the more heavily populated east, and the fact that American dietary calcium came mostly from milk products with the largest consumers being children who were actively putting down new bone. The worldwide average of ^{90}Sr was shown to be about 0.12 microcuries per gram of calcium in man or 1/10,000 of the acceptable permissible level at that time. This suggested that the atomic bomb circulating ^{90}Sr was not a global hazard. A remarkable study on baby teeth in mid-western communities of the United States compared the strontium concentrations with adult bones the late 1950s. These investigations showed that bioapatites discriminated against strontium during formation and concluded that only those individuals who obtained their total food supply from restricted areas (with low calcium in the rocks, soils, or waters) were at risk (Eckelmann et al. 1957). The furor over the nuclide hazard eventually collapsed when it was realized that strontium bioavailability was overshadowed by calcium and no one, especially children, was likely to be at risk in the United States or abroad (Eckelman et al. 1957; Fowler 1960).

Massive doses of radioactive elements from nuclear explosions, such as the Chernobyl disaster, are locally extremely hazardous as they spread the radionuclides in the soils and the plants that animals and humans ingest over time and the persistence of the nuclides in the mineralized tissues is a worry However, the potential for incorporation of several potentially harmful radioactive materials in mineralized tissues at a hazardous exposure level has yet to be documented in spite of extensive surveys (Fabrikant 1988; Elliott 2002).

30.4.3 Substitutions in the Hydroxyl Site

The hydroxyl (OH) site in the calcium apatite crystal structure can be fully occupied by fluorine or chlorine. As end members in that chemical system, they form the independent minerals fluorapatite and chlorapatite. These halogen species are the predominant forms of apatite found in sedimentary, metamorphic, and igneous rocks (Hughes et al. 1989). Bromine (Br) or iodine (I) can be incorporated in the apatite structure type, but mineral species in which the halogen site is filled entirely by bromine or iodine have not been found naturally.

Bioapatites originally precipitate and remain mostly as hydroxylapatite in human bones and teeth because of the predominance of aqueous fluid and OH concentration relative to the halogens in the human body. When higher amounts of fluorine become available the element can become incorporated substituting for part, at least, of the OH. For example, 1 mg L^{-1} fluorine added to drinking water, an amount that has become standard for many of the reservoirs that supply water to populations across the United States, appears to reduce caries and may lead to a reduction in the incidence of osteoporosis (Watts 1999; Kohn and Cerling 2002). On the other hand, the regular ingestion of greater than 100 mg L^{-1} of fluorine over a long period of time by humans leads to disease. Such high amounts of fluorine in local waters and agricultural products grown in soils irrigated with high-fluorine-containing water, or through industrial exposure, may result in fluorosis (Vischer 1970; Finkelman et al. 1999). Whether the mineral matter in the bone and tooth tissues of such exposed human populations is partially fluorapatite, i.e., a mixture of the two separate apatite species, or whether each apatite crystallite has both fluorine and hydroxyl in its channel sites is unknown. Fluoride has been used to treat osteoporosis (see Sect. 30.6.2.1).

Chlorapatite, the calcium phosphate apatite mineral in which all the channelways are filled with chlorine, has not been identified in mineralized tissues in spite of the high concentration of chlorine (Cl) in body fluids, but small amounts of chlorine can be detected on analyses of bioapatites.

30.4.4 Carbonate (CO_3^{2-}) in Apatites

One additional chemical constituent that is often detected in apatite analyses becomes important when discussing bioapatites, and that is carbonate. Two carbonate-containing calcium apatite minerals, dahllite and francolite (Table 30.1), have been described from phosphorites, fine-grained sedimentary deposits mined for fertilizer on many continents. These phosphate minerals are associated with the common calcium carbonate minerals calcite and aragonite (Gaines et al. 1997). Neither calcite nor aragonite has been identified in bone tissues, but because many bioapatites show higher than stoichiometric Ca/P ratio and CO_3^{2-} on analysis, the suggestion is that CO_3^{2-} substitutes either for PO_4^{3-} or for OH^{1-} in the apatite. The carbonate ion CO_3^{2-} has a different charge and size than the dominant phosphate groups (PO_4^{3-}). It is a planar trigonal ion with a diameter of 0.24 nm and does not easily fit in the crystal structure. The amount and disposition of CO_3^{2-} within the apatite crystal structure is, and has been, a topic of great interest for some time (McConnell 1973; Skinner 1989; Elliott 1984, 1994, 2002).

If carbonate, CO_3^{2-}, is present in the crystal lattice other ions, such as triply positively charged cations, might take the place of calcium so that local charge disruption could be balanced by coupled substitution or by vacancies in the lattice. The inclusion of CO_3 could compromise the ideal architecture of the apatite backbone and the channelways. Such destabilization may account for the very fine-grained nature of bioapatites. And aid in the mineral dissolution by osteoclastic activity (Skinner 2012).

The association of CO_2 with bioapatite is not particularly surprising. The molecular species CO_2 or bicarbonate, HCO_3^{1-}, are produced along with many others during cell metabolism and could adsorb on the high surface area of the tiny crystallites. Early deposition of mineral mechanisms probably takes place close to ph7 but if slightly acidic because of presence of orthophosphate species that may aid the nucleation of bioapatite whereas other ions in the fluid, such as bicarbonate, may be inhibitory (Glimcher 1998).

The carbonate ion and its distribution is not of major concern for this medical geology purview except that its presence makes us aware of the necessity to consider both the physical and chemical aspects of the mineralizing system. Carbonate probably aids incorporation of other elements into the lattice. Bioapatite precipitates are aggregates of crystallites, which means that the mineral mosaic of many crystallites can each present a slightly different composition and size. The lower crystallinity, and variable composition, of carbonate-containing-bioapatites may be an irritation preventing precise designation of the mineral phase, but it has an advantage for the biological system since many different proteins have other element cofactors to activation. The high surface area of the bioaptite crystallites facilitates their dissolution as required for the dynamic bone mineral formation-resorption system. Nature has utilized a solid phase that fulfills several functional roles required for bone (Skinner 2000b). Bioapatites record exposures of living creatures to the environment and particularly the bioavailability of elemental species in our diets Kohn and Cerling (2002) present insights and discussion of stable isotopes in bioapatites.

This brief summary does not do justice to investigations with a variety of techniques which include electron and X-ray diffraction analysis, infrared, polarized infrared and Raman spectroscopy, solid state carbon-13 nuclear magnetic resonance spectroscopy, and most recently atomic force microscopy. These have been and are used to detect and quantitate the amount of carbonate within the crystal lattice of bioaptite in the tissues whether as part of the composite with organic molecules apatite or as an adsorbed species.

From the above selected examples and Table 30.3, the very wide range of elements and molecular species that can be accommodated within the apatite crystal structure is summarized. Table 30.4 lists the levels of elements essential to proper body function. The remarkable match for many elements and chemicals entering the human body show that many become associated with, or become part of, the apatitic mineral matter.

30.5 Analysis of Apatitic Biominerals

To ascertain the ranges of included elements and species in bioapatites, the mineral matter must be extracted and concentrated. The techniques devised for separating mineral from the associated organic and cellular materials can, in many cases, further complicate the assays. The other option is to analyze the tissues keeping both the mineral and organic fractions associated. Either way there are special techniques required to prepare the sample for analysis. Preparation of the mineral phase and the main methods of analysis, for example diffraction, will be presented followed by the sample preparations necessary for examining whole tissues. Histology is the general name for the host of techniques using optical and/or scanning electron microscopic (SEM) analyses now utilizing synchrotron as well as electron sources to investigate thin sections of tissues.

30.5.1 Sample Preparation: Mineral

The opportunity to examine the mineral separately from surrounding organic materials, whether examining pathological aggregates from arteries or from normal bone tissues, is a non-trivial undertaking (Kim et al. 1995). Bioapatites have individual mineral grains of the order of 1 × 2 × 25 nm perhaps up to 50 nm and are loosely associated one with another as porous aggregates and random crystallite distribution or alternatively on and in a fibrous protein matrix. The latter is characteristic of bone tissues where the crystallites often align parallel to the length of the collagen molecules (Skinner 1987). Aware that there is a range of composition of bioapatites and mindful of their very small grain size, care with any extraction must be exercised.

Table 30.4 The essential elements, and their recommended daily intake (RDI)

Element	RDI
Boron[a]	(1.7–7.0 mg)
Bromine	0.3–7.0 mg
Calcium[a]	0.8–1.3 g
Cesium[a]	0.1–17.5 mcg
Chromium[a]	130 mcg
Cobalt	15–32 mcg
Copper[a]	1–2 mg
Fluorine[a]	1.0 mgt
Iodine	70–150 mcg
Iron[a]	10–18 mg
Lithium	730 mcg
Magnesium[a]	200–450 mg
Manganese[a]	3.5 mg
Molybdenum	160 mcg
Nickel[a]	(35–700)
Phosphorus[a]	0.8–1.3 G
Potassium[a]	3,500 mg
Selenium	70 mcg
Silicon[a]	(21–46 mg)
Tin[a]	0.13–12.69 mcg
Vanadium	(12.4–30.0 mcg)
Zinc[a]	8–15 mg

Note: Additional elements detected include the following non-essential elements: Ag, Af, Cd, Pb, and Sp
From the Food and Nutrition Board, National Research Council (1989), the Federal Register #2,206, and for elements in bone Bronner (1996) and Skinner et al. (1972)
[a]Indicates those detected in bone tissues and ( ) indicates non-essential elements

In addition it is necessary to record the specific tissue and site in the organ or exactly where a pathological deposit is located. The age of the individual and the date of sampling are also critical because different cell systems and tissue textures are encountered in every bone or soft tissue (Weiner and Wagner 1998).

In all normal bone the intimate association of mineral with matrix proteins and other proteins shows variations dependent on source (Miller 1973) Table 30.2 gives average amounts, but the mineral concentration as well as distribution also varies at submicron levels (Rey et al. 1996). The chemical variations detailed above reinforce the possibility that the mineral itself may also vary with growth and maturation, especially in bone where the entire structures, e.g., haversian bone, are constantly resorbed and re-deposited in a different pattern over time (Skinner 1987). The chemical analysis of mineral that may attract and adsorb transient ions from the surrounding fluids is only an indication of the compositional range of the tissue at a specific moment in place and time.

Analyses of mineralized tissues are not only compromised by the poor crystallinity of the mineral but

by the presence of non-mineral components. A sample of enamel with over 96 wt.% mineral, less than one percent protein (enamelin), and the most highly mineralized normal tissue in the human body, is the preferred choice for biomineral analyses and comparisons between individuals and populations. Enamel is also examined because it has a very restricted period of formation, and because no cells remain at maturity the final tissue is not reworked. Dentine, the mineralized tissue adjacent to enamel and the major tissue in the tooth, is maximally 75% mineral when fully mature, with collagen about 20% of the total, plus some small molecular protein species, and fluid. Bone tissues, especially in the first stages of formation, and spicular, or trabecular bone, found adjacent to the marrow cavity (Fig. 30.1b), may contain less than 50% mineral per unit area. To accurately ascertain the mineral composition and structure, it is important to extract the mineral portion from these organic moieties.

The usual way to separate the mineral fraction has been to immerse the entire sample in sodium hypochlorite or bleach. Most of the organic matter will eventually dissolve, and the time it takes depends on the size and porosity of a particular sample relative to the amount of bleach. Smaller sample size allows for a more rapid dissociation of the mineral from the intimately associated organic molecules. To concentrate the mineral fraction from the heavily mineralized cortex of a long bone will require a long soak and much decanting and re-suspending of the tissue sample in the bleach. This procedure could alter the amount of mineral, especially as smaller crystallites are likely to be more soluble.

More exotic systems of extracting the organic moieties from the mineral phase using chemical methods have been suggested. Refluxing with ethylene diamine (Skinner et al. 1972), a lengthy and possibly dangerous procedure because the solution pH is 12 or greater. An interesting observation on careful chemical extractions is that the morphology of the organ or sample, whether bone or tooth, will likely remain after virtually all of the organics (>95%) have been removed. This is an illustration of the remarkable permeation of mineral with the biological macromolecules and how difficult it is to completely extract all of the intimately associated organic phases. Pathologic spherulitic apatitic samples also have ultra small size, not only of the aggregates, but of the crystallites within them.

Several other methods of extraction have been employed. One technique, called "ashing" subjects tissues to elevated temperatures, above 500°C and often to 1,000°C, for at least an hour. Researchers who wish to assure complete disappearance of the bioorganic portions advocate up to 10 or more hours at elevated temperatures. The extended high temperature treatment certainly eliminates any organic components, but it also re-crystallizes the mineral phase. An X-ray diffraction analysis of the ashed sample (see Sect. 30.5.2) provides a sharp diffraction pattern of hydroxylapatite that is comparable to well-crystallized geologically obtained mineral. Occasionally, two mineral phases are detected with the identity of the second phase dependent on the bulk composition of the sample, the temperature, and duration of ashing. For example, above about 200°C adsorbed water and any carbonate will be lost and other calcium phosphates, $CaHPO_4$ or $Ca_2P_2O_7$, appear (Table 30.1). This indicates that the bulk Ca/P composition was lower than stoichiometric HA. At greater than 1,000°C, tricalcium phosphate, $Ca_3(PO_4)_2$, may take the place of the pyrophosphate. The relative amounts of the two species at a known temperature can be used to calculate a Ca/P for the sample. If the Ca/P is greater than 2.15, tetracalcium phosphate $Ca_4(PO_4)_2O$ may form. The appearance of different mineral phases reflects the experimental conditions that the mineralized tissue sample was subjected to, and clearly is markedly different from any of the second phase in the low-temperature-produced and extracted tissues. An alternative, low-temperature ashing using activated oxygen, has also been employed to remove the organic constituents and X-ray or electron diffraction techniques on the extracted materials is the way to determine the crystal chemical characteristics of the mineral phase (Kim et al. 1995).

30.5.2 X-ray and Electron Diffraction

Unambiguous identification of most mineral materials utilizes diffraction techniques that can easily determine the species based on the unique crystal structural characteristics of the compound (Klug and Alexander 1954). The powder diffraction method either X-ray or electron diffraction may be employed, and the choice is usually dependent on instrument availability and the specifics of the sample. Vast databases have accumulated over the past hundred years since the techniques were elucidated. The International Union Committee of Diffraction has a compendium of X-ray diffraction data on crystalline compounds, both organic and inorganic, which contains a subset for natural and synthetic mineral materials including the apatites.

The tiny crystallite size and variable composition of bioapatites are fully expressed in the diffraction analysis of the extracted materials. Well-crystallized (without vacancies and >0.5 μm in average size) hydroxylapatite geological samples give many discrete diffraction maxima from which one easily calculates unit cell parameters. Comparing a mineral apatite X-ray powder diffraction pattern (Fig. 30.5) with bioapatite, e.g., cortical bone mineral, show a few and broad maxima, poorly crystalline pattern that makes it difficult to determine any compositional details other than to say the sample 'may be apatitic' (Skinner 1968). Enamel is perhaps the only biomaterial that provides

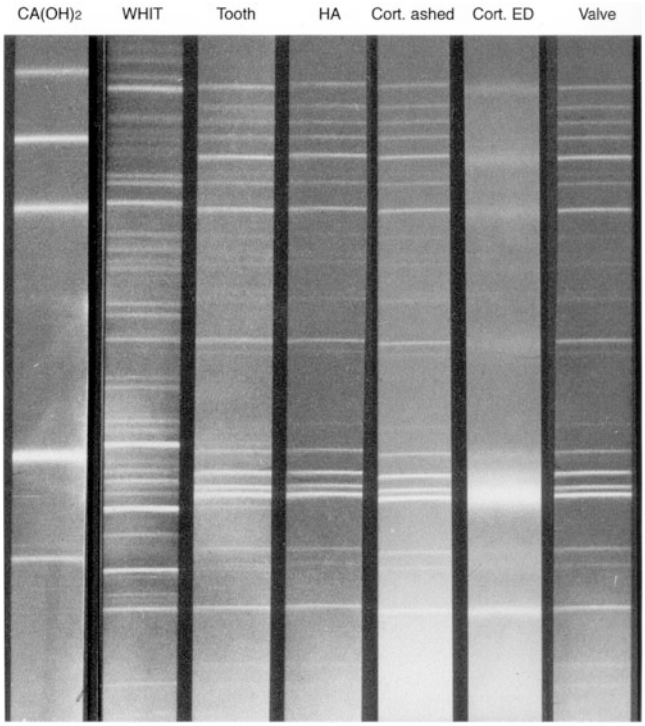


Fig. 30.5 Comparison of X-ray powder diffraction patterns of mineral and tissue bioapatites: $Ca(OH)_2$, portlandite; *WHIT* whitlockite, Tooth, bioapatite; *HA* hydroxylapatite, synthetic hydrothermal sample (From Skinner 1973a, b); Cort. ashed and Cort. ED, cortical bone ashed and ethylene diamine extracted cortical bone; Valve, human aortic valve pathological bioapatite deposit

sufficient diffraction detail to calculate lattice parameters and Table 30.5 presents the results of calculations from powder diffraction data of several different apatites. Electron diffraction does not necessarily afford more precise results over X-ray diffraction as the level of crystallinity is the important criterion for producing diffraction. The advantage of electron diffraction is that the beam may be more finely focused, and a small mineral crystal aggregate within a thin section might be separately examined rather than extracting the mineral to obtain a pure mineral sample for powder diffraction.

30.5.3 Sample Preparation: Mineralized Tissues

To examine the mineral phase in its biological surroundings requires different sample preparation. The techniques must minimize any alteration of the tiny crystallites while maintaining the organic and cellular framework in which the mineral matter is distributed. Thin sections are prepared for examination with optical and electron microscopic techniques so that morphological relationships typical of the different tissue types and chemical observations on the mineral phase can be assayed (Sect. 30.5.4).

The techniques to prepare biological tissues as thin sections parallel the methods employed in petrology (the study of rocks) (Blatt and Tracy 1996). The mineralized tissue sample is embedded to obtain a plano-parallel thin

Table 30.5 Calculated unit cell (lattice) parameters for synthetic apatites, and bioapatites

	a-Axis	*c*-Axis	Ref.
Hydroxylapatite (Synthetic)			
315¡C, 2 kbars H_2O pressure			
CaO/P_2O_5 1.61	9.421	6.882	Skinner (1968)
CaO/P_2O_5 1.12	9.416	6.883	Skinner (1968)
600¡C, 2 kbars H_2O pressure			
CaO/P_2O_5 1.61	9.4145	6.880	Skinner (1968)
CaO/P_2O_5 1.12	9.4224	6.8819	Skinner (1968)
100¡C Precipitation	9.422	6.883	Bell and Mika (1979)
(Reitveld calc.)	9.4174	6.8855	Young and Holcomb (1982)
Mineral	9.418	6.875	Gaines et al. (1997), p. 856
Carbonate apatites (synthetic)[a]			
Low temperature (<100¡C precipitates, air-dried)			
Direct (adding calcium acetate into a solution of ammonium carbonate + phosphate) 13.8% CO_3	9.373	6.897	Laborite et al. (1973)
Inverse (adding ammonium solution into calcium acetate) 10.4% CO_3	9.354	6.897	Laborite et al. (1973)
Direct using sodium carbonate	9.440	6.880	Legers et al. (1968)
Direct using sodium carbonate and fluoride 22.1% CO_3	9.268	6.924	Legers et al. (1968)
High temperatures 1.455 CO_3, the maximum determined	9.367	6.934	Rey et al. (1996)
Bone Ashed cortical	9.419	6.886	Skinner et al. (1972)
Enamel Human	9.421	6.881	Carlstrom (1955)

See Elliott (1994, pp. 234–248) for details on precipitation mechanisms and discussion of IR, X-ray diffraction analyses, and possible locations of the carbonate ion in the apatite lattice

[a]The wide variations in the values may be partially attributed to differences in composition of the starting materials and to the methods of preparation, e.g., direct versus inverse, under low temperatures for carbonate apatite synthesis

section, from 50 to less than 10 μm thick. The thickness will depend on the information desired and the method of analysis. Embedding preferentially employs plastics, rather than paraffin, the typical media for soft tissue sections examined in pathology laboratories (Malluche et al. 1982). Plastic is required because the juxtaposition of tiny crystals of hard mineral with the soft organic matter and cells minimizes differential hardness of the medium put to sectioning.

After obtaining a fresh tissue sample the first act is to soak the pieces in alcohol solution about ten times the sample volume for a few hours or overnight refreshing the solution several times. Occasionally formaldehyde is used but formalin has to be buffered or the solution will be acidic and at least some mineral crystallites may be lost or dissolve during immersion. The alcohol soak effectively lowers the water and fat content in the tissue, stops further biological degradation, and stabilizes the organic components. The second stage is to embed the whole sample with not too viscous epoxy materials such as methylmethacrylate, or other commercially available plastics such as Spurr™ or Epon™. The viscosity must be appropriate to facilitate penetration throughout the sample, and probably will require a vacuum to maximize efficiency. Most mineralized tissue laboratories will have an automated embedding system that takes the tissue through the extraction and embedding procedures and applies vacuum and heat and a "hardener" to ensure a fully homogeneous block of embedded sample. Any holes, effectively pockets of air that might incorporate dust or foreign materials, must be avoided especially if the analysis will use SEM coupled with energy dispersive analysis (SEM/EDXA) to measure the elemental composition of the mineral. Once embedded the whole tissue piece is ready for sectioning and polishing.

Sectioning requires a sharp knife, usually diamond or carborundum blades, to cut the plastic embedded tissues without tearing the crystallites from their organic matrix. Once cut the thin section is mounted on a glass or plastic slide for microscopic study, usually without the addition of a cover slip. Grinding to assure plano-parallel surface for high-resolution electron microprobe elemental analysis may be essential, but it almost always results in smearing and disrupting the crystallites, so this level of the preparation procedure again requires care. It is wise to constantly view the section under a microscope with at least 40× magnification. Any surface roughness will cause interference in elemental analyses, although background corrections can be applied to the raw data (Reed 1993).

Every laboratory that performs mineralized tissue section analyses, usually the pathology departments of hospitals, has their own sterile procedures. To prepare uniformly parallel, non-artifact-containing hard tissue thin sections is an art, not a science. Once made, sections may be stained to identify osteoid, cell walls, nuclei, etc., or the calcified portions for specialized investigations, but such treatments would compromise analysis of the mineral phase. Careful, sensitive attention to the preparation of any section is necessary for accurate documentation of the tissue components, their structures, and elemental components.

30.5.4 Histomorphometry

To study the variations in the textures found in all mineralized tissues, bone, teeth, or pathological materials, optical investigations at different magnifications are usually required. Optical light microscopy using transmitted light at approximately 40× magnification is sufficient to determine areas of mineralization in a tissue section. The mineral matter will be non-translucent and easily distinguished from organic components. By employing a polarizing light microscope, with the crossed polarizers in place, the mineral portions will show variations in birefringence on rotation of the stage, which is the characteristic used to identify any crystalline mineral (Nesse 2004). The mineral calcite, for example, can be easily differentiated from hydroxylapatite because calcite has a much higher birefringence. However, to thoroughly document the tissue components and study the remodeling of bone tissue, combinations of SEM and optical and electron microscopy are usually employed.

Some of the distinct morphologies observed in normal and pathological mineralized tissues are next. Figure 30.6a is tissue section cut longitudinally through the upper end of a long bone. Dense mineral matter and cortical bone surrounds the shaft and the open marrow cavity. It also outlines the trabecular tissue in what appears to be the porous head of the femur. The arcuate patterning of trabeculae throughout the head is an expression of mineral deposition conforming to stress in this organ; the bone tissue distribution responds to mechanical strain (Albright and Skinner 1987). In Fig. 30.6b two tissue sections through vertebrae of different ages illustrate changes in trabecular thickness and patterning. The vertical struts thicken with age as the tissues respond to the effects of gravity from our vertical posture.

Figure 30.7a is an x-ray transmission microradiograph of cortical bone tissue from a dog tibia. It illustrates variations in mineral density in the several osteons (the circular patterns), typical of haversian bone. Variation in the levels of gray in this section reflect different amounts of mineral per unit area and detected at higher magnification because of the mineral impeding the transmission of x-ray energy. Figure 30.7b is another view of the same area but illuminated with UV light. Three of the osteons show two circular dark lines around the central vascular opening, the pattern resembling an archery target. The lines are due to the incorporation of the antibiotic tetracycline as the mineral deposits and the molecule glows under UV light. The two lines mark two

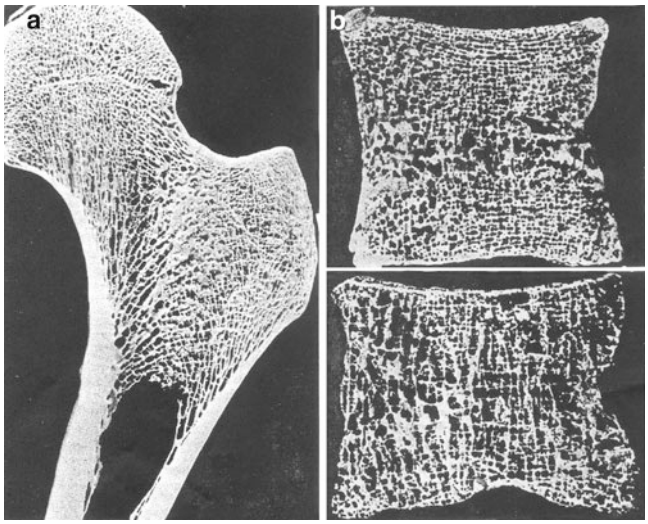


Fig. 30.6 Transmission X-ray photos of excised samples of bone. (**a**) Longitudinal section through the upper femur of normal 29-year-old male (From Albright and Skinner 1987, Figure 5–1). Note the arcuate pattern of the trabeculae. (**b**) Mid sagittal sections of first lumbar vertebra of a 20-year-old female (*above*) and 50 year old (*below*) (From Arnold et al. 1966, Figure 1.16). The trabeculae in the older bone show thickening and ÒreinforcementÓ accentuating their vertical orientation in the skeleton

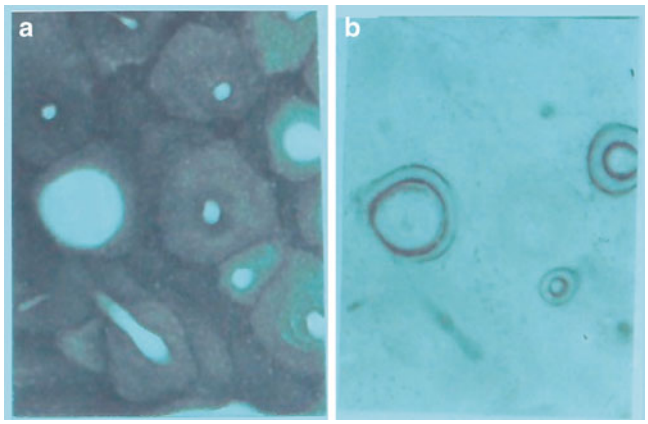


Fig. 30.7 Ground section of the cortical portion of a rib from a dog given two doses of the antibiotic tetracycline. The tetracycline is incorporated at the time of deposition of bone tissue (Skinner and Nalbandian 1975, Figure 3). Magnification 175×. (**a**) In transmitted light, the darker the color the more mineral present, an expression of decreased transmission of light. Note variations in bone mineral density between the several haversian systems in the section. Some contain more mineral than others. (**b**) Same area in ultraviolet light (tetracycline glows in UV light) which shows two fluorescent rings (*concentric dark circles*) in the haversian systems marking the times when two doses of tetracycline were added to the diet as bone tissues were developing

separate doses of tetracycline allowing the rate of mineral deposition, and the growth of osteons in the cortex, to be determined (Skinner and Nalbandian 1975).

Figure 30.8a is an image that shows spherulitic (pathologic) calcium phosphate deposits in breast tissue. The higher resolution (×1,200) and use of back scattered electron imaging (one of the modes available with SEM) shows these tiny deposits. Figure 30.8b is an electron micrograph at still higher magnification (×25,000) of the dentine-enamel junction illustrating the different size, shape, and aggregation of hydroxylapatite crystallites in these two tissues. At magnifications greater than × 200,000, transmission electron micrographs have shown hexagonal outlines of the early-formed enamel crystals.

Weiner and Wagner (1998) recently reviewed the multiple levels of structural organization seen with histological examination and some of the physical contributions of the mineral crystallites. Bone organs require strong and flexible tissues, so they do not buckle and break when subjected to torsion, tension, compression, and instant and consistent responses to applied stresses for proper organ function. Although histology demonstrates the physical textures and mineral distribution during growth, development, and aging, the incorporation of chemicals may well alter the reactions of the mineral, and hence the mineralized tissues (Bronner 1996; Skinner et al. 2004). A brief discussion of osteoporosis and outlining some treatments for this disease will illustrate how our present understanding of bone physical properties and bone mineralogy are benefiting medical care.

30.6 Osteoporosis

Osteoporosis is a metabolic bone disorder characterized by reduction in the volume of bony tissue per unit volume of bone (Fig. 30.9). The disease has been a topic of investigation for over a hundred years (Arnold et al 1966; Barzel 1970; Riggs and Melton 1995). The tissue degeneration over time usually leads to fragility of the bone organs and may cause an osteoporotic individual to sustain a fracture, perhaps without obvious trauma, pain, or other warnings. The main concern is that the usually elderly patient will be at high risk for additional fractures. In osteoporosis there is a normal mineral/collagen ratio, just less mineralized tissue per unit area which distinguishes the disorder from osteomalacia where mineral is reduced in amount relative to the bioorganic matrix. There is, as might be expected, no known cure for osteoporosis. It is part of the normal aging processes and pervasive in postmenopausal women and older men. The disease is considered a health issue in the United States and in other developed countries. It has become an important area for basic and clinical research with the incidence of fracture considered a signal of disease (Fig. 30.10). Onset of different types of osteoporosis, the relationship to nutrition, exercise, or other potential

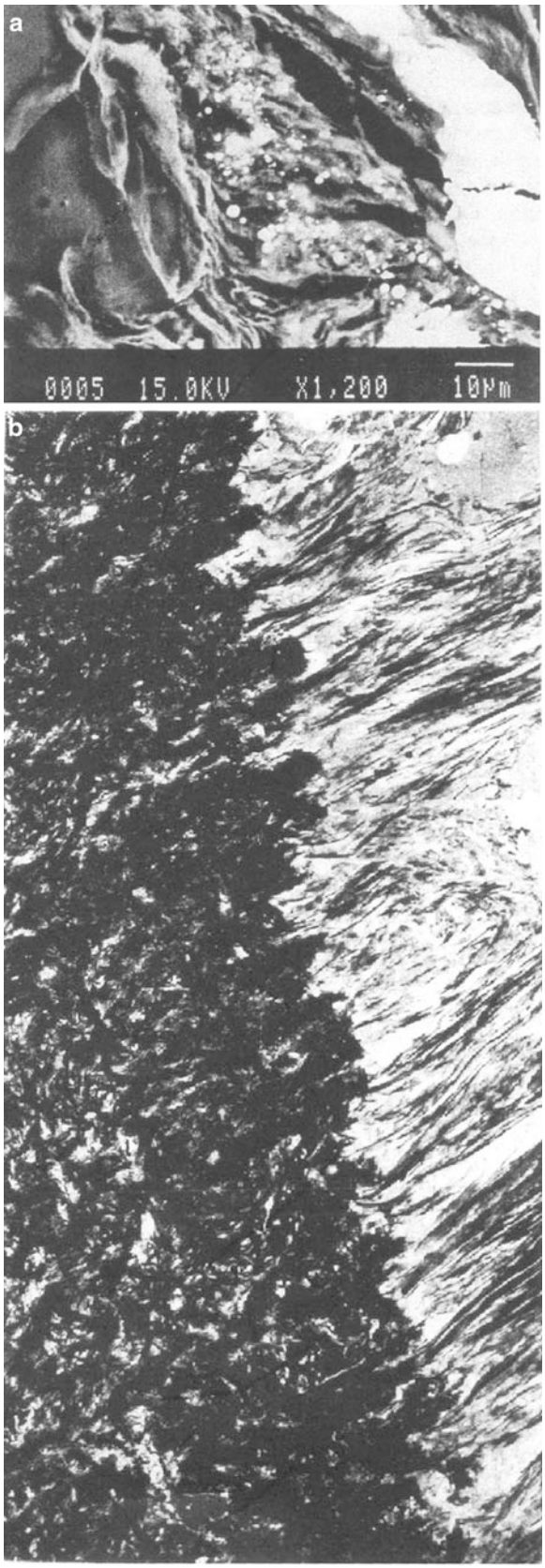


Fig. 30.8 Microradiographs taken with the scanning electron microprobe, backscattered images of breast tissue samples. (**a**) Calcium phosphate spherules deposited in breast tissue. Note the size of the spherules at this high resolution (magnification × 1,200) relative to the heavily mineralized calcium phosphate deposit (*white areas*) on the *right*. Section thickness is 6 µm (From Pogo et al. 1998, Figure 3.) (**b**) The dentin-enamel junction in a tooth showing the typical small crystallites in dentine (*left*) and the larger crystallites in enamel (*right*) (From Goose and Appleton 1982, Figure 2.6.) Enamel crystallites are about 100 µm in length

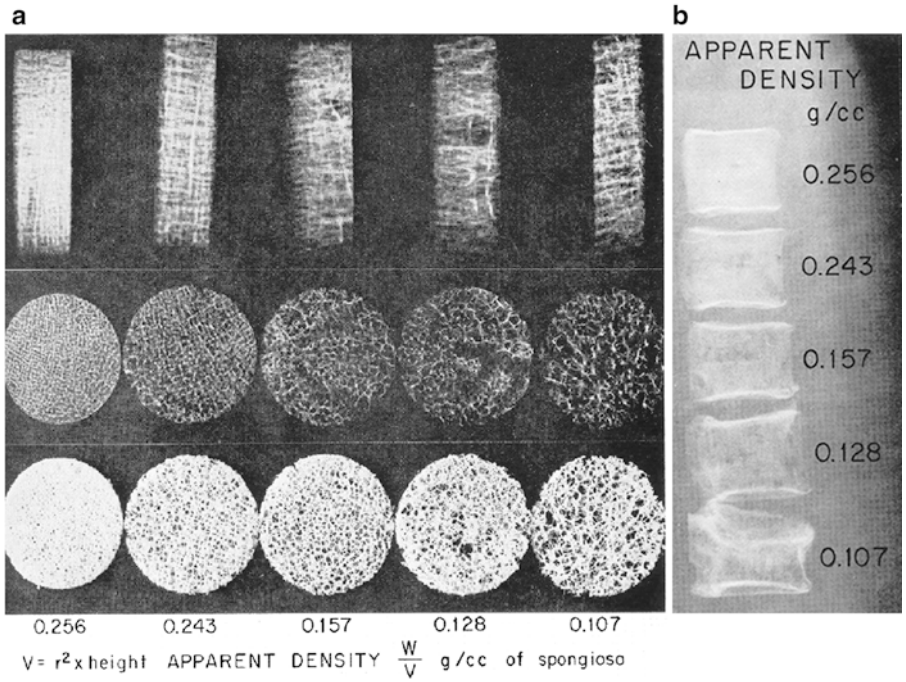


Fig. 30.9 Osteoporosis examined by transmission X-ray analysis. The first lumbar vertebra from five autopsy subjects illustrates the possibility of quantifying the level of mineral in a bone (From Barzel 1970, Figure 2B). (**a**) *Top row vertical* sections through the lumbar vertebrae showing changes in the distribution of mineral (*white*) from homogeneous fully mineralized tissues on *left part* of the figure to more porous on the *right* where the trabeculae thicken and appear more vertically oriented in the samples from older individuals. *Middle row* transverse section through the vertebrae shows increased porosity and disorder of the vertical sections. *Bottom row* transverse sections have been extracted and are now fat-free and dry. The porosity increase and the density of the mineral per unit area can be measured. Formula used to calculate the apparent density of columnar sections is weight/volume, g/cc. (**b**) Transmission X-ray of the entire vertebrae illustrating the apparent density differences that might be visible on radiologic examination (From Barzel 1970, Figure 4, *left half*). The several vertebrae depicted in this figure coincide with the mineral density calculated for the sections in Fig. 30.9a

contributing factors, and the results of a variety of treatments are the factors considered in the many epidemiological studies in several countries.

Bone mass is genetically programmed for an individual but may be modulated qualitatively and quantitatively by environmental factors. Reduction of bone tissue density in the vertebrae, a hallmark of osteoporosis, and subsequent vertebral fracture is estimated to be as high as 1 out of 4 women by age 65–70 (Wasnich 1996; Eastell 1999). This expression of compromised bone strength results in debilitation and pain with normal body movements and often premature death. Medical attention now must go beyond bone quantity (mass) into bone quality: architecture and rates of turnover (Chestnut et al. 2001) that involve molecular biochemistry and the relationships of the inorganic with organic constituents.

Osteoporosis, a focus of public attention when bone loss was recorded for the astronauts subjected to weightlessness, is also a consideration when there are long periods of inactivity such as for disabled people in wheelchairs or bedridden. The dynamic bone tissues respond to normal wear and tear. A cadre of metabolic disorders that result in bone loss in the young as well as in the old have been identified (Avioli 2000). The accumulated knowledge from abnormal situations has illuminated the complicated array of physical and chemical interactions necessary to maintain a viable skeleton. However, what constitutes effective treatment or, best of all, prevention of osteoporosis, remains elusive. A multiplicity of approaches that include not only pharmacologic intervention, but genetic evaluations (Econs 2000), diet (Marcus et al. 1996), and exercise (Riggs and Melton 1995) address the disease.

30.6.1 Detection of the Disease

Throughout the skeleton and its interrelated body systems there are dynamic changes but probably none are so obvious a sign of aging as the bend of the spine, also known as dowagers hump, because it is often typical of older women. Figures 30.6b and 30.9a, b depict the remarkable differences at the tissue level in the distribution of trabeculae in affected vertebrae. A diagnosis of osteoporosis is inferred from a clinical examination and confirmed by radiologic

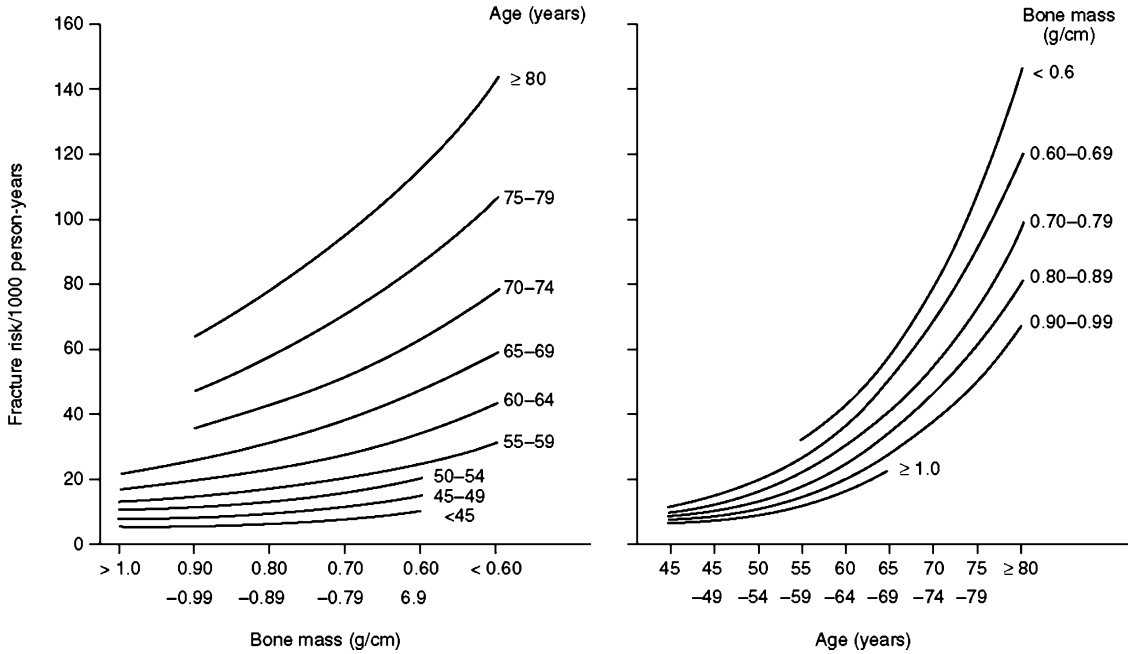


Fig. 30.10 Graphs comparing the average (over 1,000 person years) of fracture risk versus bone mass and fracture risk versus age (From Johnston et al. 1996, Figure 1)

examination using X-ray radiographic transmission analyses. A radiograph of the forearm, or leg bones, which is available as a result of an accidental fracture, may present poorly mineralized, inhomogeneous, or "porous" bone and tissues that are reminiscent of these spine photos.

The density level, or mineral content, of bones can be measured using these radiographs, and such examinations are noninvasive (Johnston et al. 1996). Digital radiographic techniques can quantify the mass of any bone or portion thereof. The results are compared with results from persons of like race, stature, and age to estimate the degree of osteoporosis and the potential risk for future fractures for a specific patient. What is actually measured is the mineral concentration per mass of bone (Fig. 30.10).

High-resolution radiographic analysis methodology, computerized axial tomography (CAT) scans, has also been used to show local differences in mineral density and distribution in the cortical bone, or the number, size, and organization of the trabeculae in a bone. These measurements are a reminder that each bone has its own biomineralization system, which is independent during formation and must be maintained at a certain level to be a contributing and effective part of the skeleton.

Radiological and densitometry analyses are useful in aiding diagnosis, but treatment for an individual depends on the patient's distinctive metabolic status. These data may inform on whether the osteoporosis is due to lowered amount or to overproduction of certain hormones that might influence the level of circulating calcium or phosphorus, for example. There are many other factors that can impact the several cells unique to the bone tissue system that must be considered in mineral formation and maintenance.

An initial designation of osteoporosis via radiographic survey and density of mineral amount per unit area may be followed with a bone biopsy. A small portion of tissue usually from the iliac crest (hip) is extracted by syringe and prepared for histological examination. Histomorphometry techniques examine texture and quantify tissue components, such as the number of the essential cells, specific hormones, and proteins, and estimate the level of bone formation and resorption. Does the tissue show normal amounts and distribution of mineral? Are there sufficient osteoblasts, the bone forming cells present, or alternatively, bone tissue resorption accelerated with a high concentration of osteoclasts (Frost 1973)? These two options are the yin and yang of bone remodeling, the dynamic system that underlies the viability of all skeletal organs (Urist 1965; Mundy 1999).

The onset of osteoporosis is asymptomatic and prevention is at least partially associated with an adequate diet and the intake, transport, and uptake of calcium, protein, and calories as well as vitamin D. The evaluation of personal choices, such as consumption of calcium-containing foods or specific trace elements via supplements may not be easily assessed for a particular patient. A variety of elements, hormones, enzymes, and proteins are essential in the complicated integration of other organs and cycles of the body system for maintaining a functional set of dynamic skeletal

tissues (Fig. 30.11). From the production of Vitamin D, or parathyroid hormone, to the absorption of calcium at the intestines, and recycling of phosphorus by the kidney, research has shown that all are important and must be properly integrated in order to arrive at the best treatment for a particular patient with osteoporosis (Avioli 2000).

30.6.2 Treatments for Osteoporosis

Over the past 50 years investigations from the cell level to animal model systems have been utilized to assess treatment regimens. Although some therapies may appear promising in animals, they may not translate to effective treatments for humans (Jee 1995). Much of the research has involved tests using pharmacologic agents that act on several body systems and effect the balance of circulating calcium and phosphorus. This seems a sensible approach applicable to all the bones in the body rather than attempting to construct treatment for specific bones.

The goal of osteoporosis treatments is to assist the body toward normal bone growth and repair and specifically to provide for normal mineral formation and retention. Because the mineral is our focus, but only one component in a complicated and multifaceted system, the following three sections on pharmacologic treatments focus on the composition and relative amounts of mineral to bring out some of the contributions provided by this basic science presentation.

30.6.2.1 Fluoride

The first thrust for making fluorine a therapeutic agent was put forward by dental scientists. Their studies led them to suggest that if the mineral in dental tissues was fluorapatite it would be less susceptible to caries as fluorapatite was the more stable apatite (Vischer 1970). To increase the formation of fluorapatite would require incrementing the amount of fluorine ingested that would then become part of the bone mineral as the inorganic portion of the tissues should respond to the nutrients supplied. However, it was known that some geographic areas with high fluorine in the waters led to a disease known as fluorosis. The simple hypothesis of more fluorine belies other aspects of the complicated bone tissue system. Fluoride may not only form fluorapatite, it may affect bone cells and the formation of the organic matrix. To obtain the optimum amount of fluorine to benefit human bones requires not only bioapatite production and composition, but also consideration of tissue recycling.

In the presence of small amounts of fluorine there was an increase in mineral matter and therefore the density of bone tissues. One study showed that fluorine was mitogenic for osteoblasts and a clinical investigation produced dramatic sustained (years) increases in bone mineral density when administered in doses of 20–30 mg fluoride daily (Riggs and Melton 1995). However, other clinical studies showed no effect and the cortical portion of the long bones decreased in density relative to an increase in trabecular bone density. The disturbance of cortical tissues could be modulated by increasing Ca and Vitamin D intake or by arranging F-free periods during treatment (Watts 1999). These interesting insights into the use of fluoride supplements as a possible means to increase fluorapatite in bone tissues also showed that only some of the fluorine found its way into bioapatite. When new mineral was formed little if any fluorine was partitioned into already existing bone mineral. The predominant bioapatite phase on bulk analyses remained hydroxylapatite.

Fluoride may be rapidly absorbed from the stomach but 50% will be eliminated via the kidneys within a few hours. Serum fluoride levels between 0.1 and 0.25 mg L^{-} and doses greater than 70 mg daily (35 mg fluorine) produce grossly abnormal bone (Riggs and Melton 1995). In areas of China with naturally high fluorine in the environment, telltale signs of brown spots (fluorosis) on the teeth and bent legs and bodies attest to its interference in the strength and architectural character of normal mineralized tissues (Finkelman et al. 1999). There is a narrow therapeutic window for fluorine. Domestic water levels at 1 μg–L^{-1} achieve dental benefits and may also provide some osteoporotic relief, but what is the level and length of treatment, or which cooperative treatments might be paired with fluorine ingestion to ensure no abnormal demineralization, and strong bone?

In shark teeth whose enameloid tissue shows the highest fluorine mineral content for any vertebrate, the uptake unfortunately appears to be related to genetics and phylogeny and not to the fluoride concentration in the aqueous environment (Aoba et al. 1998). The appropriate dose and best schedule for human fluorine ingestion have not been fully determined. The possible impact of fluoride on bone and on other body systems will require much longer and more comprehensive investigations before fluoride can be part of routine treatment of postmenopausal or other types of osteoporosis.

30.6.2.2 Bisphosphonates

Bisphosphonates are a group of compounds, analogues of the pyrophosphates, where the phosphorus atom of the phosphate group connects directly to a carbon atom. The chemical arrangement of several pharmacological bisphosphonates that are used to treat osteoporosis approximates the general structure with sites, e R1 and R2 representing other chemical substituents (see Fleisch 2000). These phosphate-containing chemicals adsorb onto mineral surfaces and are not readily degraded chemically or enzymatically, an attribute that makes them useful markers of mineral materials. Specific formulations with easily detectable elements in the "R" groups (such as radioactive species in these adducts) have been employed in nuclear medicine

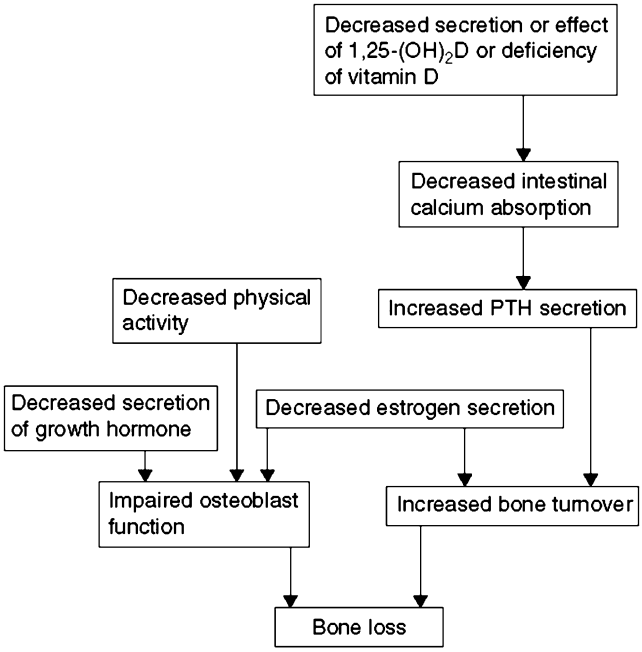


Fig. 30.11 Sketch illustrating the proposed contributions to age-related bone loss in women. The interactions of vitamin D, hormones, and activity level affect the uptake of calcium, bone turnover, and cell function, which if not in balance, may cause bone loss (From Eastell 1999, Figure 1)

experimentation. However, other, non-nuclear-containing bisphosphonates are useful in treatment of osteoporosis because they act locally by preventing the resorption of calcium phosphates by osteoclasts. Some bisphosphonates may inhibit osteoclast activity or cause apoptosis (cell death), but the major effect is to slow down the dissolution of bioapatite and therefore reduce bone remodeling, although the action is usually transient.

Only 5% of an oral dose of a bisphosphonate is absorbed from the gut, and the amount will be lower in the presence of calcium or other divalent ions. The drugs, usually taken in the morning before any food is consumed, will mostly be excreted by the kidney. Because bisphosphonates bind to the mineral they may become buried in the tissue becoming inactive. It is estimated that roughly 25% of the amount bound will be lost from the skeleton after 10 years, an indication of the slow rate of bone tissue turnover.

There are side effects of bisphosphonate ingestion: some tissues may not mineralize properly, and an increased amount of non-mineralized matrix may be the locus for fracture of the bone. Oral administration may also cause some soft tissue side effects but intravenous bisphosphonates have been effective in patients with Paget's disease (Watts 1999). Current research using animal model systems and high-resolution transmission electron microscopy show that bisphosphonates may increase the width of individual apatitic crystallites in the tissues but not markedly increase the mineralization level. Such results beg additional questions. One might be phrased as follows: In the composite tissues typical of bone where the physical and chemical properties of all the components must be maintained to ensure proper and continued interactions and organ function, is the major effect thus far noted with bisphosphonates, increasing crystallite size, an appropriate and useful contribution for long-term treatment of osteoporosis? The long term treatment has been questioned as jawbone degeneration was encountered in some patients. The latest medical recommendations are to discontinue bisphosphonate medication after 5 years.

30.6.2.3 Hormones

The most familiar pharmacologic treatment for postmenopausal osteoporosis is to augment the lowered production of estrogens in older women. Although the maximum level of skeletal bone density is affected by nutrition and usually achieved by the age of maturity (25–30 years), estrogen was shown to be important in producing and keeping the calcium levels in the circulation adequate for proper skeletal mineralization. For aging individuals the addition of estrogens, originally prescribed to relieve hot flashes and vaginal dryness, was shown to increase bone mineral density in the spine by 5–10% after age 65. Progesterone, or one of the progestins, was added to the estrogens in some formulations. Long-range clinical studies on the combined pharmaceuticals have shown no bone tissue benefits. Most recently these clinical studies have noted an increase in cardiovascular disease and breast cancer rates. This is most unfortunate and a sad commentary on hormonal replacement therapy that has been followed for many years by hundreds of thousands of women.

Although linked with skeletal health, the mechanisms for a direct action of the hormones on bone tissue remains under study. Early clinical investigations indicated reduced bone loss perhaps through suppression of osteoclastic bone resorption. Estrogens bind at receptors on the nuclei of target cells in both men and women and activate genes that affect the several different pathways required to maintain adequately mineralized bone tissues. For example, the normal production of growth factors and cytokines involved in calcium homeostasis may be altered without estrogens and other hormones. Estrogen acts at the kidneys and in the bowel, and via feedback mechanisms may sensitize the bone remodeling system cells to be more receptive to mechanically induced electrical signals (Watts 1999).

The "normal dose" of 0.625 mg daily of conjugated estrogens, usually equine derived, was lowered to 0.3 mg to prevent bone loss (Watts 1999). With a half-life of 10–18 h, estrogen compounds are easily metabolized by several different tissues once absorbed. From a series of investigations it became known that the stable equilibrium

for the hormone between estrone, the dominant form, estradiol, the active form, and other conjugated and esterified forms, is rapid. Estrogens circulate by binding to sex hormone-globins or albumin, the dominant circulating protein in blood and in tissues, but only the unbound form is biologically active. Estrogen compounds are excreted in bile, resorbed in the small intestine, and become less active in the process.

The side effects of tenderness, fluid retention, weight gain, occasional generation of deep vein thrombosis, and pulmonary emboli are some of the reasons that discouraged many women from starting or continuing the use of the estrogens post menopause. Raloxifene (Evista) acts to modulate the effect of estrogens on the surface of cells. It is an estrogen-acceptor modulator, and thus far it is the only drug approved that has been shown to prevent bone loss by causing differential expression of estrogen-regulating genes. The actions of estrogen and future use are cloudy at present and their potential contribution to minimizing bone tissue loss needs further study (Chlebowshi et al. 2003).

30.6.2.4 Summary

Bisphosphonates, possibly useful for reducing bone mineral resorption; hormones, particularly estrogen, certainly important in calcium-mineral dynamics; and fluoride, thought to increase the stability of the mineral phase; all have shown modest success for retaining or restoring mineral in bones in a variety of clinical investigations. Together with exercise and a diet that includes vitamins and certain trace elements, the battle to understand and relieve if not prevent osteoporosis continues. The lack of consistent and sustained benefits from the proscribed regimens on osteoporosis-affected populations makes it clear that additional experimental protocols are needed and will require lengthy clinical evaluation. Any new techniques or pharmacologic agents in these ongoing research efforts will provide useful data on the mineral, its roles and reactions.

30.7 Mineral and Mineralized Tissue Research and Medical Geology

In this chapter the several roles that mineral plays in mineralized tissues were discussed. Bioapatite is the stiffening agent for our skeleton and also constantly recycled in bone tissues. Bioapatite acts as a filter recording the many different ions ingested; some will be essential nutrients, and others perhaps hazardous to our health. Mineralized tissues play essential roles in growth, development, and maintenance of all bodily functions, organs, and tissues.

The fact that all naturally occurring elements and man-made chemicals are distributed throughout the environment where food and drink are obtained has recently led us to ask whether they are possibly impacting our health. I suggest in this chapter that the mineral composition of bone could be a potential signally agent to identify potential environmental problems. By exploring in detail the special physical and chemical characteristics of bioapatites, we increase our understanding of how the human body and its tissues react. For example, the addition of fluorine to domestic waters may benefit human dental health, it can also be one of the pharmacologic agents used in the treatment of osteoporotic patients still to be fully explored. The importance of minerals to the complicated and interactive metabolism of human systems provides an illustration of the information and sophisticated analyses from a host of scientificdisciplines that forge links between geological/mineralogical sciences and anthropological/health activities. Maximum integration between these disciplines is sought as a benefit for personal health and for populations at risk and to allow us to contemplate and eventually practice preventative medicine. A National Academy Report (2007) "Earth Materials and Health" discusses some of the research priorities between these disciplines.

See Also the Following Chapters. Chapter 6 (Uptake of Elements from a Biological Point of View) • Chapter 31 (Inorganic and Organic Geochemistry Techniques)

References

Albright JA, Skinner HCW (1987) Bone: structural organization and remodeling dynamics. In: Albright JA, Brand R (eds) The scientific basis of orthopaedics. Appleton & Lange, Norwalk, pp 161–198

Aoba T, Komatsu H, Shimazu Y, Yagishita H, Taya Y (1998) Enamel mineralization and an initial crystalline phase. Connect Tissue Res 38:129–134

Arnold JS, Bartley MH, Tont SA, Jenkins DP (1966) Skeletal changes in aging and disease. Clin Orthop 49:17–34

Avioli LV (ed) (2000) The osteoporotic syndrome: detection, prevention and treatment, 4th edn. Academic, San Diego

Barzel US (ed) (1970) Osteoporosis. Grune & Stratton, New York

Bell LC, Mika H (1979) The pH dependence of the surface concentration of calcium and phosphorus on hydroxylapatite in aqueous solutions. J Soil Sci 30:247–258

Blatt H, Tracy RJ (eds) (1996) Frontiers in petrology. W. H. Freeman, San Francisco

Bronner F (1996) Metals in bone: aluminum, boron, cadmium, chromium, lead, silicon and strontium. In: Bilezikian JP, Raisz LG, Rodan GA (eds) Principles of bone biology. Academic, San Diego, pp 295–303

Brown WE, Eidelman N, Tomazic B (1987) Octacalcium phosphate as a precursor in biomineral formation. Adv Dent Res 1:306–313

Carlstrom D (1955) X-ray crystallographic studies on apatites and calcified tissues. Acta Radiol Suppl 121:1–59

Chestnut CH III, Rosen CJ, Bone Quality Discussion Group (2001) Reconsidering the effects of antiresorptive therapies in recycline osteoporotic fracture. J Bone Miner Res 12:2163–2172

Chlebowshi RT, Hendrix SL, Langer RD et al (2003) Influence of estrogen plus progestin on breast cancer and mammography in

healthy postmenopausal women. The women's health randomized trial. JAMA 289:3243–3253
Driessens FCM, Verbeeck RMH (1990) Biominerals. CRC Press, Boca Raton
Eastell RC (1999) Pathogenesis of postmenopausal osteoporosis. In: Favus MJ (ed) Primer on the metabolic diseases and disorders of mineral metabolism, 4th edn. Lippincott Williams & Wilkins, Philadelphia, pp 260–262
Eckelmann WR, Kulp JL, Schulert AR (1957) Strontium–90 in man. Science 125:219
Econs MJ (ed) (2000) The genetics of osteoporosis and metabolic bone disease. Humana Press, Torowa
Elliott JC (1984) Infrared and Raman spectroscopy of calcified tissues. In: Dickson GR (ed) Methods of calcified tissue preparation. Elsevier, Amsterdam, pp 413–434
Elliott JC (1994) Structure and chemistry of the apatites and other calcium phosphates. Elsevier, Amsterdam
Elliott JC (2002) Chapter 11: Calcium phosphate biominerals. In: Kohn MJ, Rakovan J, Hughes JM (eds) Phosphates: geochemical, geobiological, and materials importance. Mineralogical Society of America, Washington, DC
Fabrikant JI (Chairman) (1988) Health risks of radon and other internally deposited α emitters. BER IV. National Academy Press, Washington, DC
Finkelman RB, Belkin HE, Zheng B (1999) Health impacts of domestic coal use in China. Proc Natl Acad Sci USA 96:3427–3431
Fleisch H (ed) (2000) Bisphosphonates in bone disease. Academic, San Diego
Food and Nutrition Board, National Research Council (1989) Recommended daily allowances, 10th edn. National Academy Press, Washington, DC, Federal Register 2206
Fowler JM (1960) Fallout: a study of superbombs, strontium 90 and survival. Basic Books, Inc., New York
Frost HM (1973) Bone remodeling and its relationship to metabolic bone disease, orthopaedic lectures, vol III. Charles C. Thomas, Springfield
Gaines RV, Skinner HCW, Foord EE, Mason B, Rosensweig A (1997) Dana's new mineralogy. Wiley, New York
Glimcher MJ (1976) Composition, structure, and organization of bone and other mineralized tissues and the mechanism of calcification. In: Greep RO, Astwood EB (eds) Handbook of physiology: endocrinology, vol 7. American Physiological Society, Washington, DC, pp 25–116
Glimcher MJ (1998) The nature of the mineral phase in bone: biological and clinical implications. In: Avioli LV, Crane SM (eds) Metabolic bone disease and clinically related disorders, 3rd edn. Academic, San Diego, pp 23–51
Goose DH, Appleton J (1982) Human dentofacial growth. Pergamon Press, Oxford
Halstead LB (1974) Vertebrate hard tissues. Wykeham Publications/Springer-Verlag, London/New York
Hughes JM, Rakovan J (2002) Chapter 1: The crystal structure of apatite, $Ca_5(PO_4)_3(F, OH, Cl)$. In: Kohn MJ, Rakovan J, Hughes JM (eds) Phosphates: geochemical, geobiological, and materials importance. Mineralogical Society of America, Washington, DC
Hughes JM, Cameron M, Crowley KD (1989) Structural variations in natural F, OH and Cl apatites. Am Mineral 74:870–876
Hughes JM, Cameron M, Crowley KD (1991a) Ordering of divalent cations in the apatite structure: crystal structure refinements of natural Mn- and Sr-bearing apatite. Am Mineral 76:1857–1862
Hughes HM, Cameron M, Mariano AN (1991b) Rare earth element ordering and structural variations in natural rare-earth bearing apatites. Am Mineral 96:1165–1173
Jee WSS (ed) (1995) Proceedings of the international conference on animal models in the prevention and treatment of osteopenia. Bone 17(Suppl):1–466
Johnston A Jr, Conrad C, Slemenda CW, Melton LJ III (1996) Bone density measurement and the management of osteoporosis. In: Favus MJ (ed) Primer on the metabolic bone diseases and disorders of mineral metabolism. American Association of Bone and Mineral Research/Lippincott Raven, Philadelphia, pp 142–151
Kim H-M, Rey C, Glimcher MJ (1995) Isolation of calcium phosphate crystals of bone by non-aqueous methods at low temperature. J Bone Miner Res 10:1589–1601
Klein C, Hurlbut CS (1985) Manual of mineralogy, 20th edn. Wiley, New York
Klug HW, Alexander WL (1954) X-ray diffraction procedures. Wiley, New York
Kohn MJ, Cerling TE (2002) Chapter 12: Stable isotope compositions of biological apatite. In: Kohn MJ, Rakovan J, Hughes JM (eds) Phosphates: geochemical, geobiological, and materials importance. Mineralogical Society of America, Washington, DC
Laborite J-C, Bone G, Mantel G (1973) Sur la structure et les proprieties des apatites carbonates de type B phospho-calcique. Ann Chime (Paris) 8:289–301
LeGeros RZ, Trautz, OR, LeGeros JP, Klein E (1968) Carbonate substitution in the apatite structure. Bull Soc Chimique. France (Spec.No.): 1712–1718
Malluche HH, Sherman D, Meyer R, Massy SG (1982) A new semi-automatic method for quantitative state and dynamic bone-histology. Calcify Tissue Into 34:439–448
Mann S (2000) Biomineralization principles and concepts in bioinorganic materials chemistry. Oxford University Press, Oxford
Marcus R, Feldman U, Kelsey J (1996) Osteoporosis. Academic, New York
McConnell D (1973) Apatite. Springer, New York
Mielke H (2003) Anthropogenic distribution of lead. In: Skinner HCW, Berger AN (eds) Geology and health: closing the gap. Oxford University Press, New York, pp 119–124
Miles AEW (ed) (1967) Structural and chemical organization of teeth. Academic, New York
Miller EJ (1973) A review of biochemical studies on the genetically distinct collagens of the skeletal system. Clin Orthop 92:260–280
Mundy GR (1999) Bone remodeling. In: Favus MJ (ed) Primer on metabolic bone diseases and disorders of mineral metabolism, 4th edn. American Society for Bone and Mineral Research/Lippincott Williams & Wilkins, Philadelphia, pp 30–38
National Academy Report (2007) Earth materials and health: research priorities for earth science and public health. In: Skinner HCW (ed) NRC and IOM. The National Academy Press, Washington, DC
Nesse WD (2004) Introduction to optical mineralogy, 3rd edn. Oxford University Press, New York
Neuman WF, Neuman MW (1958) Chemical dynamics of bone mineral. University of Chicago Press, Chicago
Pogo SH, Skinner HCW, Ague JJ, Carter D (1998) Using scanning electron microscopy to study mineral deposits in breast tissues. Am Mineral 83:1122–1126
Posner A (1977) The relation of synthetic amorphous calcium phosphate to bone mineral structure. Orthop Trans 1:78–95
Reed SJB (1993) Electron microprobe analysis. Cambridge University Press, Cambridge
Rey C, Kim HM, Gerstenfeld L, Glimcher MJ (1996) Characterization of the apatite crystals of bone and their maturation in osteoblast cell culture: comparison with native bone crystals. Connect Tissue Res 35:343–349
Riggs BL, Melton LJ III (eds) (1995) Osteoporosis: etiology, diagnosis and management. Raven, New York
Roberts JE, Bonar LC, Griffin RG, Glimcher MJ (1992) Characterization of very young mineral phases of bone by solid state ^{31}phosphorus magic angle sample spinning nuclear magnetic resonance and X-ray diffraction. Calcify Tissue Into 50:422–448

Shannon RD (1976) Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallogr A 32:751–767

Shannon RD, Prewitt CT (1969) Effective ionic radii in oxides and fluorides. Acta Crystallogr B 25:925–946

Skinner HCW (1968) X-ray diffraction analysis techniques to monitor composition fluctuations within the mineral group apatite. Appl Spectrosc 22:412–414

Skinner HCW (1973a) Phase relations in the $CaO–P_2O_5–H_2O$ system from 300–600°C at 2 kb H_2O pressure. Am J Sci 273:545–560

Skinner HCW (1973b) Studies in the basic mineralizing system $CaO–P_2O_5–H_2O$. Calcify Tissue Res 14:3–14

Skinner HCW (1987) Bone: mineral and mineralization. In: Albright JA, Brand R (eds) The scientific basis of orthopaedics, 2nd edn. Appleton & Lange, Norwalk, pp 199–211

Skinner HCW (1989) Low temperature carbonate phosphate materials or the carbonate apatite problem: a review. In: Crick R (ed) Origin, evolution and modern aspects of biomineralization in plants and animals. Plenum Press, New York, pp 251–264

Skinner HCW (2000a) Minerals and human health. In: Vaughan DJ, Wogelius RA (eds) Environmental mineralogy. European mineralogical union notes in mineralogy, vol 2. Eotvos University Press, Budapest, pp 383–412

Skinner HCW (2000b) In praise of phosphates or why vertebrates chose apatite to mineralize their skeletons. Into Geol Rev 42:232–240

Skinner HCW, Nalbandian J (1975) Tetracyclines and mineralized tissues: review and perspectives. Yale J Biol Med 48:377–397

Skinner HCW, Kemper E, Pak CYC (1972) Preparation of the mineral phase of bone using ethylene diamine extraction. Calcify Tissue Res 10:257–268

Skinner HCW, Nicolescu S, Raub TD (2004) A tale of two apatites. In: Petrescu I (ed) Environment and progress 2. Editura Fundatiel de Studii Europene, Cluj-Napoca, pp 283–288

Skinner HCW (2012) Minerals and Human Health EMU Notes in Mineralogy 11:441–485

Suich PR, LaCout JL, Hewat A, Young RA (1985) The structural position and type of Mn^{2+} partially substituted for Ca^{2+} in fluorapatite. Acta Crystallogr B 41:173–179

Urist MR (1965) Bone: formation by autoinduction. Science 150:893–899

Van Waser JR (1958) Phosphorus and its compounds. Interscience Publishers/Wiley, New York

Vischer TL (1970) Fluoride in medicine. Han Huber Publications, Bern

Wasnich RD (1996) Epidemiology of osteoporosis. In: Favus MJ (ed) Primer on the metabolic bone diseases and disorders of mineral metabolism, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, pp 249–252

Watts NB (1999) Pharmacology of agents to treat osteoporosis. In: Favus MJ (ed) Primer on the metabolic bone diseases and disorders of mineral metabolism, 4th edn. Lippincott Williams & Wilkins, Philadelphia, pp 278–283

Weiner S, Wagner HD (1998) The material bone: structure–mechanical function. Ann Rev Mater Res 28:271–298

Young RA, Holcomb DW (1982) Viability of hydroxylapatite preparations. Calif Tissue Res 34(Supplement 2):S17–S32

Zipkin I (1970) Inorganic composition of bone. In: Schraer H (ed) Biological calcification: cellular and molecular aspects. Appleton, New York, pp 69–104

Inorganic and Organic Geochemistry Techniques

31

Mitko Vutchkov, Gerald Lalor, and Stephen Macko

Contents

31.1 Introduction

Inorganic and organic geochemistry are applications of analytical chemistry to solve Earth science problems by analyzing the composition of geological and biological systems. Measurements and characterization of unknown substances are used for making policy decisions, exploring mineral resources, pollution prevention, and management of environmental hazards. Geochemical exploration methods use trace element geochemistry to identify the "fingerprints" of certain minerals and rocks and to discover new deposits. Environmental geochemistry has become an essential tool for determining where man-made or natural distributions of elements can become a potential environmental health hazard. Organic geochemistry adds a valuable contribution to determining the nature, origin, and distribution of organic constituents in the environment.

This chapter brings together descriptions of the most common inorganic and organic geochemistry techniques. The section below provides an introduction to the analytical geochemistry techniques, reference sources on the subject, and glossary of the technical terms and acronyms. The analytical techniques are described in a standardized format to facilitate the comparison between them and include a brief introduction, principle of operation, what each one does, and how it is used. Typical application examples are also provided.

The most common inorganic and organic geochemistry techniques comprise methods for analysis of major, minor, and trace elements; age dating; stable isotopes; etc. The data derived from these methods are used for interpretation of the global and local element cycles as well as the distribution of elements in the environment. Figure 31.1 compares the most common analytical techniques used for analysis of major, minor, and trace concentrations of the elements in the Earth's crust.

Inorganic geochemistry techniques for solids and liquids include a suite of analytical methods for measuring chemical parameters of materials that can be fundamentally divided

M. Vutchkov (✉) • G. Lalor
International Centre for Environmental and Nuclear Sciences, University of the West Indies, Kingston, Jamaica
e-mail: mitko.voutchkov@gmail.com

S. Macko
Department of Environmental Sciences, University of Virginia, Charlottesville, VA, USA

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_31,

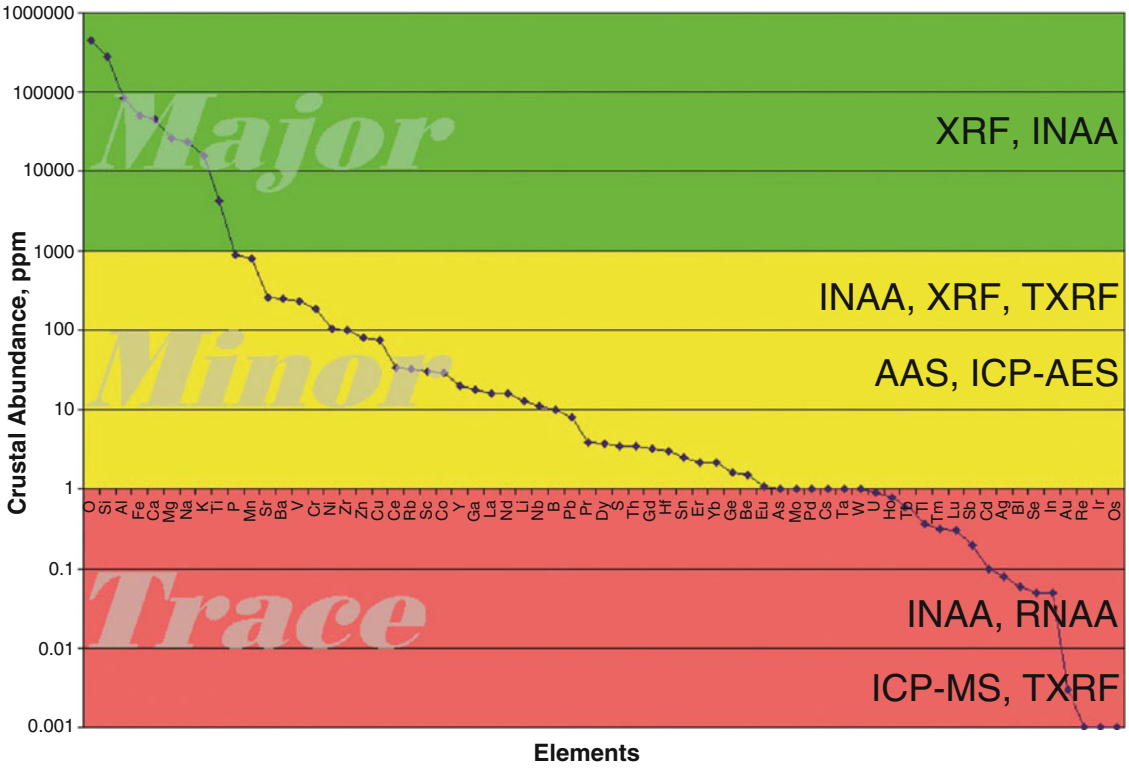


Fig. 31.1 Analytical methods selection chart for major, minor, and trace elements concentration in the Earth's crust. XRF X-ray fluorescence spectrometry; *INAA* instrumental neutron activation analysis; *TXRF* total reflection XRF; *AAS* atomic absorption spectroscopy; *ICP-AES* inductively coupled plasma atomic emission spectroscopy; *RNAA* radiochemical neutron activation analysis; *ICP-MS* inductively coupled plasma mass spectrometry

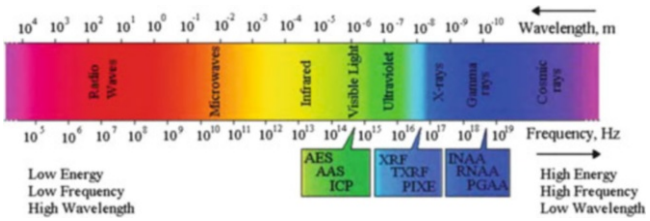


Fig. 31.2 EM spectrum and related spectroscopy techniques

into spectroscopic techniques and non-spectroscopic or classical, wet chemistry methods. Spectroscopic techniques use the interaction of electromagnetic (EM) radiation with a sample to perform an analysis, while the classical methods utilize physical means of detecting analytes, such as mass, volume, density, color, refraction, conductivity or electric charge. Figure 31.2 illustrates the different regions of the EM spectrum based on the origin of photons, i.e., radio waves, microwaves, infrared, visible, ultraviolet (UV), X-rays, gamma, and cosmic rays, and the associated analytical techniques.

The EM radiation can be described in terms of a stream of photons traveling in a wave-like pattern with the speed of light. The photons can be expressed in terms of energy (E), wavelength (λ), or frequency (n) that are mathematically related by the fundamental equation of Max Planck

$$E = h\,\nu = hc/\lambda \tag{31.1}$$

where, h = Plank's constant ($4.136 - 10^{-15}$eVs) and c is the velocity of light ($3 - 10^{8}$m s^{-1}).

The physical principles and mathematical description of radiation across the EM spectrum is the same, but due to the different energy levels, different mechanisms of interaction with matter occur. Low-energy photons behave more like waves, while higher energy photons such as X-rays and gamma rays behave more like particles, i.e., they exhibit wave-like particle duality. This is an important factor used in designing detectors and instruments for spectroscopy measurements.

The energy of the photons in various regions of the EM spectrum corresponds to different types of transitions in the atoms and molecules, which can be detected and measured using specific spectroscopy techniques. For example, microwaves and infrared spectroscopy methods are

associated with molecular rotations and vibrations; UV, visible, and atomic absorption/fluorescence spectroscopy are based on outer electron transitions, whereas X-ray fluorescence (XRF) spectroscopy is related to inner electrons of the atom and gamma rays to nuclear transitions.

A variety of analytical instruments that are commonly used in inorganic and organic biogeochemistry are based on spectroscopy, e.g., UV-visible, fluorescence, atomic, infrared, X-ray, and nuclear spectroscopy. The interpretation of the spectra thus produced can be used for chemical analysis and examining atomic energy levels and molecular structures. Qualitative and quantitative analysis of the elemental composition using these techniques is based on absorption, emission, or scattering of electromagnetic radiation by matter, e.g., atoms, molecules, atomic or molecular ions, or solids. The analysis of concentrations of elements relies on the relationship between the measured EM radiation and concentration of the analyte. The unknown concentrations are determined from calibration curves using appropriate standard reference materials (SRM).

The various analytical techniques require analytical samples to be either in a solid or liquid form. Traditional geochemical analysis essentially involves getting a sample into solution and then using an appropriate method to measure the elemental concentration in the solution. These methods are generally referred to as destructive methods (sample dissolution) and include the classical wet chemistry methods: Atomic spectroscopy (AS), inductively coupled plasma (ICP) spectroscopy, and some other methods such as the radiochemical neutron activation analysis (RNAA). Nuclear and related analytical techniques depend on the physical properties of the atomic nucleus and do not need preparation of solutions, i.e., they are non-destructive. The most common non-destructive techniques include instrumental neutron activation analysis (INAA) and XRF spectrometry.

For detailed information on chemical analysis in pathology see Chap. 32, this volume.

31.2 Inorganic Geochemistry Techniques for Solids

31.2.1 Neutron Activation Analysis

NAA is a nuclear analytical technique used for non-destructive multi-element analysis of solids, liquids, or gases. Theoretically, about 70% of naturally occurring elements can be analyzed by NAA at concentration levels down to 1 μ kg^{-1}. Analysis can be performed instrumentally without chemical pretreatment of the sample, thus avoiding problems related to incomplete dissolution or loss of volatile elements. Since neutrons interact directly with the atomic nuclei, the NAA measures the "total" amount of an element, regardless of oxidation state or chemical form.

The neutron activation method was discovered in 1936 by Hevesy and Levi, and commercialized in the mid-1950s and 1960s. In the last three decades, NAA has been widely applied to mineral exploration, environmental biogeochemistry, and health-related studies. NAA is well recognized as a referee method of choice in the certification of new reference materials or quality control trials.

31.2.1.1 How It Works

Principle of NAA

Neutron activation is a physical method for analysis of elemental composition that uses neutron irradiation of a sample to convert the elements into radioactive isotopes. The radioactive elements can then be detected and quantified by gamma ray spectroscopy. The basic principle of NAA, shown in Fig. 31.3, illustrates how the radioactivity of atomic species grows and decays after exposure to neutrons.

The induced activity during the irradiation follows an exponential growth curve and depends on the half-life ($t_{1/2}$) of the particular nuclide. Therefore, the isotopes with longer half-lives ($t_{1/2}$) will need longer irradiation times compared to those of short-lived species. The induced activity of the sample will decrease with time according to the disintegration rates of isotopes present in the sample. The short-lived isotopes, used in NAA, have half-lives of less than an hour; medium-lived isotopes range from an hour to several days; and long-lived isotopes range from several weeks to 2–3 months.

The most common type of nuclear reaction used for NAA is the neutron capture or (n, gamma) reaction. When a naturally occurring stable isotope of an element (target nucleus) absorbs a neutron, it is transformed into higher mass unstable nucleus as shown in Fig. 31.4.

The excited nucleus instantaneously decays through emission of prompt gamma rays or, in most cases, is

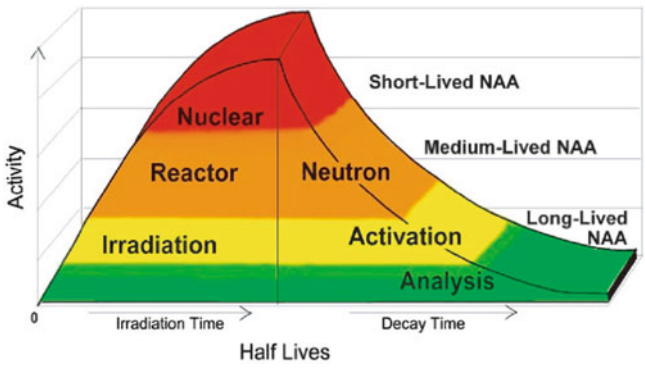


Fig. 31.3 Growth and decay curves of radioactivity of a natural sample exposed to neutrons

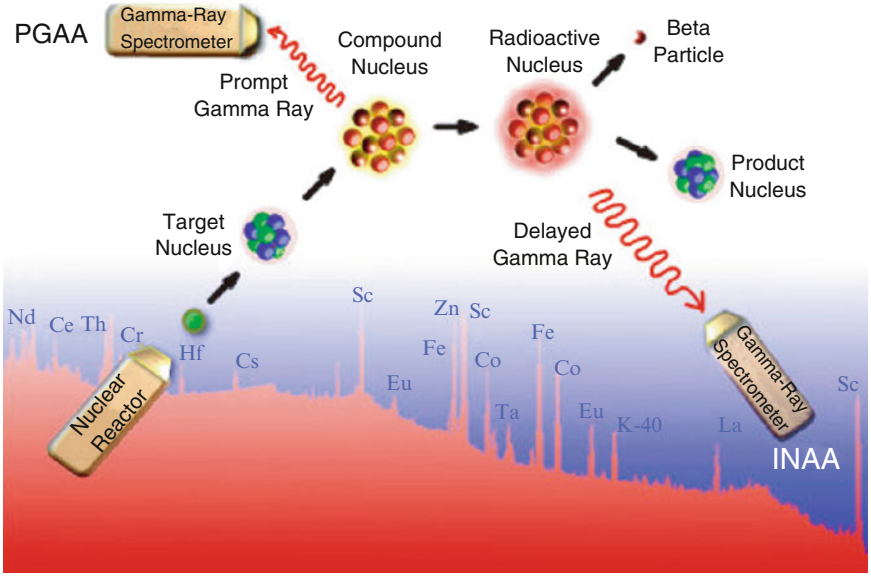


Fig. 31.4 A schematic diagram illustrating the principle of neutron activation

Table 31.1 Classification of most common NAA methods

Neutron activation analysis				
Gamma ray spectroscopy				Neutron counting
Prompt gamma	*Delayed gamma*			*Delayed neutron*
PGAA	INAA and RNAA	ENAA	FNAA	DNC

converted to a radioactive nucleus, which decays primarily by emission of beta particles and/or gamma rays.

The fundamental equation for activation, decay, and counting of a radionuclide with activity A and mass M (g) can be described as follows (Soete et al. 1972):

$$A = \frac{N\theta\sigma\gamma\rho W\phi\varepsilon}{M\lambda}\left(1 - e^{-\lambda t_i}\right)\left(e^{-\lambda t_d}\right)\left(1 - e^{-\lambda t_c}\right) \tag{31.2}$$

where: t_i, t_d and t_c = irradiation, cooling, and counting times, s; φ = thermal neutron flux, neutrons cm^{-2} s^{-1}, N = Avogadro's number, $6.023 - 10^{23}$ atoms g^{-1} atom; θ = abundance of the activated nuclide; σ = absorption cross-section of the irradiated species, cm^2; γ = gamma emission probability; ρ = concentration of analyte, μg kg^{-1}; W = sample weight, g; Φ = effective thermal neutron flux, neutrons cm^{-2} s^{-1}; ε = efficiency of the counting system; M = molecular mass of the target element; λ = decay constant of the radioisotope produced, s^{-1}.

The basic instrumentation necessary to carry out NAA includes a source of neutrons, gamma ray spectrometer, and data processing software. Nuclear reactors are most commonly used for NAA, providing high and well thermalized neutron fluxes. The neutron energy spectrum consists of three principal components sorted in increasing energy order: thermal, epithermal, and fast neutrons. The thermal neutrons comprise about 90–95% of the total neutron flux and induce the main (n, gamma) reactions of target nuclei used in conventional NAA. The epithermal and fast neutrons induce nuclear reactions, which release gamma rays or nuclear particles and the technique associated with them is called epithermal neutron activation analysis (ENAA) and fast neutron activation analysis (FNAA), respectively.

In principle, with respect to the underlying nuclear reactions and their measurement, neutron activation analysis methods are classified into three categories (Table 31.1): (1) prompt gamma NAA (PGAA), (2) delayed gamma NAA (DGNAA), and (3) delayed neutron counting (DNC).

Neutron-induced PGAA uses prompt gamma rays released from excited nuclei during sample irradiation (Fig. 31.4) and is manly applicable to elements that do not form radioactive products after irradiation (e.g., hydrogen and boron) or elements whose half-life is too short or long to be measured by NAA. DGNAA technique measures the delayed gamma rays obtained during the radioactive decay of nucleus and includes the conventional NAA, ENAA, and FNAA techniques. Depending on the treatment of the samples, the NAA technique is generally referred to as INAA when analysis is done without chemical processing and as RNAA when post-irradiated radiochemical separation of analyte or interfering nuclide is applied. DNC is used for quick determination of uranium and other fissionable radionuclides with improved detection limits.

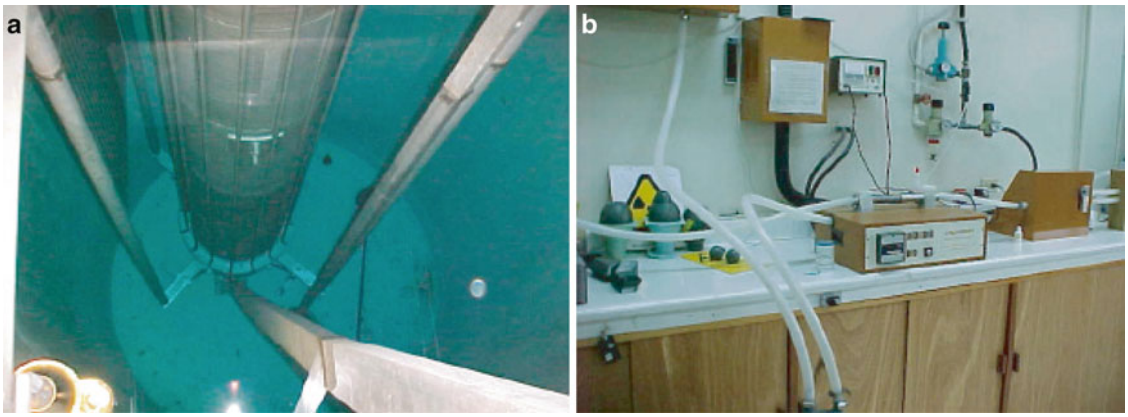


Fig. 31.5 Neutron activation analysis using Slowpoke-2 nuclear reactor in Jamaica: (**a**) vertical cross-section view of the reactor core-reflector configuration and (**b**) irradiation-transfer system for NAA

Irradiation Facilities

The irradiation of samples is typically performed in nuclear reactors with neutron fluxes ranging from 10^{11} n.cm^{-2}.s^{-1} to 10^{16} n.cm^{-2}.s^{-1}. The low power nuclear reactors such as TRIGA (training research isotopes general atomics) and Slowpoke (safe low power kritical experiment) with typical neutron fluxes 10^{13}–10^{12} n.cm^{-2}.s^{-1} are most commonly used in NAA. Figure 31.5a, b show the Slowpoke-2 reactor facility in Jamaica, which is used for NAA (Lalor 1995).

For neutron activation analysis samples are irradiated in or near the core of the nuclear reactor for an appropriate period of time, from seconds (short-irradiations) to several weeks (long-irradiations). This period of time depends on the half-lives ($t_{1/2}$) of the elements of interest. The irradiation sites designed for short-lived radionuclides are equipped with computer-controlled pneumatic transfer systems for quick transport and measurement of the samples. Activation of long-lived nuclides usually require much higher flux levels and longer irradiation times. Samples are loaded/unloaded automatically or manually depending on the application and irradiation facility. Neutron irradiation for PGAA is limited to reactors with external neutron beam facilities designed to perform simultaneous irradiation and counting of the samples.

Gamma Ray and Neutron Counting

Simultaneous neutron irradiation of natural materials generates numerous isotopes of elements that decay with different half-lives. Elements of interest are usually measured in several steps with different cooling times to allow the interfering elements to decay. The typical gamma ray counting equipment comprises a high-resolution germanium detector (<2 keV for ^{60}Co 1.33 MeV line), pulse processing electronics including an amplifier, an analog-to-digital converter (ADC), and a multichannel analyzer (MCA) interfaced to a desktop computer (Fig. 31.6).

Measurement of prompt gamma rays is performed using a high-purity germanium detector placed nearby to the neutron beam port. The geometry arrangement and shielding of the prompt gamma ray counting system is designed to minimize the background due to neutrons and gamma rays.

The delayed neutrons are detected with boron trifluoride (BF_3) proportional detectors surrounded by double cadmium sheets filled with paraffin wax. The paraffin wax is used to thermalize the fast neutrons while cadmium absorbs the thermal neutrons, thus ensuring low-neutron background around the BF_3 detectors.

Data Processing

Data processing software for neutron activation analysis includes programs for gamma ray spectrum analysis and qualitative and quantitative analysis. Spectrum analysis programs perform identification of nuclides (qualitative analysis), background subtraction, and calculation of the net peak areas of elements. The quantitative analysis software employs the net peak areas to calculate the elemental concentrations using standards.

31.2.1.2 What It Does

The principal objective of the neutron activation analysis, qualitative and quantitative multi-element analysis, can be completed in several basic steps, which are summarized in Table 31.2.

Sample Preparation

Solid samples are analyzed by NAA as received without any chemical pretreatment, which avoids losses of volatile elements (e.g., arsenic, selenium, cadmium, and mercury) and contamination of the sample. Small volumes of liquid samples can be analyzed directly provided that their boiling point is greater than 60°C, or alternatively they are can be evaporated to dryness before analysis. The common sample preparation procedure for NAA involves weighing about 100- to −500-mg samples into heat-sealed polyethylene or quartz vials. As a rule, the vials should not be filled more than three quarters of the volume to prevent pressure

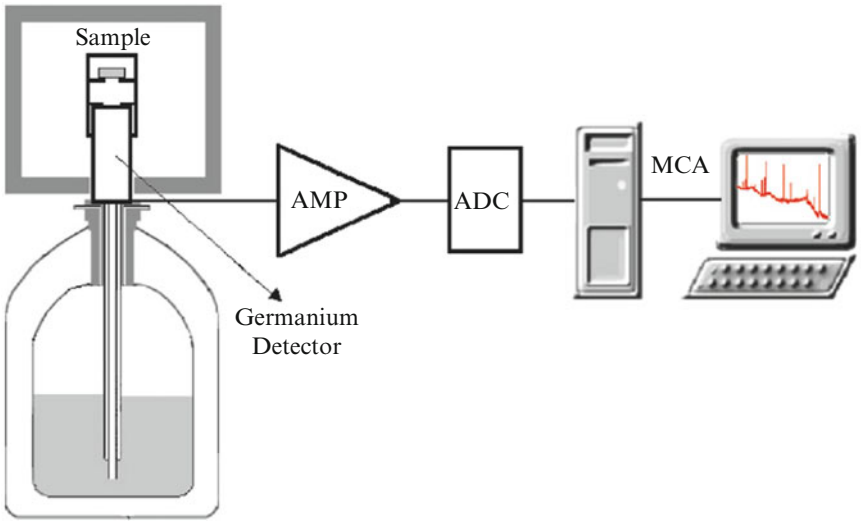


Fig. 31.6 Schematic representation of a gamma ray counting system

Table 31.2 Basic steps for neutron activation analysis

Step	INAA	RNAA	DNC	PGAA
1. Sample preparation	Yes	Yes	Yes	Yes
2. Irradiation	Yes	Yes	Yes	Yes
3. Separation	No	Yes	No	No
4. Cooling periods	Multiple	Multiple	Single	No
5. Measurement(s)	Multiple	Multiple	Single	Single
6. Quantitative analysis	Yes	Yes	Yes	Yes

buildup. The encapsulated samples are packed together with standards and a flux monitor into larger vials used for irradiation. All sample containers should not be handled with bare hands to avoid contamination of the vial surface with sodium and chlorine.

Irradiation of Samples

The "unknown" samples are activated in the nuclear reactor for different irradiation times and neutron flux levels taking into consideration nuclear properties of the element of interest and the expected concentration of elements. As a rule, the irradiation time should be several times greater than the half-life of the short-lived radionuclides and as long as practically possible for the long-lived isotopes. There are two main irradiation schemes employed in NAA, short- and long-irradiations. Short-irradiations are used for NAA of isotopes with half-lives less than 1 day, whereas long-irradiations are applied for medium- and long-lived nuclides. The overall activation time varies from seconds to several days and weeks. Table 31.3 lists the typical irradiation, decay, and counting times employed for activation of the very-short, short-, medium-, and long-lived nuclides using the Slowpoke-2 reactor (Lalor et al. 2000; Vutchkov et al. 2000).

As shown, the complete analysis of the elements listed in Table 31.3 can be performed after two irradiations, for short- and long-lived elements, respectively.

Radiochemical Separation

Radiochemical NAA involves isolation of one or a group of elements from the irradiated sample before measurement by gamma ray spectroscopy. Chemical separation is usually done after irradiation of the sample in order to remove interfering elements and/or to concentrate the analyte isotope(s). This procedure can improve the detection limits of elements by several orders of magnitude. RNAA requires special facilities and well-trained personnel in handling open radioactive materials. For this reason it is mainly used for certification of new reference materials, analysis of human tissue samples, and other special projects.

Cooling

Samples removed from the reactor are frequently highly radioactive due to the activation of elements with high-neutron cross-sections and/or high concentrations. The irradiated samples are stored in shielded lead containers for a specific period of time referred to as "cooling," before gamma ray spectrometry measurement. The cooling period allows the short-lived radioactive species that might interfere with the analyzed element to decay and to therefore reduce the overall radioactivity of the sample. If the half-life of the analyte is much longer than that of the interfering nuclide, after sufficient cooling time the interference can be completely eliminated. Because the production and decay rates of gamma ray radiation are dependent on the half-life of the nuclide, measurement of the elements of interest can be optimized by varying irradiation and cooling times. As shown in Table 31.3, the sequential cooling and counting after several periods of time can optimize the determination of about 30–50 elements in various materials. Depending on the decay time of the elements of interest, the complete analysis report can be obtained from 10 min to 30 days after irradiation.

Table 31.3 Typical irradiation-decay-counting scheme for multi-element NAA of geological and biological materials using Slowpoke-2 reactor in Jamaica

	Analysed elements			
Nuclides	Very-short	Short	Medium	Long
Product half-life	<12 min	12–1440 min	1–3 day	>3 day
Neutron flux ($n.cm^{-2}.s^{-1}$)	2×10^{11}	5×10^{11}	10×10^{11}	10×10^{11}
Irradiation time	5 min	5 min	4 h	4 h
Cooling time	3–6 min	15 min	2–5 day	20–30 day
Counting time	5 min	30 min	1 h	3 h
Elements	Magnesium, calcium aluminium, cooper, titanium, vanadium	Chlorine, potassium, sodium, iodine, dysprosium, manganese	Arsenic, bromine, sodium, cadmium, gold, lanthanum, samarium, etc	Selenium, scandium, chromium, caesium iron, europium, zinc, uranium, thorium, hafnium, tantalum, cerium, barium, etc

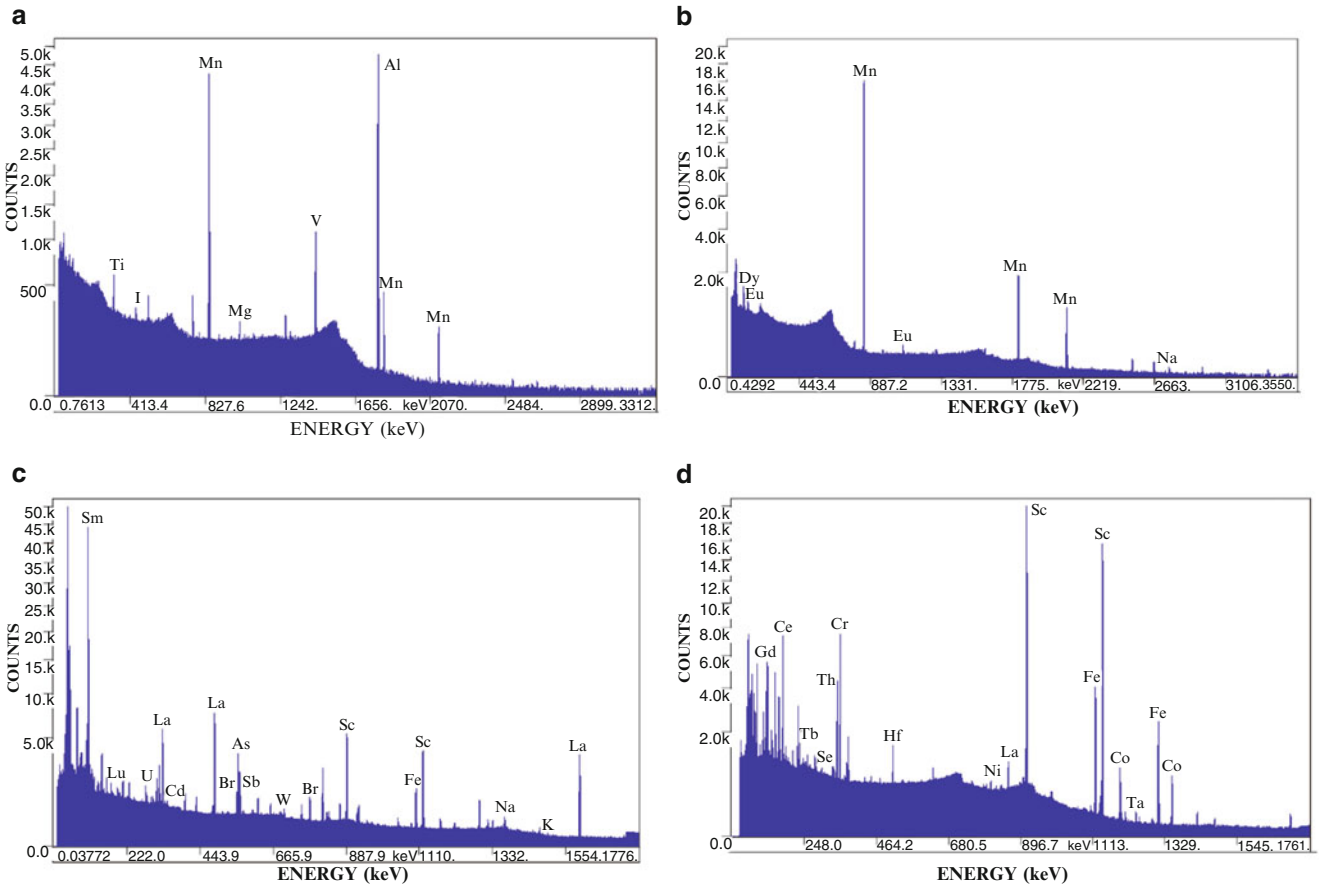


Fig. 31.7 Gamma ray spectra of a Jamaican soil sample measured for (**a**) very-short, (**b**) short-, (**c**) medium-, and (**d**) long-lived elements.

Measurement

In most cases, measurement of delayed gamma ray spectra of irradiated samples is performed repeatedly after different cooling times as indicated in Table 31.3. The counting time of different groups of elements depends on their half-lives and the overall activity of the sample. Typical gamma ray spectra of Jamaican soil sample measured for very-short, short-, medium-, and long-lived elements are shown in Fig. 31.7.

As seen from Fig. 31.7a, b, elements such as aluminum, magnesium, vanadium, manganese, cooper, and dysprosium can be determined in two measurements after 5 and 30 min of cooling. Medium-lived nuclides of sodium, potassium, bromine, arsenic, cadmium, gold, and tungsten (Fig. 31.7c) are counted after 2–5 days cooling time. Very-long lived elements such as chromium, iron, barium, cesium, rubidium, selenium, scandium, rare earth elements, etc. (Fig. 31.7d) are typically counted after 3–4 weeks decay time.

Measurement of prompt gamma ray spectra is performed during irradiation of the sample with neutrons. In prompt gamma activation the multi-element analysis is performed in one measurement, but the counting times may range from minutes to several hours per sample. The detection limits of PGNAA can be improved by increasing the sample mass or by using longer measurements and irradiation times.

Counting of delayed neutrons is performed after a decay time of 10–20 s for about 1 min. The DNC method using a pneumatic transfer system can be fully automated and has the highest turnaround time among the NAA methods.

Qualitative Analysis

Qualitative neutron activation analysis involves processing of the gamma ray spectrum and identification of activated nuclides. Spectrum analysis normally performs smoothing of spectra, background subtraction, peak searching, determination of energies of the photo peaks, multiplet deconvolution, and net peak area determination. The qualitative NAA is accomplished by library matching of the energies of identified nuclides in the sample.

Quantitative Analysis

Quantitative NAA is generally accomplished by using the relative comparator method. The most common quantitative analysis approach is the comparator method in which the unknown sample (unk) is co-irradiated with comparator standards (std) and measured under identical counting conditions. The main advantage of this procedure is that it eliminates the nuclear and instrumental parameters and simplifies Eq. 31.2 as follows:

$$C_{unk} = \left(C_{st}\frac{W_{std}}{W_{unk}}\right)\left(\frac{A_{unk}e^{-\lambda t_{std}}}{A_{std}e^{-\lambda t_{unk}}}\right) \qquad (31.3)$$

Calculations of elemental concentrations include corrections for decays of unknown and standards, sample weights, and other parameters as indicated in Eq. 31.3. The comparator single and multi-element standards are prepared from certified solutions of a known concentration of the elements of interest. Each sample analysis batch also includes a certified standard reference material similar to the matrix of the analyzed samples such as soil, sediments, rock, coal, plant tissue, etc. In multi-element analysis mode the comparator method is a time-consuming procedure and requires the measurement of a large amount of standards.

An alternative quantitative analysis approach gaining large popularity is the so-called k0-method, which eliminates the need for co-irradiation of comparator standards. By measuring a single comparator (e.g., gold, zinc, zirconium, etc.) and the efficiency of the counting system, quantitative analysis can be performed using a library with k0-factors of elements. The k0-method has a great advantage in multi-element analysis of samples with various origin.

Accuracy and Detection Limit

The accuracy of NAA depends on several factors such as the counting statistics and geometry, gamma ray self-absorption, neutron self-shielding, spectral and nuclear interferences, etc. The principal error in the analysis is the counting statistics of the total and net peak area determination. In most cases, this error is better than 5–10%. The errors due to sample preparation in NAA are usually negligible because the samples do not require any special treatment. Errors due to neutron self-shielding and gamma ray self-absorption can be neglected when using small sample masses. Spectral interferences are negligible because most spectrum analysis programs perform accurate peak deconvolution and calculation of net peak area. Nuclear interferences are important for samples with elevated uranium concentrations and include corrections for the ^{235}U fission products while analyzing lanthanum, cerium, neodymium, molybdenum, and zirconium. The accuracy of quantitative NAA of trace metals is generally less than 5–10% for most elements, depending on the concentration of the element of interest and the sample matrix.

The overall detection limits for most elements range from less than one microgram per kilogram to several hundreds of milligrams per kilogram. The lowest detection limits are observed for gold, iridium, and samarium, which are important to the exploration geology. The second group of elements includes a number of potentially hazardous substances (e.g., arsenic, antimony), the radioactive elements uranium and thorium, and some of the rare earths. The third and fourth groups include the majority of the essential and toxic elements and some precious metals. The elements in the fifth group—calcium, potassium, and magnesium—exhibit poor detection limits, but due to their usually high concentrations in geological materials, these elements are readily quantified. Some elements such as silicon, tin, and sulfur have very poor detection limits and they are usually analyzed using other techniques.

The detection limits of elements in biological materials are much lower than those obtained for soils (Lalor et al. 2000) because the matrix of biological materials consists mainly of elements that are not activated with the conventional analytical schemes (e.g., carbon, nitrogen, oxygen, and sulfur).

31.2.1.3 Summary and Conclusions

The neuron activation analysis method is currently well recognized as a referee method because of the traceability of analytical results. The uncertainty in NAA can be fully assessed because the analytical procedures including activation, decay, and measurement are mathematically well described and defined. The main advantages and disadvantages of the neutron activation methods are summarized in Table 31.4.

Table 31.4 Advantages and disadvantages of NAA

Advantages	Disadvantages
Multi-element analysis of >50 elements	Long turnaround times for multielement NAA
Nondestructive, no need for sample dissolution	Some common elements e.g. lead, sulphur and silicon cannot be determined by NAA or have poor detection limits
Analysis of solids, liquids and gases	Analysed samples should be treated as a radioactive waste
Analysis of "total" concentrations	
Low detection limits for most elements	
Dynamic concentration ranges, from μg/kg to 100%	

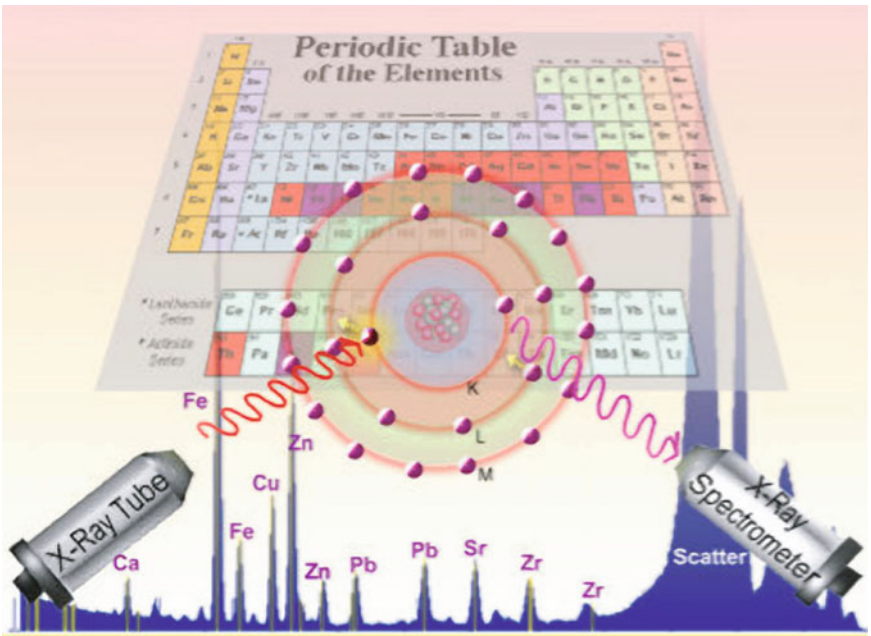


Fig. 31.8 A simplified representation of XRF

31.2.2 X-Ray Fluorescence Spectrometry

X-ray fluorescence is a spectroscopic method for multi-element analysis of solids, powders, and liquids using the X-ray fluorescence radiation of elements present in a sample. X-ray spectroscopy dates back to 1913, when Moseley photographed the X-ray spectrum of several elements and demonstrated the relationship between the atomic number of elements and the wavelength of their emission lines. This relationship not only underlined the foundation of the quantitative XRF analysis, but was also useful in the discovery of new elements in the periodic table, such as hafnium (Bertin 1975). Bragg constructed the first X-ray analytical device and received the Nobel Prize in physics in 1915. The XRF technique did not become widespread until the 1950s when Friedman and Birks built a prototype of the modern wavelength-dispersive XRF (WDXRF) spectrometer with multi-element capabilities and the sensitivity to analyze various materials. Twenty years later, the advances in solid-state detectors made possible the construction of the energy-dispersive XRF (EDXRF) spectrometers. In the early 1980s an ultratrace XRF analyzer was commercially introduced using the principle of the total reflection XRF (TXRF). Today, the laboratory XRF spectrometers can analyze all elements with atomic number $Z = 4$ (beryllium) to $Z = 92$ (uranium) in various materials. Portable XRF instruments are widely used for *in situ* analysis of lead and other heavy metal contamination; the alpha-proton X-ray spectrometer on the Mars Pathfinder was used to analyze chemical composition of Martian soil and rocks.

31.2.2.1 How It Works

Principle of the XRF

X-rays cover an approximately 1- to 100-keV region of the EM spectrum and are associated with atomic electron transitions between different shells of the atom. The primary interaction of the X-rays with matter includes photoelectric absorption and incoherent and coherent scattering. The high-energy excitation radiation interacts with inner-shell orbitals of the atom as illustrated in Fig. 31.8. If the energy of the incident radiation E_p is greater than the binding energy (E_b) of an inner-shell electron, an electron is ejected in the form of a photoelectron. The atom remains ionized for a very short time (about 10^{-14} s) until the vacancy is filled by a higher energy electron from outer orbits. The difference between the two energy levels of the transferred electron results in emission of a photoelectron of a specific wavelength or energy. This photon will either escape from the

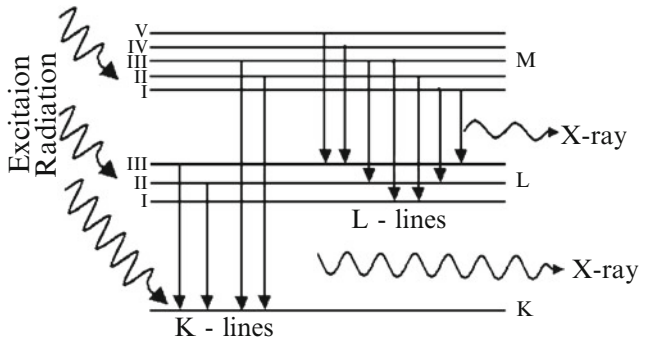


Fig. 31.9 Principal K and L XRF lines and their electronic transitions

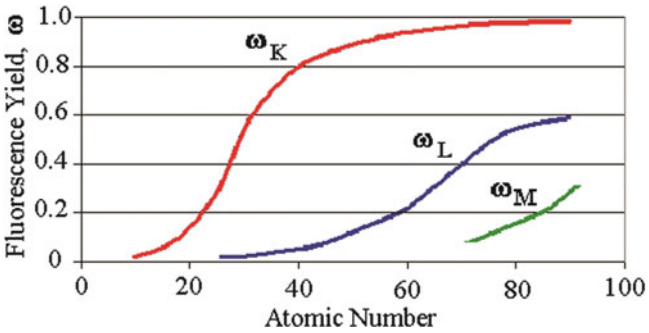


Fig. 31.10 Fluorescence yield for K, L, and M lines as a function of atomic number

atom in the form of characteristic X-rays or will be absorbed within the atom by ejection of an electron called an Auger electron.

In the stable atom, electrons occupy discrete energy levels that are labeled in order of decreasing binding energy as K, L, M, . . ., and the corresponding characteristic X-ray lines in these shells are known as the K, L, M, . . ., series of lines. Figure 31.9 shows the electronic transitions of the principal K and L emission lines used in XRF.

The probability of emission of the characteristic K, L, M, . . . lines, called fluorescence yield, increases with the atomic number of elements and decreases in order K > L > M series as shown in Fig. 31.10. Because production of Auger electrons is the only other competing reaction, the yield of X-ray photons will be primarily reduced by the Auger effect.

The X-ray scattering process occurs in two modes, with and without changes in incident photon energy or wavelength, referred to as incoherent (Compton) and coherent (Rayleigh) scattering, respectively. Compton scattering is due to collision between an X-ray photon with a loosely bound outer shell electron in which the incident photon transfers a portion of its energy to the electron and scatters at angle θ following the relationship (Jenkins et al. 1981):

$$E_{Com} = \frac{E_{exc}}{1 + \frac{E_{exc}}{511}(1 - \cos\theta)} \tag{31.4}$$

where E_{Com} and E_{exc} are the Compton and excitation photon energies in keV, respectively.

In Rayleigh scattering the incident photons interact coherently with all the electrons of the atom without loss of energy. The probability of either types of scattering depends on the X-ray energy and sample composition. Thus, the effect of Compton scattering increases with decreasing atomic number, whereas Rayleigh scattering is most prominent for low photon energies and high atomic number elements.

The absorption of X-rays through a sample with thickness *d* is described by an expression similar to Lambert-Beer law, i.e.,

$$I_d = I_o e^{-\mu\rho d} \tag{31.5}$$

where, I_o and I_d are the intensities of incident and transmitted X-rays and ρ is the density (g.cm^{-2}) of the specimen. The mass absorption coefficients μ of absorber (cm^{2}g^{-1}) is a function of the atomic number and the energy (wavelength) of the X-rays, and exhibits sharp discontinuities at particular energy levels, known as K, L, M, etc., absorption edges. It accounts for the major interactions that can occur in the absorber, i.e.,

$$\mu(E) = \tau_{ph}(E) + \sigma_{inc}(E) + \sigma_{coh}(E) \tag{31.6}$$

where $\tau_{ph}(E)$ is the photoelectric mass absorption coefficient and $\sigma_{inc}(E)$ and $\sigma_{coh}(E)$ are the incoherent and coherent mass scattering coefficients.

The mass attenuation coefficient μ_{mix} for a multicomponent mixture of pure elements is calculated as the weighted average of the absorption coefficients μ_i of the different constituents i, i.e., $\mu_{mix} = \Sigma(W_i \cdot \mu_i)$, with W_i = weight fraction of component i.

X-Ray Fluorescence Spectrometry

X-ray fluorescence works by irradiation of a sample with a primary EM radiation generated with radioisotopes, X-ray tubes, or charged particles (electrons, protons, and alpha particles), which in turn causes an element to emit characteristic X-ray lines. By measuring the intensities of the X-ray fluorescent lines against their wavelengths or energies, one can identify and quantify the concentrations of elements in a sample.

A typical X-ray fluorescent spectrometer includes an excitation source and apparatus for separation, detection, and measurement of the fluorescent X-rays. The intensities of the secondary X-rays are typically reduced by two to three orders of magnitude compared to the primary excitation radiation. Therefore, to perform an XRF analysis a high-intensity excitation source and a sensitive X-ray detector are necessary. There are two common types of spectrometers in use, the wavelength-dispersive (WD) and energy-dispersive (ED) XRF. The term dispersion here refers to different methods of separation and measurement of the characteristic

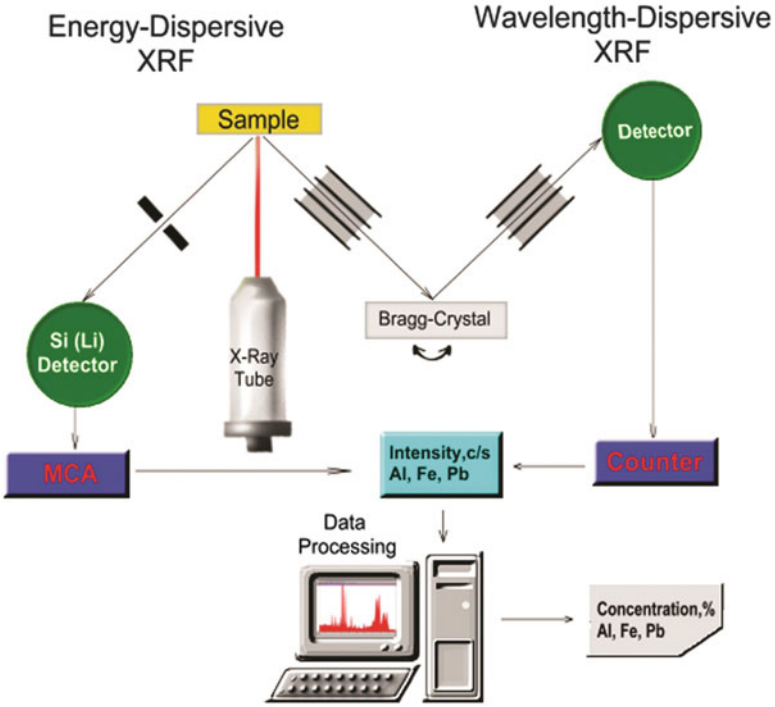


Fig. 31.11 Schematic diagram of a simplified wavelength-dispersive and energy-dispersive XRF spectrometer

X-ray lines of elements by means of wavelength or energy dispersion.

X-ray fluorescence generally uses two types of excitation sources: radioisotopes with discrete gamma or X-ray emission lines, and sources of continuous radiation generated by X-ray tubes. WDXRF spectrometers use X-ray tubes, whereas the EDXRF spectrometers employ both radioisotope and X-ray tube excitations. The basic components of a typical wavelength-dispersive and energy-dispersive XRF spectrometer are shown in Fig. 31.11.

The X-ray tube with selected anode materials (chromium, molybdenum, tungsten, etc.) excites the characteristic lines of elements present in the sample, which are then separated and dispersed into a pulse-height distribution spectrum called a characteristic X-ray spectrum. Collimators are placed on both optical paths of the sample, including the primary (excitation) and secondary (fluorescent) radiation. In WDXRF the separation of wavelengths of secondary X-rays is done by Bragg's diffraction from natural or synthetic crystals according to the equation

$$n\lambda = 2d.\sin(\theta) \tag{31.7}$$

where n is the diffraction order, λ is the wavelength of the characteristic X-ray, d is the lattice spacing of the analyzing crystal, and θ is the angle of incidence.

By rotating a crystal with specific d spacing at an angle θ along with the detector and associated collimators at diffraction angle (2θ), it is possible to separate and measure the characteristic wavelengths of different elements. To cover the entire range of X-ray wavelengths of elements, different diffraction crystals are usually used such as lithium fluoride (LiF), quartz (SiO_2), penta-erythritol (PET), NH_6PO_3 (ADP), and acid phthalate (KAP). The intensity of the wavelengths are detected by either NaI(Tl) scintillation or flow-proportional counters or both. Gas-flow proportional detectors are filled with a mixture of two gases such as helium, argon, krypton, or xenon and used for long-wavelength X-rays, but the scintillation detector is used for counting of short wavelengths.

In energy-dispersive XRF spectrometry, the wavelength-dispersive crystal-detector system is replaced by a solid-state energy-dispersive detector that can simultaneously separate and detect all X-rays from a sample. The common solid-state detectors used in XRF include lithium-drifted silicon semiconductor detectors Si(Li), HgI_2, and silicon pin diode. Among those detectors, the Si(Li) has the best resolution (less than 150 eV) but needs to be operated at $-273°C$ to maintain the diffusion of lithium ions. The Si(Li) semiconductor detector detects the X-ray photons, which are then converted by an ADC and an MCA directly into an X-ray spectrum.

Compared to wavelength-dispersive crystal-detector systems, the solid-state detectors in EDXRF have much higher dispersive efficiency because they perform direct separation and measurement of the secondary X-ray spectrum. This makes possible the use of low-powered X-ray sources, such as radioisotopes or X-ray tubes with secondary targets. The most common radioisotope sources employed in X-ray spectrometry are ^{55}Fe, used for analysis of elements

with atomic number $12 < Z < 23$; ^{109}Cd, for analysis of elements with $20 < Z < 42$; and ^{241}Am, for elements with $47 < Z < 58$. The secondary-target excitation mode allows modification of broadband radiation of an X-ray tube to narrow band energy, suitable for efficient excitation of a particular group of elements.

The conventional WDXRF measurement electronics are generally referred to as a single-channel pulse-height analyzer in contrast to the MCA used for EDXRF spectrometry. Thus, the wavelength-dispersive XRF measures the intensity of the characteristic X-ray lines of elements in sequential mode, whereas the EDXRF records the entire X-ray spectrum at once. The EDXRF spectrum is further processed with a computer program that performs background subtraction, correction for escape peaks and other interferences, and calculates the net peak intensities of elements. In both cases, wavelength- and energy-dispersive XRF spectrometry provide information about the intensities of the characteristic X-ray emission lines, which is used to calibrate the equipment and calculate the concentrations of unknown samples.

31.2.2.2 What It Does

The X-ray excitation sources used in wavelength- and energy-dispersive XRF provide adequate analytical sensitivity for quantitative analysis across the elements sodium to uranium. The elements of interest are identified by their spectral wavelengths or energies for qualitative analysis and the intensities of the emitted spectral lines are used for quantitative analysis. The quantity of the characteristic X-rays is proportional to the amount of elements in the sample. The quantitative XRF analysis of unknown samples is usually performed using calibrations with matrix-matched standards.

Sample Preparation

Most solid or liquid samples can be analyzed by XRF as received, without any special sample preparation. However, for quantitative analysis the samples and standards must be homogeneous with an optically flat surface. Table 31.5 summarizes the most common specimen preparation procedures for solids, liquids and gases.

Solid geological and biological samples are previously oven- or freeze-dried and milled to a fine powder (<200 mesh) before preparation for XRF analysis. For the pressed powder pellet about 3–7 g of sample is homogenized with a binding agent and pressed into a solid pellet using a hydraulic press. For preparation of the glass discs the sample material is weighed and fused at a temperature of 900–1,200°C using a suitable flux, e.g., lithium tetraborate.

Sample preparation is a critical step in XRF analysis. By choosing the most appropriate sample preparation procedure (e.g., fusion, pelletizing, etc.) some problems associated with sample homogeneity, particle size effects, nonlinear working range of concentrations, etc., could be greatly overcome (Buhrke et al. 1997).

Table 31.5 Some common specimen preparation methods for XRF analysis

Sample type	Sample preparation
Powders (geological)	Bulk sample in a sample cup
	Press pellet
	Fused glass beads
Powders (biological)	Bulk sample in a sample cup
	Press pellet
	Dry ashing and pelletising
Metals	Cutting, grinding and polishing
Liquids	Measurement in a sample cup
	Pre-concentration
	Filtering through a 0.45 μ filter
Air particulates	Aspiration through a filter paper
	Dust wipes with filters

Qualitative XRF Analysis

The qualitative XRF analysis is an important step in geochemical studies, which makes it possible to determine the overall composition of samples. The problems associated with qualitative analysis depend on the X-ray dispersion method. The qualitative WDXRF analysis is performed by stepwise scanning over the entire wavelength range by rotating the crystal detectors in small increments of angle θ. Modern WDXRF software includes advanced peak search routines to analyze the peak shape and to locate the true position of the peaks. The EDXRF spectrometer simultaneously collects the X-ray spectrum of elements for a short period of time; however, to cover the entire range of elements, usually two to three different excitation conditions and measurements are applied.

There are several interferences common to wavelength and energy-dispersive XRF that need to be considered in qualitative analysis. These include the incoherent and coherent scattering, K, L, and M line overlaps, tube emission lines, background distribution, higher order of diffraction lines in WDXRF, and the presence of escape and sum peaks in EDXRF. Figure 31.12 shows the X-ray spectrum of a typical Jamaican soil sample with characteristic peaks and scattered radiation (Johnson et al. 1996).

Quantitative XRF Analysis

Quantitative XRF analysis using wavelength and energy-dispersive spectrometry involves a series of steps designed to prepare, measure, and process the X-ray spectral data. Table 31.6 outlines the basic procedures used to perform a quantitative XRF analysis.

Although the measurement principle of the wavelength and energy-dispersive XRF spectrometers differ, the

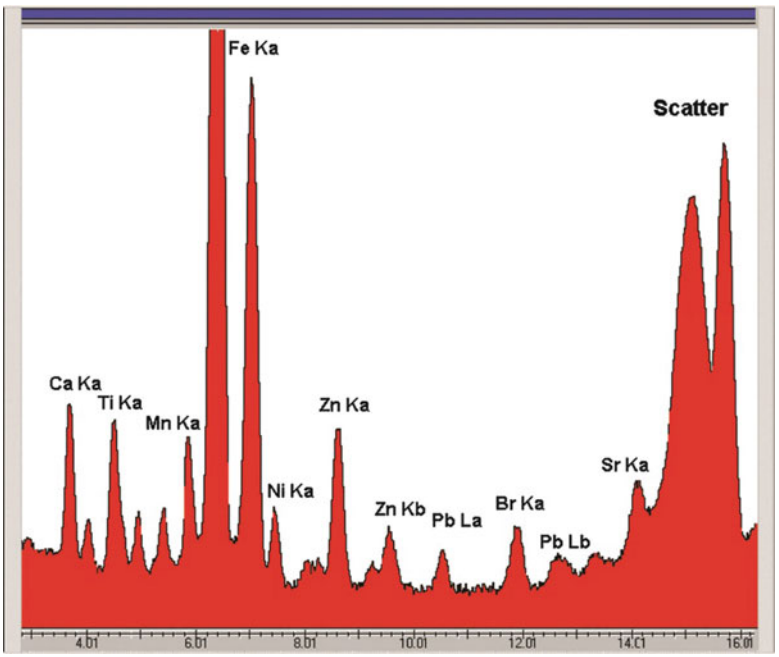


Fig. 31.12 A typical X-ray spectrum of Jamaican soil sample with characteristic X-rays and backscattered radiation

Table 31.6 Basic steps for quantitative X-ray analysis

Step	Procedure
1. Preparation	Perform qualitative analysis of unknown samples
	Select calibration standards with composition similar to unknowns
	Prepare samples and standards using similar procedures
2. Measurement	Measure the samples and standards under identical operating conditions
3. Data processing	Perform spectrum analysis to obtain the net intensities of elements
	Perform quantitative calibration for each element
	Calculate concentrations of unknown samples
	Carry out appropriate quality control checks

quantitative analysis process, converting the intensities into elemental concentration, is similar for both techniques.

In the case of monochromatic excitation of homogeneous samples, the relationship between the concentration C_i of an element i and intensity I_i can be generalized in the form

$$C_i = K_i \,.\, I_i \,.\, M_i \tag{31.8}$$

where K_i is a sensitivity factor that depends upon instrumental parameters and physical constants of the analyte i, and M_i is a matrix correction factor accounting for the absorption and enhancement of the primary and secondary radiation.

Equation 31.8 shows that the concentration of an element can be determined by measuring the intensity I_i of the analyte and by calculation of K_i and M_i factors from standards. The accuracy of the analytical results to a large extent depends on the efficiency of the interelement correction factor M_i. The interelement or matrix interferences result from variations in the physical and chemical properties of the sample, such as particle size, mineralogy, chemical composition of major constituents, etc. The most common matrix effects encountered in XRF are the absorption and tertiary fluorescence or enhancement of X-rays, which can be corrected by using appropriate quantitative analysis methods (Jenkins et al. 1981; Tertian and Claisse 1982).

Thin-Film Calibration

In thin-film samples the interelement effects associated with the absorption and enhancement effects are reduced significantly because each atom interacts with the primary and secondary (characteristic X-ray) radiation independently. A thin specimen is defined as satisfying the relationship $\rho d < 1\ \mu^{-1}$ (Eq. 31.8), where ρd is the mass thickness of the sample per unit area. The intensities of the analyte lines for thin samples are directly proportional to analyte concentrations.

Empirical Calibration Method

The empirical calibration is performed using matrix-similar standards or samples already analyzed by other techniques. The standards should include the elements of interest and cover the expected concentration range of elements. Calibration curves are typically generated using 10–15 calibration standards. The calibration coefficients are calculated through a least-squared regression analysis with a correlation coefficient greater than 0.95.

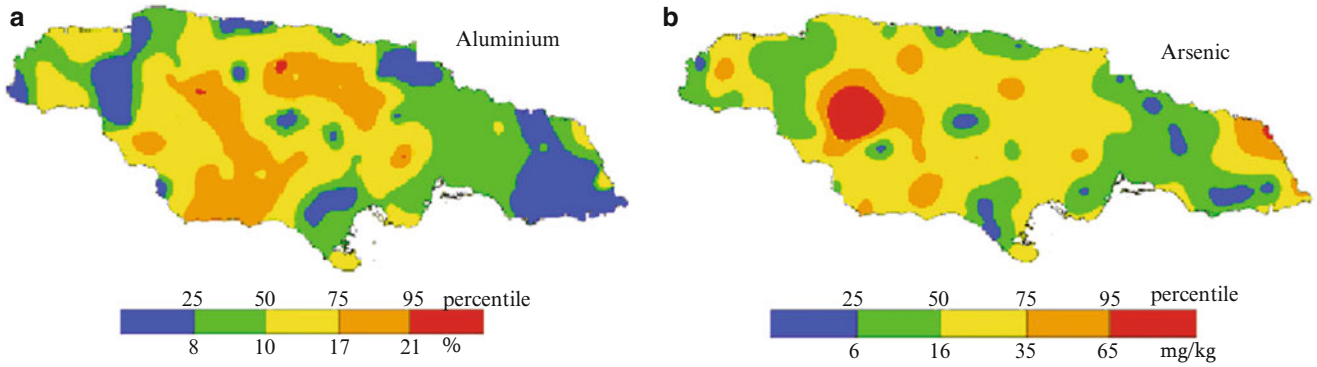


Fig. 31.13 (**a** and **b**): Map of aluminum and arsenic distributions in Jamaican soils

Dilution Methods

The sample dilution method is used to minimize the matrix differences between samples and comparator standards. This can be accomplished by adding either an element with a very high (e.g., lanthanum, barium, etc.) or a very low (e.g., lithium borate) absorption coefficient. This changes the overall mass absorption coefficient of the sample to a level that the effect of matrix elements on the intensity of the analyte line are negligible. Dilution methods also allow for obtaining linear calibrations over wider concentration ranges.

Matrix Correction Methods

The common approach for compensating the matrix effect includes the use of the incoherent (Compton) scattering of the primary radiation, which is measured in each XRF spectrum. The Compton method applies the inverse relationship between the Compton peak intensity and the mass absorption coefficient of the sample for correction of the matrix effects. The ratio of elemental intensities to the Compton peak is used to generate calibration curves. An alternative approach for matrix correction is the internal standard method, i.e., addition of an element to a sample that is affected in the same way as the analyte element by the matrix. The ratio of the line intensities of analyte element to the internal standard will then be independent of the matrix.

Mathematical Methods

Mathematical correction methods were introduced in an attempt to develop absolute methods for quantitative XRF analysis. In 1955 Sherman published the fundamental relationship between X-ray intensities and sample composition, but due to the lack of adequate computational power this method did not find practical application. Bettie and Brissey used the regression analysis to solve equations relating measured intensities of elements with corresponding concentrations. Other approaches include the intensity-correction model of Lucas-Touth and Price, the alpha-correction model of Lachance and Trail, the Rasberry and Heinrich enhancement correction model, and the fundamental parameters method.

31.2.2.3 Summary and Conclusions

Compared to NAA methods, XRF has the advantage of being a cost-effective technique providing good detection limits across the elements of the periodic table and applicable to a wide range of concentrations from milligram per kilogram to 100%. WD and EDXRF are not the methods of choice for sub-trace elements, nor are they for those in which calibration standards with similar matrix composition cannot be obtained.

The choice of a wavelength- or energy-dispersive spectrometer depends on several factors, such as cost, speed, and application type. Thus for high throughput quantitative analysis of a limited number of elements where the initial cost is not important, the WD XRF instruments are probably the best choice, whereas for multi-element analysis with some limitations in the detection limits or accuracy, the EDXRF is preferable.

31.2.3 Applications of NAA and XRF

Neutron activation and XRF analysis are widely used in environmental geochemistry and health. Some examples are shown below.

31.2.3.1 Geochemical Mapping of Soils

Surface soil samples were collected on an 8 × 8 km grid across Jamaica which resulted in some 204 samples, including field duplicates. Samples were prepared by wet sieving to −150 μ, drying at 80°C, and grinding to less than 50 μ. The soil samples were analyzed by NAA and XRF for about 40 elements and the results were used to produce geochemical maps of soils (Lalor 1995).

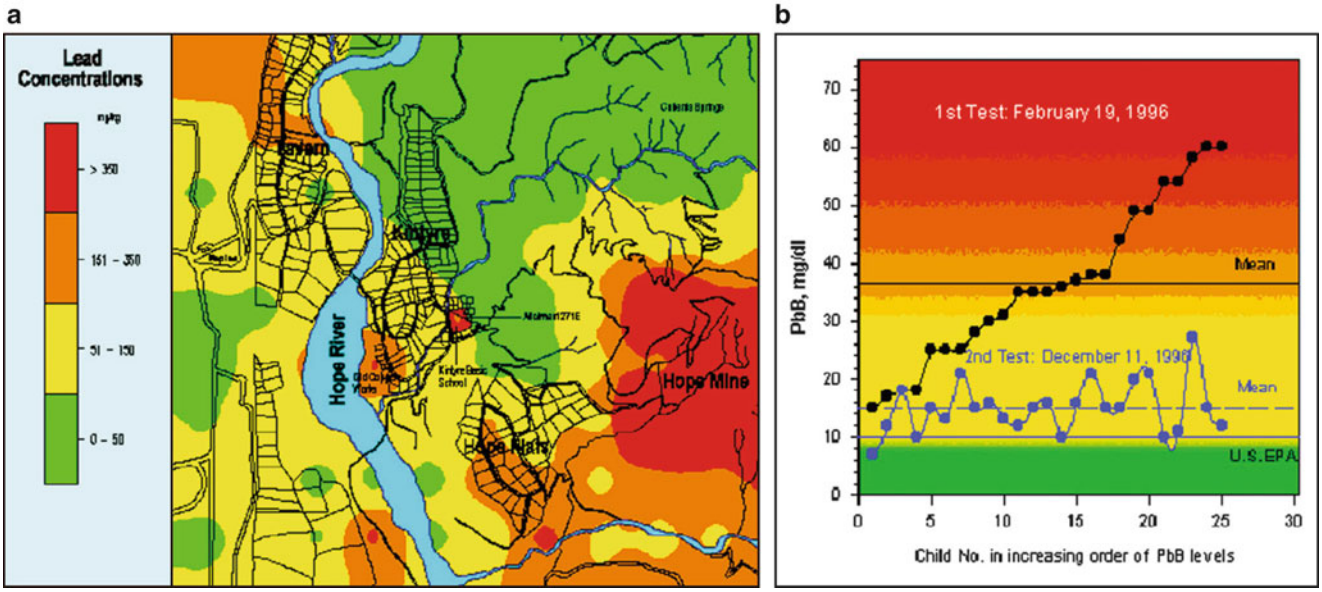


Fig. 31.14 (**a**) Map of lead distribution in contaminated soils; (**b**) blood levels of children before and after remediation

Figure 31.13a shows the geochemical map of aluminum in Jamaican soils. The highest concentrations of aluminum in soils (Al > 15%) overlap the areas of known bauxite deposits. As this is also an agricultural area the soil map provides useful information on possible aluminum intake by plants.

The arsenic distribution map in Fig. 31.13b shows an anomalous area with arsenic concentrations as high as 373 mg kg^{-1}. A followup study of the "hot spot" defines the boundaries of the anomalous area and indicates that under the present land use conditions arsenic does not present environmental hazard problems (Lalor et al. 1999).

31.2.3.2 Lead Pollution and Health

EDXRF spectrometry was used to provide data for mapping of lead contamination from an old mine. Figure 31.14a shows the distribution of lead in soils. The highest lead levels were around a school for preschool children 3–6 years old. Blood lead (PbB) screening of the children showed a mean level of 37 µg dL^{-1}. The environmental intervention included isolation of lead contamination by means of paving, nutritional supplements of food rich in iron and calcium, and a lead-safe education campaign for teachers, parents, and children. The upper curve in Fig. 31.14b shows the blood lead levels of children sorted in increasing order. Ten months after intervention the PbB levels of children were greatly reduced as shown by the lower curve in the same figure (Lalor et al. 2001). Five years later a new generation of children attending this school showed a mean PbB level of 8.3 µg dL^{-1}. This illustrates the effectiveness of the mitigation campaign undertaken in the lead-contaminated area.

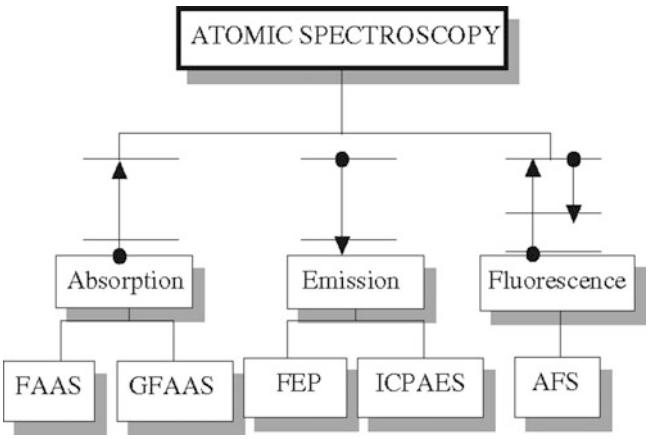


Fig. 31.15 Block Diagram of atomic spectroscopy techniques

31.3 Inorganic Geochemistry Techniques for Liquids

31.3.1 Introduction

There are a number of methods used to analyze inorganic elements in liquid samples. The most widely used are atomic spectroscopy (AS), mass spectroscopy (MS), and electrochemical techniques (ET). According to the principle of interaction of atoms with light, the AS methods are classified as absorption, emission, and fluorescence (Fig. 31.15).

- Atomic absorption spectroscopy (AAS) measures the intensity of absorbed radiation of atomic species. Two AAS techniques are commonly used: flame (FAAS) and graphite-furnace (GFAAS).

- Atomic emission spectroscopy (AES) involves excitation of the analyte atoms and measurement of their emission as they revert to their ground state. The classical emission spectroscopy, known as flame emission photometry (FEP), has been further developed to accommodate multi-element capabilities using inductively coupled plasma (ICP) as the atomization source.
- Atomic fluorescence spectroscopy (AFS) is based on measurement of radiation given off as excited atoms relax to a state higher than the ground state. Even though AFS is the most sensitive of the above techniques, it has not gained widespread use for routine analysis, because it is based on complex and expensive instrumentation technology.

Among the different spectroscopy techniques for liquids, the most commonly used are AAS, ICP-AES, and ICP-MS, each of which has particular advantages and disadvantages. ICP mass spectrometry (ICP-MS) utilizes hot plasma as an ionization source and has been extensively developed as a multi-element spectroscopic technique capable of analyzing both elemental concentrations and isotopic ratios. Along with modern spectroscopic techniques, classical electrochemical methods are increasingly used due to their exceptionally low detection limits for many elements and possibility of *in situ* analysis.

31.3.1.1 Sample Dissolution

Liquid analytical techniques require prior dissolution of the solid samples using appropriate digestion procedures. The choice of a particular dissolution technique depends on the sample type, the elements of interest, and their concentration levels. Metals in biological materials can be digested more easily than those in geological matrices. Water samples usually require relatively minimal treatment prior to analysis.

31.3.1.2 Geological Materials

The decomposition of geological samples is commonly classified as selective, partial, or total, according to the mineral composition of the sample and analytical requirements.

- *Selective extraction* is used to extract specific elements leaving the remaining sample matrix in the residue. This procedure usually involves single or sequential attack of the sample by weak acids. The most popular procedure involves the use of sodium acetate, which dissolves only alkali metals and carbonates. Sodium pyrophosphate is also used for liberating organically bound heavy metals.
- *Partial digestion* is usually accomplished using aqua-regia (3:1 mixture of concentrated hydrochloric and nitric acids). Attack with aqua-regia causes loosely bound metals to enter into solution while those embedded in mineral particles remain in the solid residue. Generally, about 30 elements can be digested using aqua-regia, with approximately half of them fully dissolved.
- *Total digestion* and analysis of minerals is important for geological exploration and industrial purposes. It usually involves oxidation of the organic matter with nitric acid, followed by hydrofluoric acid (HF) digestion, which attacks the silicate matrix. The excess silicon is removed during evaporation of volatile silicon fluoride (SiF_4). For digestion of easily volatilized elements closed digestion vessels are required.

Certain minerals such as barite and some refractory elements (e.g., zircon) will dissolve only partially under these conditions. In such cases, sodium peroxide (Na_2O_2) decomposition is required, which provides basic chemical attack. Generally, samples with high sulfide content require higher reaction temperatures for full dissolution, which can be achieved using microwave high-pressure bomb digestion (Kingston and Jassie 1988). This method, however, cannot be used for determination of rare earth elements (REE) and some other resistant minerals. The only attack in which REE, major oxides such as silicon dioxide (SiO_2), and refractory minerals in geological samples are fully dissolved is lithium metaborate/tetraborate fusion. This technique involves preliminary drying of the sample at 105°C, followed by an attack of the geological matrix with a high-temperature (about 1,050°C) molten flux of lithium metaborate/tetraborate. The fusion process produces a homogenous amorphous solid solution (bead), which then may be easily dissolved in dilute acid and used for analysis.

31.3.1.3 Biological Materials

The inorganic content of most biological materials is a minor constituent and in most instances the organic matter is removed prior to analysis. Organic matter is eliminated by oxidation of carbon and hydrogen to carbon dioxide and water, which is done by heating in the presence of oxygen (dry ashing), or by the action of oxidizing acids (wet ashing), such as sulfuric, nitric, and perchloric acids. The dry ashing procedure is usually carried out at about 450–600°C for 4–12 h. The inorganic residue (ash) is easily dissolved and used for analysis. This procedure is not applicable for some volatile elements such as arsenic, selenium, lead, and cadmium for which closed vessel microwave digestion in Teflon containers is recommended. Samples intended for analysis of mercury are oven dried at 60°C or freeze-dried.

31.3.1.4 Water Samples

Water samples are usually analyzed without sample digestion, unless they contain high concentrations of suspended solids. The standard methods of water analysis require prior filtration of the sample through a 0.45-μ filter. Freshwater

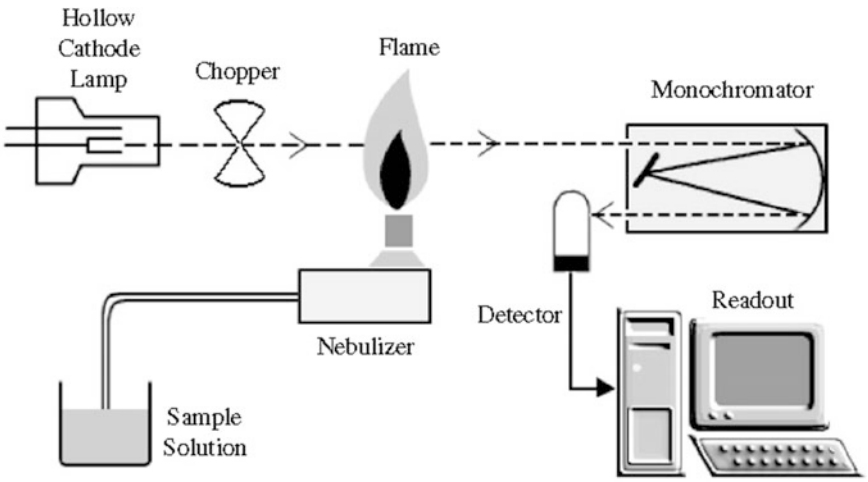


Fig. 31.16 General setup of FAAS

samples are usually filtered and acidified in the field to prevent loss of metals. Wastewater samples generally require digestion, as they contain high concentrations of metals in a variety of matrices.

31.3.2 Atomic Absorption Spectroscopy

AAS is the most frequently used method for trace element analysis of liquid samples. This method involves an introduction of an atomized sample into a beam of appropriate EM wavelength, which causes the atoms to absorb light and become excited. The measured absorption of the light beam is compared with that of known calibration standards, and the concentration of the elemental species is obtained.

31.3.2.1 How It Works

Historically, two main types of AAS instruments have been developed, FAAS and GFAAS. These two AAS techniques have similar measurement principles, but they differ in the methods of sample introduction and atomization. The analysis is normally completed in a single element mode because of the limitation imposed by the excitation source (Welz and Sperling 1999).

Flame Atomic Absorption Spectroscopy

FAAS was developed by the Australian scientist Alan Walsh in the mid-1950s (Walsh 1955). FAAS is theoretically applicable to most metals, provided that a light source is available for that element. A schematic representation of an FAAS system is shown in Fig. 31.16.

The liquid sample is initially converted to a fine aerosol by a nebulizer, which increases the surface area of the solution sample and facilitates evaporation and volatilization. The sample aerosol is transported to the flame, where most of the metal ions are reduced to elemental atoms. Air-acetylene and nitrogen oxide-acetylene are the most commonly used flame mixtures. The universal source of EM radiation used in AAS is the hollow cathode lamp, which consists of a cathode filament made of the element to be analyzed and filled with argon or neon gas. When a high voltage is applied, the lamp emits characteristic radiation of the analyte element. To isolate the radiation of the lamp from the radiation emitted by the atoms in the sample cell, a mechanical device (chopper) located between the light source and the sample cell is used.

The radiation from the burner is directed to the monochromator or wavelength selector, which isolates the resonant spectral line from the background radiation of the sample. The spectral line is measured by a detector, usually a photomultiplier tube, which converts incident EM energy into an electrical signal. Modern instruments are operated by a computer, which acts as an output device and is used to process the analytical information.

Graphite Furnace Atomic Absorption Spectroscopy

GFAAS was introduced in the early 1970s as a method with greatly improved sensitivity compared to FAAS. It is typically applicable to most elements analyzed by FAAS. A schematic representation of a GFAAS system is shown in Fig. 31.17.

GFAAS instrumentation uses an electrically heated polycrystalline furnace, mounted in the graphite tube, which allows the sample to be heated by radiant heat rather than convective energy. This delays atomization and minimizes some non-spectral interferences. The electrothermal atomization allows almost full atomization of the sample, which makes this technique more sensitive than FAAS. The sample is placed directly on the platform of the graphite furnace, which is heated stepwise to dry the sample, ash the organic matter, and vaporize the atoms. At the end of the heating cycle, the graphite tube is externally flushed with argon gas to prevent the tube from burning away.

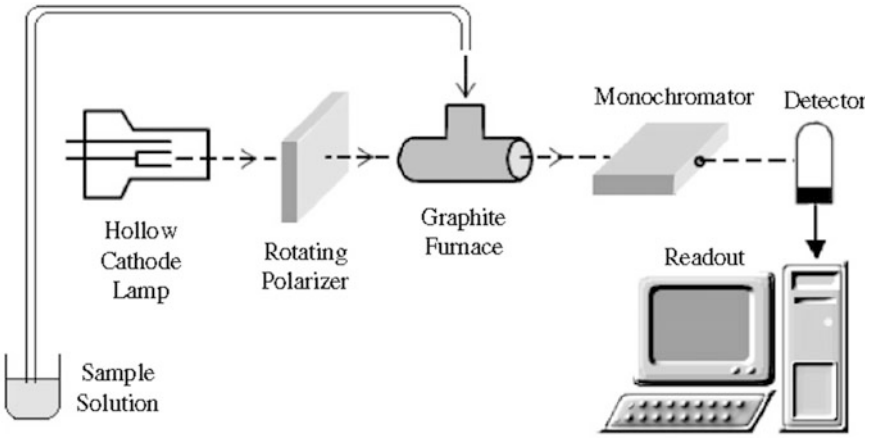


Fig. 31.17 General setup of GFAAS

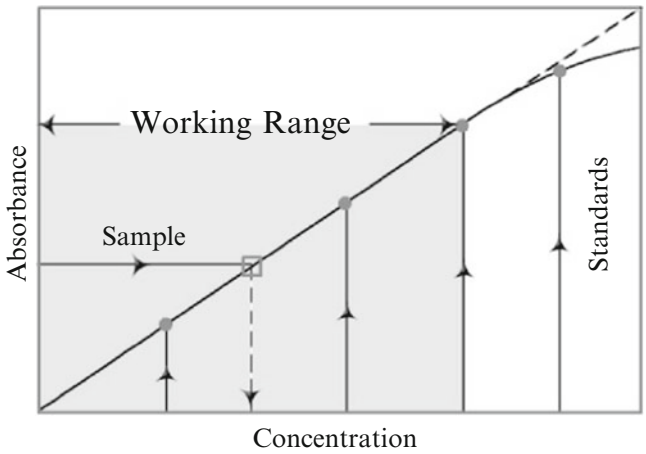


Fig. 31.18 A typical calibration curve used for AAS analysis

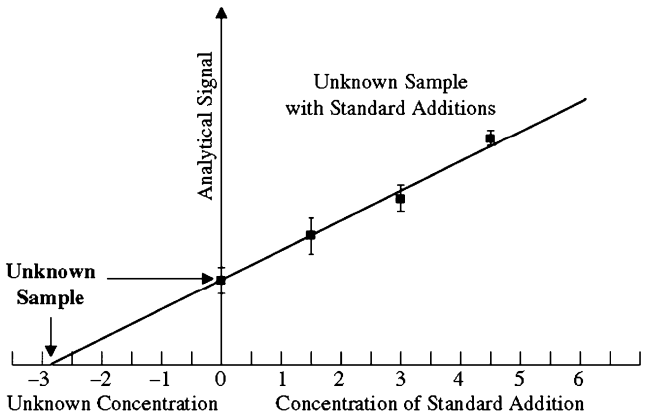


Fig. 31.19 Standard addition calibration curve

31.3.2.2 What It Does

Flame and graphite AAS techniques are primarily used for chemical analysis of metallic elements in a variety of materials. The quantitative analysis can be accomplished by using a calibration curve or standard addition method. The choice of quantification method depends on the sample type, potential interferences, and matrix effects.

A calibration curve is a plot of measured analytical signal as a function of known concentrations of standards. Typically three standards, covering the expected concentration range of the sample, as well as a blank, are used to produce a working calibration curve (Fig. 31.18). The working range is defined within the linear relationship of the calibration curve. If the sample concentration is too high, there are three alternatives that can help bring the absorbance into the linear working range; namely sample dilution, alternative wavelength with a lower absorptivity, or reduction of the path length by rotating the burner head.

The standard addition method is usually applied to quantification of samples that have substantial matrix effects. In this method, a known amount of analyte is added (spiked) to an aliquot of the sample. The concentrations and absorbance of the original and spiked samples are plotted on a graph, which is used to determine the concentration of the unspiked sample, as shown in Fig. 31.19. Because the sample aliquots differ only in analyte concentration and have relatively no matrix differences, this method is often referred to as an "internal" calibration method, and it is more accurate than using external calibrations.

There are five major types of interferences encountered in atomic absorption spectroscopy: physical, chemical, spectral, ionization, and background. The physical, chemical, and ionization interferences influence the atomization process, whereas spectral and background interferences change the atomic absorption signal.

- *Physical* characteristics of dissolved samples, such as viscosity or density, may influence the aspiration rates or droplet size and therefore the measurement signal. This may change the slope of the calibration curve and produce lower or higher analytical results.
- *Chemical* interferences occur when the atoms are not completely free or in their ground state. This may be due to the presence of certain chemical compounds in a

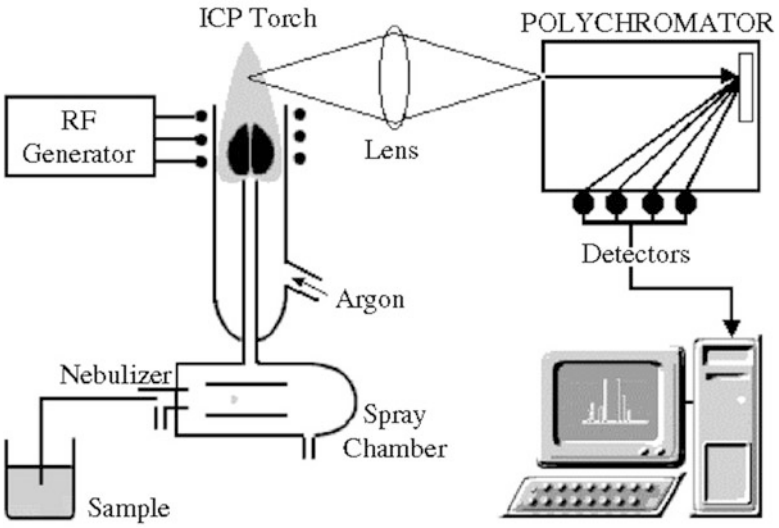


Fig. 31.20 Schematic diagram of simultaneous ICP-AES

sample matrix, which can change the sensitivity and the response between samples and standards.

- *Spectral* interferences exist when an emission line of atomic or molecular species other than the element being analyzed is present within a band pass of the wavelength that the analyte absorbs.
- *Ionization* interferences occur when the flame or furnace causes complete removal of atomic electrons. The ionized atoms of the analyte element will not absorb light, thereby lowering the concentrations of that element.
- *Background* interferences are due to undissociated molecules of the matrix that can scatter or absorb the light. In either case, the apparent concentration of the analyte will be much higher. Most GFAAS perform automatic background corrections using the Zeeman method.

The common procedures used to correct for interferences include: (1) serial dilution and re-analysis of the sample to determine whether the interference can be eliminated, (2) use of matrix modifiers to compensate for potential interferences, and (3) use of standard addition calibration method (Schlemmer and Radziuk 1999).

31.3.3 Atomic Emission Spectroscopy

Atomic emission spectroscopy uses optical emission of excited atoms for quantitative elemental analysis. The earliest emission technique, flame emission photometry, was limited mainly to sodium and potassium, as most atoms of non-alkali metals remain in their ground state at normal flame temperatures. The emission source should ideally be able to vaporize the sample, breakdown all compounds, and excite the atoms and ions. To achieve this, temperatures much higher than those of the flame are required. Thus, the use of alternative atomization sources, such as electrical arcs and plasma, are required. ICP has revolutionized AES by making it applicable to a wide range of elements. Because all atoms in the sample can be excited at the same time, they can be detected simultaneously, giving ICP-AES multi-element capabilities.

31.3.3.1 How It Works

ICP-AES was introduced in 1975 as a multi-element trace element technique utilizing argon plasma for atomization and excitation. The ICP source consists of a torch with three concentric quartz tubes through which ionized argon gas flows at high speed, reaching temperatures between 6,000 and 10,000 K (Fig. 31.20). The liquid sample is introduced into the plasma in the form of an aerosol. The high temperature of the plasma excites the atoms and they emit light of characteristic wavelengths. Because emission lines are quite narrow, there are less chances of interelement overlaps, and multiple elements can be determined simultaneously. A polychromator, which isolates all wavelengths of the analyte, is used in conjunction with multiple detectors that measure the wavelengths and intensities of analyte lines.

31.3.3.2 What It Does

ICP-AES is mainly used for analysis of dissolved/suspended samples, as solid samples require costly introduction techniques such as laser ablation. The element range spans about 60 elements, including some non-metals such as sulfur and some halogens, which cannot be analyzed by AAS. The calibration curves are generally linear in wide dynamic ranges. Due to the high atomization temperatures, even most refractory elements—such as boron, phosphorous, tungsten, uranium, zirconium, and niobium—are atomized efficiently. As a result, the minimum detection limits of these and most other elements can be orders of magnitude lower with ICP-AES than with FAAS.

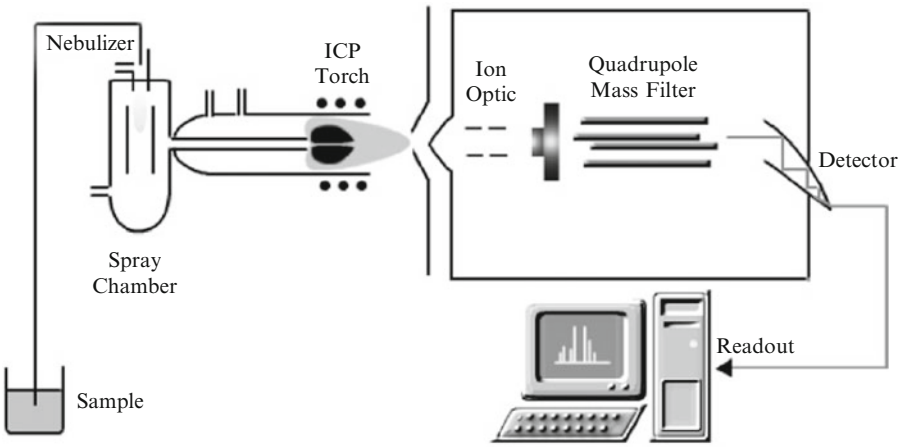


Fig. 31.21 Schematic diagram of an ICP-MS

Spectral interference problems are fairly common in ICP-AES due to the line-rich spectra produced by the hot plasma source. High-resolution spectrometers can be used to minimize spectral interferences or an alternative analyte emission line can be used to detect the element. Chemical matrix effects are not as profound in ICP-AES as in other techniques, and they can be moderated by use of internal standards. Detection limits are typically less than 1 μg L^{-1}, which is about at least ten times higher than those of obtained by GFAAS.

31.3.4 Inductively Coupled Plasma Mass Spectrometry

ICP-MS was developed in the early 1980s and has become increasingly popular for the analysis of geological, environmental, and medical materials. The ICP-MS technique uses high-temperature plasma discharge to generate positively charged ions, which can be separated and quantified based on their mass-to-charge ratios by means of a mass spectrometer. Like AES, atomic mass spectrometry is inherently a multi-element analytical technique, which can also provide information about isotopes present in the sample. Conventional ICP-MS analyzes liquid samples, but techniques for direct analysis of solid samples such as laser ablation ICP-MS have also been developed.

31.3.4.1 How It Works

The main steps in ICP-MS operation involve (1) ionization of the liquid or solid material introduced into the ICP using a stream of argon carrier gas, (2) separation of the ions based on their mass-to-charge ratio using a mass spectrometer, and (3) measurement of the relative quantities of ions by an electron multiplier detector and counting system. A schematic diagram of an ICP-MS is shown in Fig. 31.21. It consists of a sample introduction system, an excitation source, ion transport system, mass separation device, detector, and computer.

The sample introduction system includes a nebulizer and a spray chamber. The nebulizer converts liquid samples into a fine aerosol by mixing it with argon in a spray chamber. The primary function of the spray chamber is to isolate fine aerosols to enter into the plasma. The sample aerosol is injected into the middle of the high-temperature plasma (5,000–10,000 K), where it is vaporized, atomized, and ionized. Recombination of sample, matrix, and argon ions may occur in the cooler parts of the plasma, which leads to the formation of various molecular species.

The ions produced in the central part of the argon plasma are transferred into a mass spectrometer by an ion transport system consisting of an interface region and an ion lens system. The interface region extracts the ions from the plasma and directs the ion beam to the front section of the mass spectrometer chamber using rotary and turbomolecular vacuum pumps. The ion lens focuses and collimates the positive ions from the sample ion beam into the mass spectrometer, which separates them according to their mass-to-charge ratio. The most frequently used mass separation device is a quadrupole mass spectrometer, which consists of four parallel conducting rods to which radiofrequency (RF) and direct current (DC) potentials are applied. By changing the RF and DC potentials on the quadrupole rods, the mass spectrometer is capable of sequential scanning across the mass-to-charge ratio range of all naturally occurring elements. The ions sorted by the mass spectrometer are detected and measured by a photomultiplier tube detector.

The computer is the final constituent of an ICP-MS instrument. It controls all hardware components of the equipment, sample handling, measurement, data acquisition, and processing of analytical information.

Table 31.7 Summary of the key advantages and disadvantages of atomic and mass spectroscopy techniques

	FAAS	GFAAS	ICPAES	ICPMS
Elements analysed	60+ metals	50+ metals	70+ metals, some non-metals	75+ metals, non-metals
Multi-element	No	No	Yes	Yes
Sample throughput	Fast (for <5 elements per sample)	Slow (3–5 min. per element)	Fast (multi- element analysis)	Fast (multi-element analysis)
Semi-quantitative analysis	No	No	Yes	Yes
Isotopic analysis	No	No	No	Yes
Detection limit	Good	Excellent	Very good	Excellent
Dynamic range	10^3	10^2	10^5	10^5–10^8
Precision	<1 %	<5 %	<2 %	<3 %
Sample volumes	Large	Small	Small	Small
Dissolved solids	<5 %	<20 %	<20 %	<0.5 %
Interferences				
Spectral	Very few	Very few	Many	Few
Chemical	Many	Many	Few	Some
Physical	Some	Very few	Very few	Some
Method development	Easy	Difficult	Moderately easy	Difficult
Ease of use	Very easy	Moderate	Easy	Moderate
Capital and running costs	Low	Medium	High	Very high

31.3.4.2 What It Does

In ICP-MS, the intensities of the analytes are measured by continuous scanning of the mass spectrometer, which generates spectra of signal intensity versus mass. The peak positions of the spectra are used for identification of the elements present in a sample, whereas the signal intensity is used for quantitative analysis. Semi-quantitative multi-element analysis is often used for rapid screening of the unknown sample composition. A single multi-element standard containing most of the elements of interest is generally used for this purpose. The quantitative analysis is generally achieved by internal, external, or standard addition calibration. An internal standardization is used for correction of instrument drift by normalization of all analyte data to non-analyte isotopes present in known concentrations in both the samples and standards. External calibration uses intensities of analyte isotopes in a number of standards that cover the concentration range of interest. The blank-corrected intensity of standards versus known concentrations is used to produce a calibration curve applied for calculation of unknown concentrations. The standard addition calibration includes spiking of sample solutions with known concentrations of analyte elements. An advantage of this method is that it also provides an effective way to minimize any matrix effects of the sample.

Interferences in ICP-MS are generally low compared to the other atomic spectroscopy methods. Spectral interferences occur when equal mass isotopes of different elements are detected (e.g., ^{64}Ni on ^{64}Zn) or when ions with twice the mass of the analyte are produced. They can be avoided by using alternative interference-free analyte isotopes or corrected by measuring the intensity of isotopes of the interfering element and calculating a correction factor. A common matrix effect in ICP-MS takes place when abundant and easily ionized elements at high concentrations are present in the sample. These effects may be generally overcome by sample dilution or separation of the interfering matrix component.

ICP-MS instruments may also be used to determine the natural isotopic ratios of specific elements such as lead and strontium. Although no standards are required for ratio determinations, the application of the isotope dilution method can significantly improve the accuracy of analysis.

31.3.5 Summary and Conclusions

Atomic and mass spectroscopy techniques are widely used for analysis of trace metals in various materials. Selecting the most appropriate method for a particular analysis depends on many factors, such as sample type, elements of interest, required accuracy, sample volume, speed, cost, etc. This task is further complicated by the fact that the capabilities of the current analytical techniques such as FAAS, GFAAS, ICP-AES, and ICP-MS overlap to a great extent. Table 31.7 summarizes the key advantages and disadvantages of these techniques, which might aid the selection process.

FAAS is the most commonly used analytical technique, mainly due to the lower capital and running cost of the instrument and the simplicity of operation. It is the technique

of choice for single element analyses of small batches of samples. The GFAAS technique has a greater sensitivity and lower detection limits than FAAS and is preferred for analysis of few elements at very low levels using small sample sizes. The main advantages of ICP-AES and ICP-MS over AAS techniques are their multi-element capabilities, larger linear dynamic ranges, and reduced interferences. ICP-AES detection limits are comparable to those obtained by FAAS, and the technique is preferred when a high sample throughput with moderate sensitivity is required. ICP-MS is best for simultaneous analysis of a large number of elements with very low concentrations. However, sometimes the analytical requirements will not be fully satisfied using a single technique, and a combination of two techniques is used to meet all requirements.

31.4 Organic Geochemistry Techniques

31.4.1 Chromatography

31.4.1.1 Principles

Chromatography is essentially a mechanism for the separation of compounds within a mixture. Originally, chromatography was used for the isolation of plant pigments (hence chroma or color). Different components of a mixture will migrate over a surface at different rates, depending on the chemical reactivity, or other interactions with the substrate. Depending on the interactions, the components will move at different speeds, with greater interactions resulting in slower speeds. Essentially, chromatography involves the use of two phases, one stationary and the other mobile (Fig. 31.22). The mobile phase can be either liquid or gas, whereas the stationary phase is typically solid. The stationary phase can have a wide range of polarities that range from non-polar to strongly ionic in nature, and may chemically bind to compounds in the mixture. Those compounds are then released by the introduction of more of the mobile phase, or may utilize chemical changes in the mobile phase. If the mobile phase is liquid, the technique is liquid chromatography (LC); if the mobile phase is gas, it is gas chromatography (GC). As the compounds are eluted from the stationary phase, they are detected by a wide array of systems that include, among others, spectrophotometric or refractive index detection, electron capture detectors (ECD), flame ionization detectors (FID), thermal conductivity detectors (TCD), or mass spectrometers.

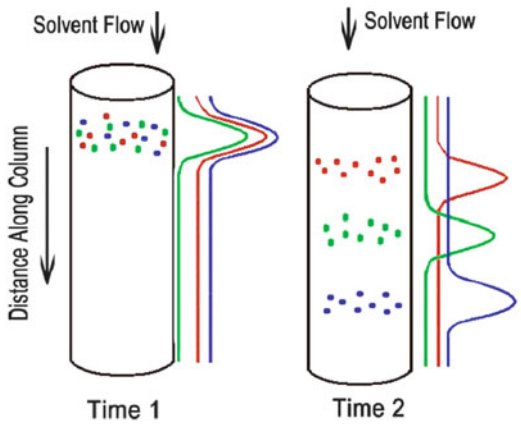


Fig. 31.22 A schematic of the basic principles of chromatography, in which mixtures of compounds can be separated into individual molecular components (the compounds that travel more quickly through the column interact less strongly with the phase in the column)

Environmental samples are extracted from the native matrix using water or organic solvents, and they are concentrated through a process like evaporation of the solvent. The concentrated material is introduced onto the column, which is essentially a tube filled or coated with the stationary phase. The tube can be a few centimeters in length to 50 m or more, and from millimeters in diameter to tens of micrometers. Very small volumes are typically placed on the column, perhaps fractions of microliters in the case of GC or LC. In LC, the mobile phase may be water, or an organic solvent like hexane or benzene. In GC the mobile phase is typically helium. The chromatographic separation scheme is optimized for the kinds of components analyzed, with the stationary phase altered to enhance the resolution of the individual components. Typical separations of a mixture (in these cases polychlorinated biphenyls; PCBs, and fatty acids captured from an aerosol) are shown in Fig. 31.23a, b, respectively.

31.4.2 Mass Spectrometry

31.4.2.1 Principles

Mass spectrometers are used both for the identification and characterization of organic compounds as well as the determination of the stable isotopic abundances of both bulk and molecular components. They are essentially instruments that have the capability to separate charged ions of either atoms or molecules based on their masses, through the use of electrical or magnetic fields (Waples 1981). In the 1920s, F. W. Aston reported the first use of mass spectrometers in the precise determination of the masses of neon isotopes. Many of the developments in the early days of mass spectrometry were aimed at cataloguing and measuring the isotopic compositions of the elements. Following World War II, the use of mass spectrometers was extended to determine the molecular weights of compounds, and the instruments were coupled to chromatographic technologies in order to isolate single components.

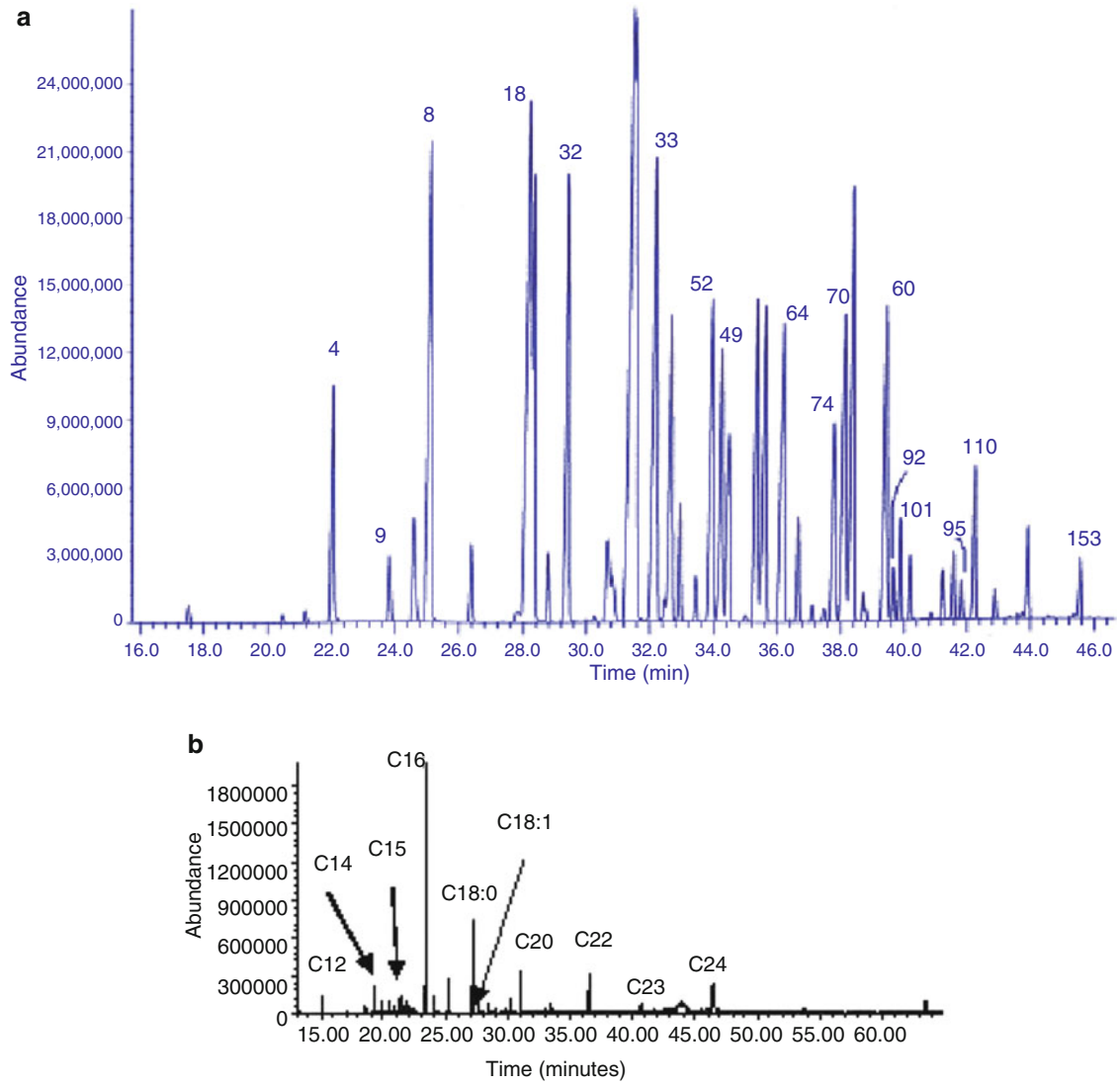


Fig. 31.23 (**a**) GC-MS chromatogram of PCBs. Peaks of interest are labeled with their Ballschmiter numbers (From Yanik et al. 2003a). (**b**) GC-MS chromatogram of fatty acid methyl esters from an aerosol. Peaks are identified by their carbon chain length (From Billmark et al. 2003)

Mass spectrometers used in geochemical research generally have four basic components (Fig. 31.24): an inlet system, an ion source, a magnet, and ion collectors. All of these parts are housed in a system that allows for the maintenance of a high vacuum (typically 10^{-6}–10^{-9} mmHg pressure). High vacuums can be achieved by turbomolecular pumping, or diffusion pumps, which are backed by standard rotary pumps. The high vacuum is important for a number of reasons including the diminishment of interferences, enhancement of ion separation, etc. The inlet system can allow for either static or dynamic introduction of samples. Samples can be converted offline, for example, through acidification, or high temperature combustion, for the introduction of the products as purified gases. The mass spectrometer can also be coupled to a chromatographic system like a GC, which allows for the continuous introduction of species as they are separated.

The source is where the gas is ionized during static introduction of the sample, and it is also the location where the molecules are struck by electrons, which can cause them to fragment into ions. The electron source, which must produce electrons with sufficient energy to accomplish the ionization, is a heated filament.

Ions are deflected by the magnetic field and the radius of that deflection can be precisely calculated using the mass of the ions, the accelerating voltage in the source, and the charge on the ion.

All of these facets are carefully controlled for in the mass spectrometer, and the result is very precise measurements of the mass, as well as the ability for an abundance of the ions to be made. The magnetic field, produced by a permanent or an electromagnet, can either be carefully maintained or rapidly varied in order to accomplish the desired separation. By varying the magnetic field used in coupled GC/MS, the

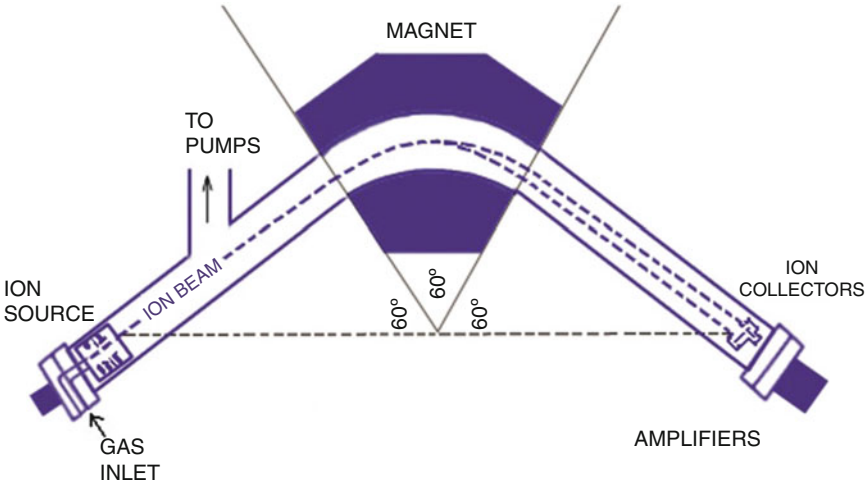


Fig. 31.24 Basic components of a mass spectrometer

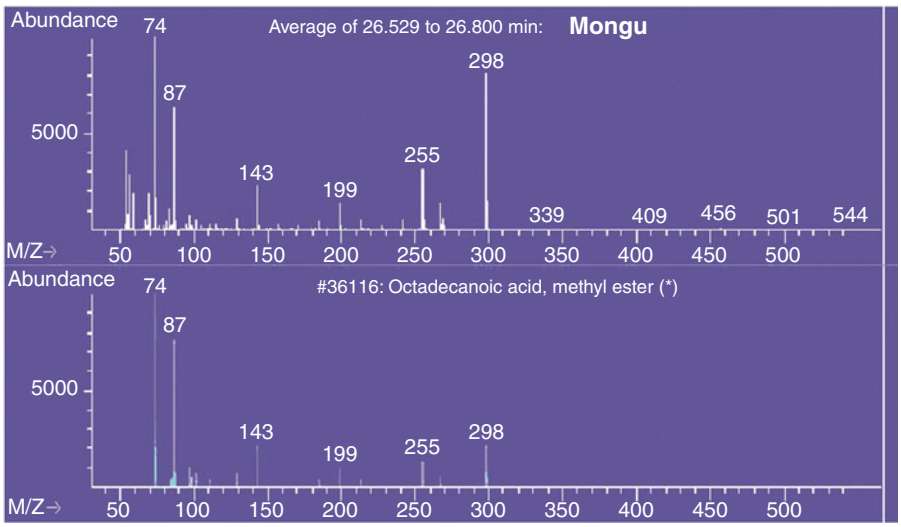


Fig. 31.25 Mass spectral comparison of the fragments of a natural fatty acid methyl ester from the library of compounds that shows the match between the two (From Billmark et al. 2003)

ions of different masses can be directed into a single collector and counted. In such a system, an entire mass spectrum can be observed (Fig. 31.25) for the molecules and their fragments, which can be used to identify the parent molecule. With a static magnetic field, used in stable isotope mass spectrometry, the ions can be directed into multiple collectors, which have been positioned inside the flight tube of the mass spectrometer so that the ion masses are uniquely identified, and are able to be quantified.

The ions are focused and are counted in a collector, which is essentially a metal cup. The detector cup is connected to amplifiers that feed into a device for reporting the signal (typically a computer). In stable isotope mass spectrometry, the signals are observed relative to a standard signal for each cup, and are reported as ratios; hence the name isotope ratio mass spectrometry (IR-MS). More recent modifications of IR-MS technology include the coupling of a GC to the front end of the MS through a combustion system that converts the compounds coming from the GC separation into pulses of gas, which is isotopically assessed (Fig. 31.26).

31.4.2.2 Stable IR-MS

Stable carbon and nitrogen isotope analysis of bulk organic materials is a well-established method for tracing biosynthesis as well as the sources and history of organic matter in the geosphere (Engel and Macko 1993). For example, carbon isotopes have been routinely used to distinguish the biosynthetic pathway of the carbon to be either C3 or C4 synthesis. Stable nitrogen isotopes, sometimes in combination with carbon isotopes, have been used to assess trends in early diagenesis and to elucidate conditions on the early Earth, and to assess the origins of organic nitrogen in extraterrestrial materials as well as to establish trophic orders in modern and fossil food chains (Macko and Engel 1991; Macko et al. 1999; Engel and Macko 2001; MacAvoy et al. 2002).

31.4.2.3 Bulk Isotope Techniques

Samples for bulk organic isotope analysis are usually refluxed in distilled dichloromethane prior to isotope analysis in order to remove lipids. The lipid-extracted samples are converted to carbon dioxide and nitrogen gas for isotope

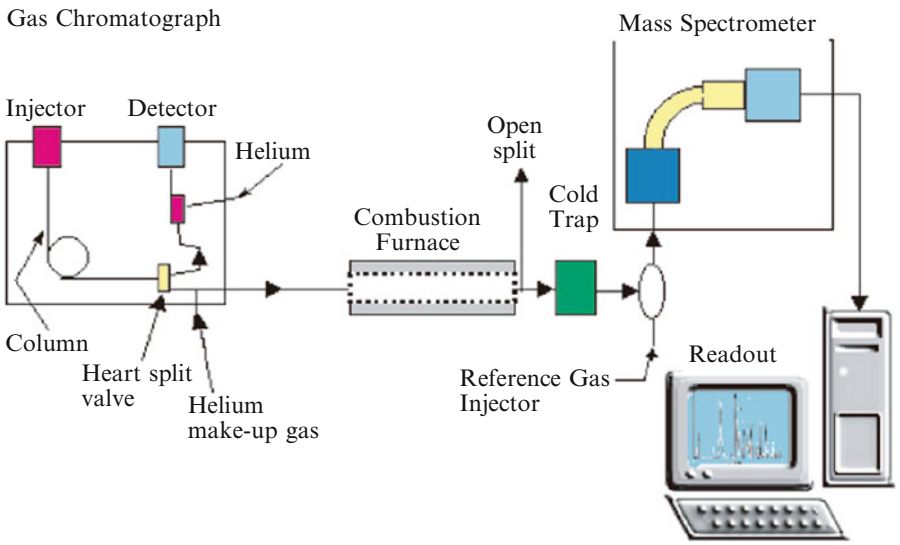


Fig. 31.26 Schematic of the combustion interface between the gas chromatograph and a mass spectrometer designed for stable isotope characterization

analysis using an elemental analyzer, which was coupled to a stable isotope ratio mass spectrometer. Generally, this conversion is a high-temperature combustion involving strong oxidation at high temperature (1,020°C) followed by reduction at lower temperature (650°C). The effluent gases are introduced into the mass spectrometer usin‰g a continuous flow interface. The stable isotopic ratio is reported as follows:

$$\delta^{N}E = \left[R_{sample}/R_{standard} - 1\right]\ 10^{3}(‰) \qquad (31.9)$$

where N is the heavy isotope of the element E and R is the abundance ratio of the heavy to light isotopes (for example, $^{13}C/^{12}C$ or $^{15}N/^{14}N$) of that element. The standard for carbon is the Pee Dee Belemnite limestone (PDB) and for nitrogen the standard is atmospheric nitrogen gas. These are assigned $\delta^{N}E$ values of 0.0‰. The reproducibility of the measurement is typically better than ±0.2‰ for the elements using the continuous flow interface on the mass spectrometer. In the laboratory, the samples are commonly measured against tanks of carbon dioxide and nitrogen gases that have been calibrated against the standards NBS 22 and atmospheric nitrogen, respectively.

31.4.2.4 Compound Specific Isotope Analysis

Stable isotopic determinations made on bulk materials are the weighted averages of the isotopic compositions of mixtures of hundreds to thousands of chemical compounds, each of which has its own isotopic abundances. The relative contribution of each of these materials to the isotopic content of the bulk material could theoretically be quantified through mass-balance or isotopic-mixing equations. The stable isotope analysis of individual molecular components holds great potential as a method of tracing the source, biochemistry, diagenesis, or indigeneity of a material.

The compounds studied to date include hydrocarbons, tetrapyrroles (chlorophyll derivatives), fatty acids, carbohydrates, and amino acids. In recent studies, nitrogen and carbon isotope analyses of components of petroleum and hydrocarbon extracts of sediments have indicated the preservation of original source materials. Isotope analyses of individual amino acids using both carbon and nitrogen isotopes have been useful in detailing indigeneity of organic matter in meteorites and fossils as well as helping to understand diagenesis. Inscribed in the isotopic signature is an indication of the biosynthetic pathway used in the formation of the compound. The transfer of nitrogen and carbon within the organism forming the component can thus be better understood. These pathways in turn imprint the signature of the organism in the rocks and sediments from which the compounds can later be isolated.

Over the years, numerous attempts have been made to isolate individual molecular components using liquid and gas techniques in order to better interpret or trace the source of history of an organic material. The possibility of comparative biochemistry in modern or fossil organisms has been suggested through the assessment of the isotopic differences between compounds of a family of components. Such differences are the result of enzymatic fractionation effects during synthesis or metabolism of the compound. Isotopic compositions of individual hydrocarbons have the potential for establishing sources for the materials, bacterial or otherwise, and have been useful in correlation techniques both in the petroleum industry and in pollution assessment. Individual carbohydrate isotope compositions also show great potential in metabolic and diagenetic studies. Depletions in the isotopic compositions of the products of reactions permit calculations to be done that quantify use and production of new organic materials and resolve them from native materials, even though the chemical compositions of the substances are identical. The studies to date have yielded

important information regarding the source and history of the compounds characterized.

Through recent technological advancements, GC effluents can be combusted and the resulting carbon dioxide directly introduced into an IR-MS. This modification, GC-IRMS, allows for rapid analysis of the carbon and nitrogen isotopes on components in a mixture, and with increased sensitivity, on the order of 0.5 nmol L^{-1} of each compound. GC-based systems are presently constrained by the volatility of the components investigated. Compounds that do not have this constraint include hydrocarbons, the analyses of which have already clearly demonstrated the power of such technology and measurements in the assessments of source and the history of organic materials.

Nonvolatile, multifunctional molecules, including carbohydrates, fatty acids, and amino acids, require derivatization prior to GC analysis to increase volatility. Through the derivatization, additional carbon (but not nitrogen) is thus added to the parent compound. This addition, as well as fractionations associated with derivatization procedures involving bond rupture and formation in, for example, esterification and acylation, needs to be corrected in order to ascertain the original isotopic composition of the compound. The original carbon and nitrogen isotopic compositions of individual amino acids and their stereoisomers have been able to be computed, however, through analysis of standards prepared in a similar fashion. Through GC-IRMS analysis of amino acids, enrichments of ^{13}C and ^{15}N for both stereoisomers of the same amino acid from a meteorite have confirmed, for example, the extraterrestrial origin of those components and supported the lack of contamination by terrestrial compounds in the absolute concentrations and stereoisomer relationships.

Lipids that originate from tree waxes are more depleted (have less of the heavy isotope) than those derived from marine plants. Thus, in sediments, resolution of chemically indistinguishable sources is now possible. Additionally, pollutants, including PCBs and phenylalanine hydroxylase (PAHs), are ideally suited for the source and history analysis through the GC-IRMS analysis of extracts of natural materials (Zepp and Macko 1997; Yanik et al. 2003a, b). Hydrocarbon components from fossil fuels have isotopic compositions that are readily resolvable from those of "natural" lipids, and thus have potential in the analysis and tracking of pollutants. Atmospheric contaminants from biomass burning have been able to be tracked to source materials because of the selective isotope signals of both the PAHs produced and the volatilization of fatty acids.

31.4.2.5 GC and GC-IRMS Procedures

The chromatographic separation techniques have a wide field of applicability in biochemistry, such as some of the examples given below (see also Knapp 1979).

Amino Acids

Solutions of amino acids are extracted from mineral matrices using 6 N HCl and dried under a stream of N_2 at 40°C. The dried samples are esterified with acidified 2-propanol for 1 h at 110°C. The solvent is removed by evaporation under a gentle stream of nitrogen gas at 25°C. The amino acid isopropyl esters are acylated with trifluoroacetic anhydride (TFAA) at 110°C. Next, the excess TFAA and CH_2Cl_2 are removed by evaporation under N_2. The derivatives are redissolved in CH_2Cl_2. The TFAA isopropyl esters of the individual enantiomers of amino acids are analyzed directly for their stable carbon or nitrogen isotope compositions by using the GC-IRMS system (Macko et al. 1997).

Carbohydrates

Alditol acetates of the individual sugars are prepared with the sample reacted with sodium borohydride to reduce the aldehyde group. Following neutralization with acetic acid, which destructs the carbohydrate-borate complex, and vacuum rotary evaporation, the carbohydrate mixture is reacted with acetic anhydride. The acetylated products are then rotary evaporated and subsequently washed with methanol and chloroform and lastly, filtered prior to analysis by GC (Macko et al. 1998).

Fatty Acids and Hydrocarbons

The fatty acids are derivatized to fatty acid methyl esters (FAME) with BF_3 in methanol and extracted with hexane. The excess solvent is evaporated. Hydrocarbons can be analyzed directly in the GC-IRMS system without derivatization.

GC-MS Techniques

For the separation of the individual components by GC, the columns and conditions are optimized for the types of components. As an example, amino acid stereoisomers are separated using a chiral phase column and a temperature program. The GC conditions are as follows: ~1–2 nmol L^{-1} of each derivative is injected and subsequently introduced directly into the source of the MS; the carrier gas is ultrapure He (99.9999%) at a head pressure of 80 kPa; the injector temperature is 200°C; the GC temperature program is 45°C for 3 min, 45–90°C at 45 min, 90°C isothermal for 15 min, 90–190°C at 3°C per minute, and then 190°C isothermal for 30 min. The solvent (ethyl acetate) peaks are removed from the effluent of the GC through a heart split valve, which is open at the time of injection. The valve is programmed to close to allow the column effluent to be directed to the mass spectrometer.

GC-IRMS Techniques

For isotope analysis, an IR-MS is interfaced to a GC through a combustion furnace (copper/nichrome wire, at 850°C) and

a cryogenic trap (−90°C). The GC is equipped with a fused silica capillary chromatographic column with the appropriate phase, temperature, and flow rates again optimized for the separation. The performance of the GC-IRMS system, including the combustion furnace, is evaluated by injection of a laboratory standard of known isotopic composition. For all runs, background subtraction is performed using the parameters supplied by the GC-IRMS software. Two to five replicate GC-IRMS runs are performed for each sample. The reproducibility ranges between 0.1 and 0.5% (1 standard deviation). The accuracy of the samples' analyses is also monitored by co-injection of a laboratory standard of known isotopic composition. The isotopic shift due to the carbon introduced, for example, in the fatty acids methylation is corrected by the following relationship (MacAvoy et al. 2002):

$$\delta^{13}C(\text{FAME}) = f(\text{FA})\delta^{13}C(\text{FA}) + f(\text{MeOH})\delta^{13}C(\text{MeOH})Z \quad (31.10)$$

where $\delta^{13}C$(FAME), $\delta^{13}C$(FA), and $\delta^{13}C$(MeOH) are the carbon isotope compositions of the fatty acid methyl ester, the fatty acid, and the methanol used for methylation of the fatty acid, respectively; f(FA) and f(MeOH) are the carbon fractions in the fatty acid methyl ester in the underivatizad fatty acid and methanol, respectively. Amino acids and other less volatile components require derivatization for GC analysis, and will have similar correction strategies to obtain the naturally occurring isotopic composition.

Analysis of the stable nitrogen or carbon isotope composition of each component is accomplished by comparison to reference gas pulses introduced at the start of the run and following the opening of the heart split valve at the end of each run, i.e., after 4,500s.

Conclusion

The potential for application of organic geochemical analysis to any multitude of environmental, ecological, or biochemical research areas is only beginning to be realized. Extension of compound-specific isotope analytical data derived from modern organisms and settings to yield interpretations of ancient depositional environments certainly appears possible. Further application of the technologies to understand the cycling of carbon and nitrogen, the identification and alteration of pollutants, or resolve metabolic relationships between compounds in living or extinct organisms are all within the scope of future research.

See Also the Following Chapters

Chapter 25 (Environmental Pathology) • Chapter 27 (Speciation of Trace Elements) • Chapter 30 (Mineralogy of Bones) • Chapter 32 (Histochemical and Microprobe Analysis in Medical Geology)

Further Reading

Bertin EP (1975) Principles and practice of X-ray spectrometric analysis. Plenum Press, New York

Billmark KA, Swap RJ, Macko SA (2003) Characterization of sources for southern African aerosols through fatty acid and trajectory analyses. J Geophys Res 108D13, 8503 SAF 39:1–9

Buhrke VE, Jenkins R, Smith DK (1997) A practical guide for the preparation of specimens for X-ray fluorescence and X-ray diffraction analysis. Wiley, New York

De Soete D, Gijbels R, Hoste J (1972) Neutron activation analysis. Wiley, New York

Engel MH, Macko SA (1993) Organic geochemistry, principles and applications. Plenum Press, New York

Engel MH, Macko SA (2001) The stereochemistry of amino acids in the Murchison meteorite. Precambrian Res 106:35–46

Jenkins R, Gould RW, Gedcke D (1981) Quantitative X-ray spectrometry. Marcel Dekker, New York

Johnson A, Lalor GC, Robotham H, Vutchkov MK (1996) Analysis of Jamaican soils and sediments by energy-dispersive X-ray fluorescence spectrometry. J Radioanal Nucl Chem 209(1):101–111

Kingston HM, Jassie LB (eds) (1988) Introduction to microwave sample and preparation: theory and practice. American Chemical Society, Washington, DC

Knapp DR (1979) Handbook of analytical derivatization reactions. Wiley, New York

Lalor GC (1995) A geochemical atlas of Jamaica. Canoe Press/University of the West Indies, Kingston

Lalor GC, Rattray R, Simpson P, Vutchkov MK (1999) Geochemistry of an arsenic anomaly in St. Elizabeth, Jamaica. Environ Geochem Health 21(1):3–11

Lalor GC, Vutchkov MK, Grant C, Preston J, Figueiredo AMG, Favaro DIT (2000) INAA of trace elements in biological materials using the SLOWPOKE-2 reactor in Jamaica. J Radioanal Nucl Chem 244 (2):263–266

Lalor GC, Rattray R, Vutchkov M, Campbell B, Lewis-Bell B (2001) Blood lead levels in Jamaican school children. Sci Total Environ 269:171–181

MacAvoy SE, Macko SA, Joye SB (2002) Fatty acid carbon isotope signatures in chemosynthetic mussels and tube worms from the Gulf of Mexico hydrocarbon seep communities. Chem Geol 185:1–8

Macko SA, Engel MH (1991) Assessment of indigeneity in fossil organic matter: amino acids and stable isotopes. Philos Trans R Soc Lond B 333:367–374

Macko SA, Uhle ME, Engel MH, Andrusevich V (1997) Stable nitrogen isotope analysis of amino acid enantiomers by gas chromatography/combustion/isotope ratio mass spectrometry. Anal Chem 69:926–929

Macko SA, Ryan M, Engel MH (1998) Stable isotopic analysis of individual carbohydrates by GC/C/IRMS. Chem Geol 152:205–210

Macko SA, Engel MH, Andrusevich V, Lubec G, O'Connell TC, Hedges REM (1999) Documenting the diet in ancient human populations through stable isotope analysis of hair. Philos Trans R Soc 353:1–12

Schlemmer G, Radziuk B (1999) Analytical graphite furnace atomic absorption spectrometry: a laboratory guide. Springer, Berlin

Tertian R, Claisse F (1982) Principles of quantitative X-ray fluorescence analysis. Heyden & Son, Philadelphia

Vutchkov M, Grant C, Lalor GC, Preston J (2000) Standardization of the SLOWPOKE-2 reactor in Jamaica for routine NAA. J Radioanal Nucl Chem 244(2):355–359

Walsh A (1955) The application of atomic absorption spectra to chemical analysis. Spectrochim Acta 7:108–117

Waples D (1981) Organic geochemistry for exploration geologists. Burgess Publishers, New York

Welz B, Sperling M (1999) Atomic absorption spectrometry. Wiley-VCH, Weinheim

Yanik PJ, O'Donnell TH, Macko SA, Qian Y, Kennicutt MC (2003a) Source apportionment of polychlorinated biphenyls using compound specific isotope analysis. Org Geochem 34:239–251

Yanik PJ, O'Donnell TH, Macko SA, Qian Y, Kennicutt MC (2003b) The isotopic composition of PAHs during biodegradation. Org Geochem 34:165–183

Zepp RG, Macko SA (1997) Polycyclic aromatic compounds in sedimentary records of biomass burning. In: Clark J, Cachier H, Goldammer JG (eds) Sediment record of biomass burning and global change. Springer, Berlin, pp 145–168

Histochemical and Microprobe Analysis in Medical Geology

32

José A. Centeno, Todor Todorov, Joseph P. Pestaner, Florabel G. Mullick, and Wayne B. Jonas

Contents

32.1 Introduction

Understanding the detrimental effects that geochemical processes and environmental pollutants may have on the health of humans and animals has been the subject of extensive study in medical geology and environmental pathology. For example, information obtained from the chemical analysis of mineral deposits in tissues may provide insight into a particular disease state and assist in the development of new treatments and therapy. In many cases, diseases related to the environment can be directly linked to the presence and distribution of toxic chemical elements in the soil, air, or water, such as chronic exposure to arsenic through contaminated drinking water (Centeno et al. 2002). To demonstrate such links, it is necessary to analyze minerals and geo-environmental toxins to obtain information on the possible origins of such diseases. Because of the ever-growing complexity of geological sources and toxic environmental, biological, and chemical agents, accurate, rapid, and nondestructive techniques for qualitative and quantitative analysis of these materials are essential.

The identification and quantification of biogeochemical mineral deposits in human tissues are often valuable adjuncts to defining the diagnosis or understanding the pathogenesis and etiology of a disease. In the chemical pathology laboratory, analyses of tissues may proceed in a variety of ways but generally can be classified into three types: (1) histochemical/light microscopy, (2) ultrastructural methods (*e.g.*, scanning electron microscopy, energy dispersive x-ray microanalysis), and (3) spectroscopic techniques.

Utilizing light microscopy with histochemical stains on paraffin block tissue sections is often the simplest and least expensive approach for rendering a histopathological diagnosis. This process involves the use of selective stains that are sensitive to the presence of particular components within tissues; a well-known example is the Prussian blue stain used to highlight the presence and location of iron deposits in tissues. Because histochemical methods often utilize morphological abnormalities rather than biochemical changes,

The opinions and assertions expressed herein are those of the authors and are not be constructed as official or as representing the views of the Joint Pathology Center, the Department of the Army, or the US Department of Defense.

J.A. Centeno (✉)
The Joint Pathology Center, Division of Biophysical Toxicology, Malcolm Grow Medical Clinic, Joint Base Andrews Air Naval Facility, Washington, DC, USA
e-mail: Jose.a.Centeno@us.army.mil

T. Todorov
Crustal Geophysics and Geochemistry Science Center, United States Geological Survey, P.O. BOX 25046, Denver, CO, USA

J.P. Pestaner
Medical Examiner Office, Washington, DC, USA

F.G. Mullick
The Joint Pathology Center, Silver Spring, DC

W.B. Jonas
East Coast U.S. Office, Samueli Institute for Information Biology, Alexandria, VA, USA

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_32,

however, it is frequently difficult to use such methods to completely assess the identities of complex foreign materials and endogenous cellular components in tissues.

On the other hand, ultrastructural and microprobe methods, such as scanning electron microscopy (SEM) combined with energy-dispersive x-ray analysis (EDXA), are useful for identification of mineral deposits (Goldstein et al. 2002). Despite the wide use of these electron-based optical systems, the SEM and EDXA techniques do have some disadvantages. Both techniques require extensive sample preparation (*e.g.*, paraffin sections must be mounted on pure carbon disks and coated with a layer of gold or carbon to avoid electron beam charging effects), and, although these techniques provide useful information on the elemental composition and morphological features of the sample, they may lack the required structural, molecular, and chemical sensitivity for *in situ* identification and localization of foreign substances in tissues.

Other physical and chemical techniques that have been traditionally used for the study of foreign materials and mineral deposits in tissues include x-ray diffraction, x-ray absorption, ultraviolet absorption, fluorescence, interference microscopy, polarization microscopy, and autoradiography. The powder method of x-ray diffraction employing the Gandolfi (1967) camera has been particularly useful for the analysis of crystals that have been separated from tissues or obtained from tissue biopsy specimens. In contrast to the combined SEM–EDXA technique, which provides information on the elemental composition of mineral specimens, x-ray diffraction has been widely used for the identification of individual chemicals or mixture of chemicals (Willard et al. 1988).

Recent technological advances in optical instrumentation have led to the development of new microprobe techniques that are capable of providing *in situ* analysis, chemical distribution (mapping), and quantitative information of tissues. Chemical and molecular spectroscopic techniques, such as Fourier transform infrared (FTIR) microspectroscopy and Raman microprobe spectroscopy, provide accurate, rapid, and selective identification by virtue of a molecule's characteristic spectrum of vibrational frequencies (Gillie et al. 2000; Mulvaney and Keating 2000). Raman microprobe spectroscopy is a light-scattering technique in which a laser of known wavelength is scattered off a molecule of interest (*e.g.*, mineral particle) to produce a spectrum of bands representing movements of atoms in the molecule (McCreery 2000). A Raman microprobe instrument provides 1-μm spatial resolution on histological sections and does not require staining for identification of chemical species (Centeno et al. 1999; Hanlon et al. 2000). Analysis is nondestructive and is easily correlated with morphological features to aid in the evaluation of a disease process. The Raman microprobe can be used to identify the microscopic contents of various organelles and diseased versus healthy tissue through determination of differences in vibrational spectra of intrinsic molecules; as a result, it is now possible to provide full chemical mapping of specimens of interest without the use of histochemical stains.

Infrared microspectroscopy, another vibrational technique, in principle has characteristics similar to those of Raman scattering (Diem 1999). In comparison with Raman microprobe spectroscopy, however, infrared microscopy suffers from two disadvantages: Spectral resolution is one order of magnitude poorer, and aqueous-based systems are intense absorbers. In contrast, water is a weak scatterer in Raman spectroscopy, which makes it an excellent medium for studying biological systems. Both infrared microscopy and Raman microprobes allow the analysis of individual particles as seen in tissue sections or of digested bulk samples in cases where digestion of tissues is required (Centeno et al. 1994).

This chapter describes the use of histochemical, microspectroscopic, ultrastructural, and microprobe tools available to biomedical scientists that are essentially identical to those used by geoscientists. The preferred technique in a particular situation depends on the analyte; thus having some idea of what is being examined is important. The primary focus here is on the Raman microprobe because of its nondestructive operation and simple sample preparation. The Raman microprobe can be used to assist in the identification of such materials as calcium oxalate inclusions in tissues as well as for myriad other applications that are discussed in detail later in this chapter. For detailed information on chemical analyses in the geosciences, see Chap. 29.

32.2 Histochemical Techniques

Histochemical examination of tissue sections is the simplest and least expensive technique for analyzing minerals in tissue. The process involves the use of selective stains that are sensitive to the presence of particular components, such as minerals within tissues. This short review discusses the most prominent histochemical techniques; for greater detail on the various staining techniques, see *Armed Forces Institute of Pathology Special Staining Techniques* (1995). The three methods used today for analyzing minerals in tissue section are microincineration, digestion of tissues, and the use of special stains.

32.2.1 Microincineration

Microincineration is a procedure employed to study mineral content without the interference of organic elements; in particular, iron, calcium, magnesium, and silicon can be identified in incinerated material. Special stains allow dyes to bind to tissue elements to enhance microscopic details of specific elements in the tissue and to assist in ruling out other possibilities, such as infection. Incineration is accomplished

by increasing the temperature of the tissue on a slide to as high as 650°C; the sample is then cooled. A Bunsen burner or appropriate furnace may be used for the incineration. Incinerated organic material (*i.e.*, ashes) appears gray to colorless; illumination and transmitted light microscopy are reduced. Calcium and magnesium oxides in tissues appear as a white ash, silicon is a white birefringent crystalline material, and iron appears as a red to yellow oxide.

32.2.2 Digestion of Tissues

Digestion of tissues in 1-*N* sodium hydroxide is another histochemical technique used to separate foreign materials from fresh or fixed tissues. The procedure involves heating the tissue in NaOH for approximately 1 h, after which the digested material is rinsed in absolute ethanol and washed with distilled deionized water. The resulting sediment is examined using light microscopy, which includes the use of polarized light and darkfield microscopy. In addition, residues can also be placed on a carbon disk and examined by SEM–EDXA.

32.2.3 Special Stains

The evaluation of toxic tissue reactions to foreign or apparently foreign materials in tissues is a frequent challenge to pathologists. Typically, such reactions range from minimal hemorrhage to extensive inflammatory reaction, and the question becomes one of determining the etiology of the reaction. Generally, infectious agents and exogenous minerals and elements generate similar inflammatory reactions; thus the pathologist's first approach to such histological analysis is to stain for infectious agents so that prompt and proper medical therapy may be administered. When inflammatory reactions are due to infectious agents, stains that are helpful in this diagnostic endeavor include hematoxylin and eosin (H&E), periodic acid–Schiff reaction (PAS stain), Gomori's methenamine silver, and Brown–Brenn Gram stain, as well as acid-fast staining techniques. Such staining techniques are beyond the scope of this chapter, but in-depth treatments of these staining techniques can be found in staining manuals (see, for example, Prophet et al. 1995). The exclusion of infectious agents is clinically important before more extensive evaluation for minerals is performed.

32.2.4 Selected Histological Techniques for Iron, Calcium, Copper, and Zinc

A deficiency or excess of iron in developing countries has been associated with the occurrence of a wide range of human diseases and health conditions, including anemia, growth impairment, and even liver diseases. In individuals affected with diseases caused by excess iron, iron is found in tissue in both ferric and ferrous forms. Perl's Prussian blue reaction (see Prophet et al. 1995), which is useful for identifying ferric ions, is a very sensitive histochemical technique that forms a stable pigment that is preserved even after other stains are performed. In addition, single granules of iron can be identified that might otherwise remain undetected. Perl's stain for ferric iron involves treatment of tissue sections with acid solutions of potassium ferrocyanide, which reacts with ferric ions in the tissue to form ferric ferrocyanide which possesses a blue color (*i.e.*, the Prussian blue pigment). An example of the use of this stain is shown in Fig. 32.1, taken during the study of a case of ferric poisoning. The ferric ions have penetrated through the stomach wall and are observed in the adjacent blood vessels (blue stains around numerous vessels). Figure 32.1 demonstrates that, in this particular case, the excess iron has been absorbed through the stomach wall surrounding the blood vessels, which probably resulted in internal hemorrhage. In contrast, Turnbull's blue reaction utilizes potassium ferricyanide to demonstrate the presence of ferrous ions in tissue. The ferricyanide combines with any ferrous ions present in the tissue, resulting in Turnbull's blue pigment.

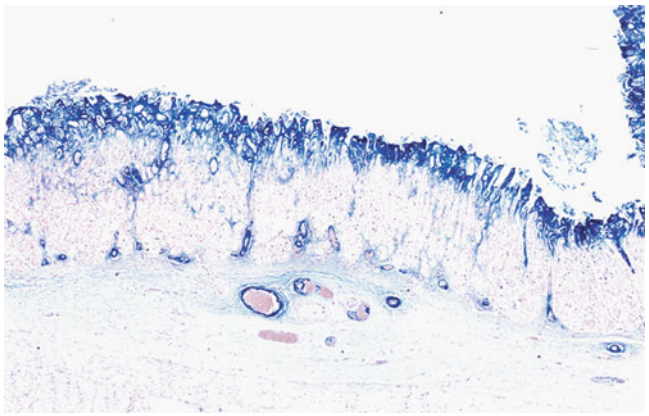


Fig. 32.1 Examination of gastric tissue by the Pearl's Prussian blue histochemical stain for the presence and location of ferric iron (magnification 10×)

A link between calcium and cardiovascular disease has been the subject of many studies (see also Chap. 13, this volume). Because calcium levels in drinking water correlate with the extent of cardiovascular disease, identifying calcium ions in tissues is a helpful diagnostic tool. Alizarin red S and Von Kossa are special stains that may be used to demonstrate calcium in tissues. The Alizarin red reaction is not specific for calcium, as magnesium, manganese, barium, strontium, and iron may show similar reactions; positive staining depends on the chelation process with the dye, and calcium appears as an orange-red precipitate. The Von Kossa silver test is a metal substitution reaction where tissue sections are treated with a silver nitrate solution and silver is deposited in place of the calcium.

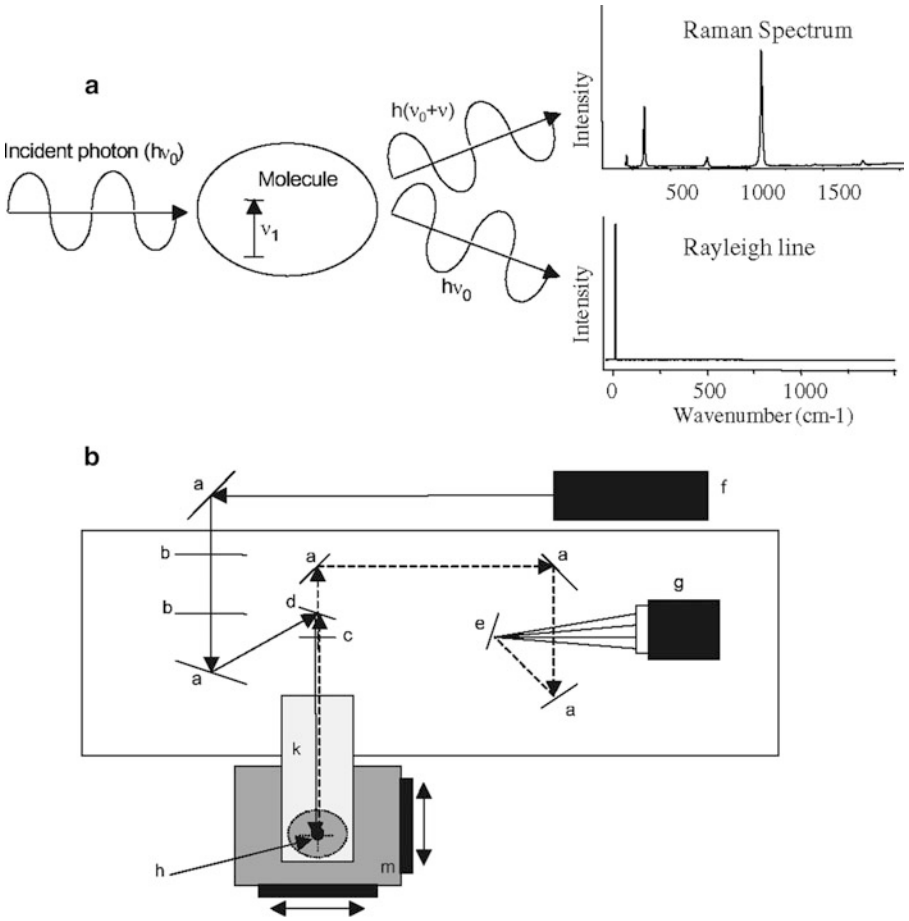


Fig. 32.2 Schematic representation of the Raman effect (**a**) and the Raman microprobe system used in our laboratory (**b**). The excitation of the molecule by incident photons produces the Rayleigh line (*bottom trace*) and Raman scatter (*top trace*). Components of the Raman microscope setup (**b**) include: *a*, mirrors; *b*, beam-focusing lenses; *c*, holographic notch filter; *d*, beamsplitter; *e*, holographic grating; *f*, laser light source; *g*, low noise charge-coupled device (CCD) camera; *h*, specimen to be analyzed; *k*, microscope; and *m*, X–Y motorized stage

Copper is generally demonstrable only in diseases where it is present in relatively high concentrations, such as in Wilson's disease. Copper can be demonstrated by its reaction with rubeanic acid. Rubeanic acid reacts with copper to form a dark precipitate consisting of copper rubeanate. Nickel and cobalt have similar reactions, but these metals are soluble when acetate is in the staining solution; thus, using acetate in the staining solution makes the reaction with rubeanic acid more specific for copper. Nevertheless, variability in staining may be attributed to the protein binding of copper ions in tissue.

Zinc is an important element in cell metabolism, and its deficiency affects rapidly growing cells. Zinc toxicity is an unusual occurrence and usually has mild toxic effects. Zinc may be demonstrated by the dithizone reaction, where dithizone reacts with zinc to form an insoluble red-purple complex, although other metals can interfere with the reaction.

32.3 Microprobe (Vibrational) Spectroscopy

Definitive identification of foreign material for geologic and medical investigations can be problematic. Raman microprobe spectroscopy is a well-established vibrational technique for the study and identification of organic, polymeric, and inorganic materials (McCreery 2000). Its utility in mineralogical analysis has increased significantly, as sample preparation is minimal, the test results are obtained rapidly, and the sample is not destroyed. In chemical pathology, the Raman microprobe has been used for more than 10 years for identification of chemical species of foreign inclusions and metabolic accumulations in histological sections of patients biopsied in order to understand their disease state. The microprobe has been fairly successful in identifying the chemical species of inclusions of unknown origin. Chemical mapping of such samples not only confirms the original identification but also provides additional information for the clinician regarding morphology, location, and distribution of species that can be useful in understanding the causes of diseases.

32.3.1 Raman Microspectroscopy

In Raman spectroscopy, the Raman effect or scattering signal arises due to the interaction of an incident laser beam (photons) ($h\nu_0$) that is focused onto a sample with the molecular vibrational modes (ν_1) of the interrogating chemical species (Fig. 32.2a). If no interaction takes place, the resulting Raman signal, known as Rayleigh scattering, occurs at the same frequency ($h\nu_0$) as the incident laser

beam; hence, no information can be obtained from the chemical system. The result of the interaction, though, may produce Raman signals at frequencies higher (anti-Stokes) and lower (Stokes) than the frequency of the incident beam. At room temperature, the Stokes scattering is more intense and is used to characterize the chemical composition of the sample. The peaks or lines observed in a Raman spectrum are Raman active vibrations attributed to structurally active molecular bonds. The shift from the frequency of the incident wavelength is known as the Raman shift (*x*-axis). Although Raman spectroscopy has been widely used for the study of macrochemical systems, only recently has wide use of the technique coupled to high-resolution (1 μm) optical microscopes been demonstrated. In contrast to SEM and EDXA, which require complex processing and handling of the tissue, Raman microprobe spectroscopy can be used with fresh tissues as well as paraffin-embedded tissue sections without having to destroy, decompose, or extract the tissue to study the foreign material.

The Raman microprobe experiments reported in this chapter were conducted employing the system shown schematically in Fig. 32.2b. Briefly, the system consists of a spectrograph (model LabRam, Jobin Yvon, Edison, NJ) equipped with holographic gratings blazed at 1,800 and 600 grooves mm^{-1} and interfaced to an Olympus microscope (Model BH-40). Specimens can be visualized using white light or laser source at magnification of 100× objective (N.A. = 0.90), 50× objective (N.A. = 0.75), or 10×. For light excitation, several laser lines may be used with wavelengths at 514, 632, and 785 nm. As demonstrated in Fig. 32.2b, during the Raman microprobe measurement, a laser is focused onto the specimen using one of the objectives. The backscattered Raman signal is collected by the same objective, collimated, and directed into the spectrometer to determine the frequencies or wavelengths. The minimum size of the laser focus spot is 1 μm. The Raman scatter signal is then analyzed by a detector system consisting of a charged coupled device (CCD). The following examples illustrate the use of Raman microprobe spectroscopy.

32.3.2 Identification of Mineral Materials in Inorganic Particulate Tissues and Geo-Environmental Samples by Raman Microprobe Spectroscopy

The chemical, elemental, and/or molecular identification of mineral deposits in human tissues is an important step in defining a diagnosis or understanding the pathogenesis of a disease. Raman microprobe provides a useful analytical and diagnostic approach for the qualitative and quantitative identification of mineral deposits in human tissues. The formation of minerals such as calcium oxalate crystals in tissues serves as a simple example for the *in situ* nondestructive and noninvasive application of this technique in human pathology. The occurrence of calcium oxalate crystals in tissues is frequently associated with the formation of chronic granulomatous inflammatory lesions from a wide range of causes. As such, these crystals may be regarded as products of inflammatory cells. Identification of these crystals in tissues has been demonstrated by the use of histochemical methods based on the microincineration of oxalate to produce carbonate (Johnson and Pani 1962). SEM–EDXA has also been used to demonstrate the presence of calcium and oxygen. Each of these techniques requires extensive sample preparation procedures, and none of these methods provides the capability of studying the localization *in situ* of these crystals within tissues. The Raman microprobe, on the other hand, not only is capable of providing information about the localization of crystals but also provides multidimensional information on the chemical nature of the crystal. Figure 32.3 demonstrates the use of the Raman microprobe for the study of calcium oxalate crystals in kidney tissue (Pestaner et al. 1996). Under ordinary light microscopy and crossed polarized image (Fig. 32.3a, b, respectively; H&E stain), crystals of calcium oxalate are observed. In this case, Raman spectroscopy proved to be very useful in the identification of not one but two forms of calcium oxalate present in the mineral inclusion: the monohydrate and dihydrate forms. These species are difficult to distinguish by classical histochemical methods. The monohydrate is characterized by a doublet at 1,463/1,491 cm^{-1}. The two bands merge into one band at 1,476 cm^{-1} for dihydrate calcium oxalate. This spectral difference provides rapid identification of calcium monohydrate and dihydrate mineral inclusions in tissue.

The characterization and identification of inorganic atmospheric particulate that may be involved in the development of respiratory diseases is an increasing concern for environmental pathologists and geoscientists. Formation of the so-called "intercontinental dust" or "African dust" has been described (Prospero 2001); however, its composition and impact on human health have not been fully characterized. Applying the Raman microprobe to improve our understanding of the composition and behavior of this type of contaminant is expected to play an increasing role in medical geology and environmental pathology. Figure 32.4 demonstrates the histological and *in situ* analysis of dust particulates in lung tissues employing the confocal Raman microprobe. The histological appearance of dust particles in this particular case is characterized by the presence of opaque granular black material and is morphologically consistent with carbon and silica (Fig. 32.4a). The opaque black granules are easily discernible under white light image (Fig. 32.4b). Using the Raman microprobe, the foreign material was identified as a carbon-based dust particle with characteristic peaks at 1,326, 1,574, and 2,656 cm^{-1} (Fig. 32.4c).

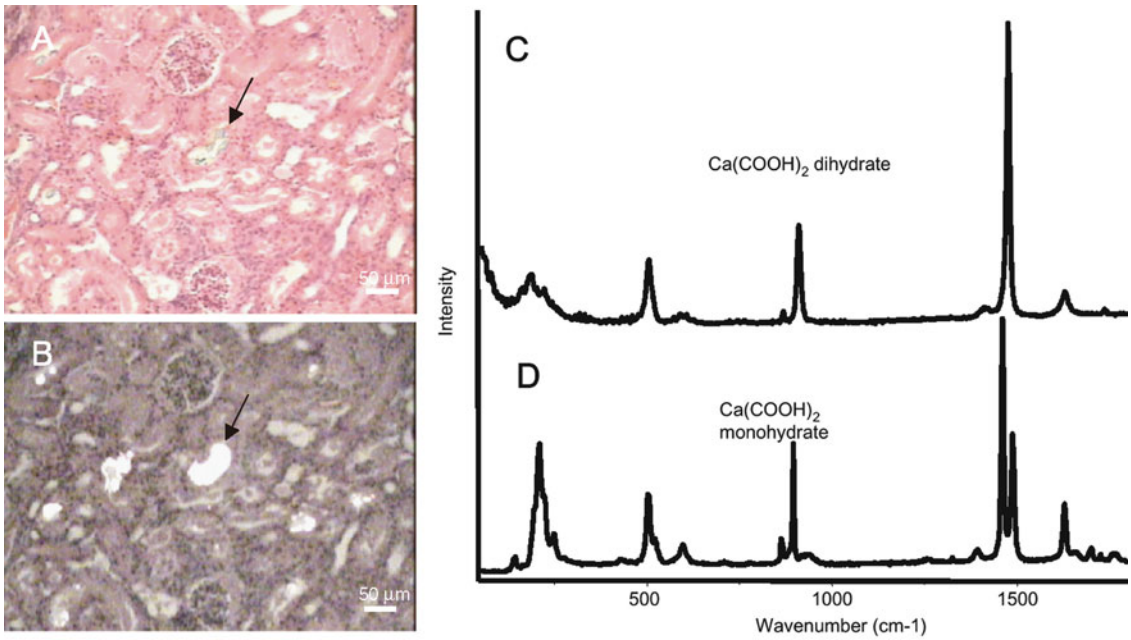


Fig. 32.3 Raman spectroscopic investigation of $Ca(COOH)_2$ inclusions in kidney tissues. *Panel A* shows the light microscopy (H&E) and *panel B* the crossed polarized (H&E) images of calcium oxalate crystals in renal tissues. The *arrow* in the *white light* image (**A**) indicates the particle for which spectra were taken. Both forms of calcium oxalate—the monohydrate ($Ca(COOH)_2 \cdot H_2O$) and dihydrate ($Ca(COOH)_2 \cdot 2H_2O$)—were identified on the same crystal, which reveals the sensitive nature of Raman microprobe. The spectra were obtained using a 632-nm laser excitation and 100× magnification

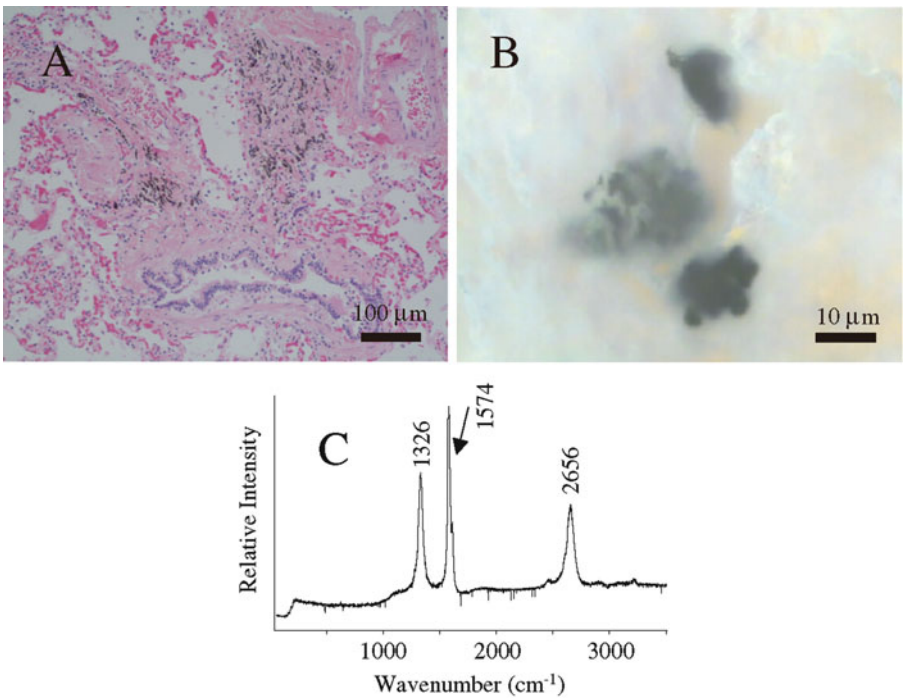


Fig. 32.4 Analyses of inorganic particulate in lung tissues by Raman microprobe spectroscopy. (**a**) Histological section (H&E stain) at 10× magnification; (**b**) light microscopy image of unstained tissue demonstrating the *black* opaque dust inclusions (100× magnification); (**c**) Raman spectra of a carbon-based dust inclusion in the tissue section shown in (**b**)

In addition to point acquisition, as demonstrated in Figs. 32.3 and 32.4, Raman microspectroscopy can be used to generate maps or images based on different Raman spectral features (*i.e.*, vibrations from the various components) (Carden and Morris 2000; Ling et al. 2002; Shafer-Peltier et al. 2002; Uzunbajakava et al. 2003). Figure 32.5 shows an example of Raman-based images obtained from a commercially available asbestos reference material. The particular type of asbestos reference material used in this study was amosite, with a representative

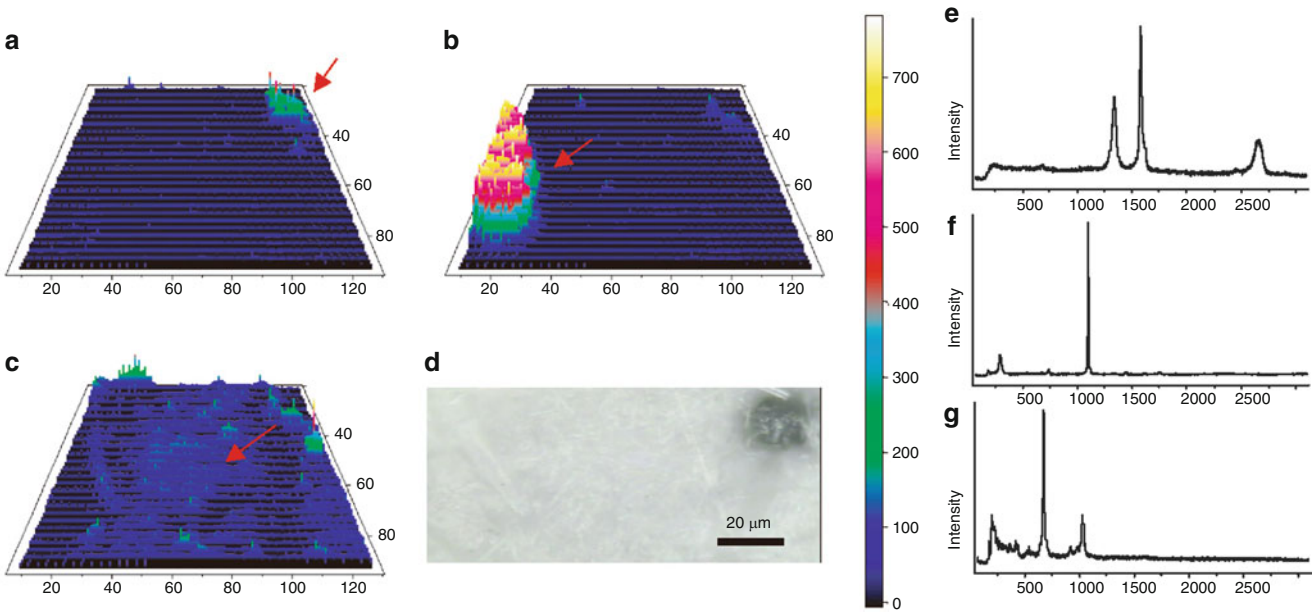


Fig. 32.5 Chemical and morphological imaging analysis of asbestos (amosite) material by Raman microprobe spectroscopy. (**a**) Three-dimensional Raman image corresponding to spectrum (**d**) (dust/charcoal particle); (**b**) three-dimensional Raman image corresponding to spectrum (**e**) (unidentified material); (**c**) three-dimensional Raman image corresponding to spectrum (**f**) (the asbestos component); (**d**) *white* light image of area analyzed by Raman spectroscopy. The laser excitation line at 632 nm was used with a 50× magnification objective

spectrum as indicated in Fig. 32.5g. The Raman chemical imaging analysis demonstrated the presence of two other components within the area investigated. The first component (upper right corner of the white light image, Fig. 32.5d) is identified as a dust/coal particle. The Raman image generated based on the spectrum of the particle (Fig. 32.5e) is shown in Fig. 32.5a, and it can be observed that dust/coal is present only in that region. A second impurity was also found (lower left corner of the white light image, Fig. 32.5d). This unidentified particle is difficult to distinguish from the asbestos material based on white light microscopy, but it is readily distinguished by Raman spectroscopy (Fig. 32.5b, f). The remaining part of the image area shows amosite crystals as observed both in the Raman and the white light images.

32.3.3 Infrared Spectroscopy

In addition to Raman microprobe spectroscopy, infrared (IR) spectroscopy is another useful vibrational technique for the *in situ* identification of mineral materials in tissues. IR spectroscopy is based on the absorption of energy by vibrating chemical bonds (primarily stretching and bending motions). As discussed earlier in this chapter, Raman scattering results from the same type of transitions, but the selection rules are different so the weak bands in the IR may be strong in the Raman and *vice versa*; thus, IR and Raman spectroscopies are complementary and when used together can provide a powerful tool for chemical compositional analysis.

32.4 Electron Microscopy

Ultrastructural analysis allows minerals to be analyzed in an efficient and sensitive manner. The scanning electron microscope is a versatile approach for the study of microinclusions in tissues, as it has the capability of examining objects at low and high magnifications with substantial depth of field and a three-dimensional appearance. SEM combined with EDXA is a very useful procedure for identifying mineral substances even though EDXA provides information on elemental composition rather than specific compounds.

To prepare samples for SEM–EDXA of tissue sections, 4- to 6-μm-thick sections are placed on pure carbon disks. Prior to mounting the section on the disc, the carbon disk is washed with concentrated H_2SO_4 (to remove iron), thoroughly rinsed with distilled deionized water, placed in acetone, and ultrasonicated for 2 h. The carbon disk is then washed with distilled deionized water again and placed in a vacuum oven to dry for 12 h. Prior to SEM–EDXA, tissue sections mounted on the disks are deparaffinized with xylene and absolute ethyl alcohol. A Hitachi S-3500 N scanning electron microscope with one energy-dispersive spectrometer (EDS) and KEVEX software were used to examine the tissue sections discussed here. An accelerating voltage of 20 keV and a 90-nA current were used for most of the measurements. Backscattered electron images at magnifications from 60 to 6,000× were used to observe morphologic characteristics of tissue and to record the composition of the inorganic material.

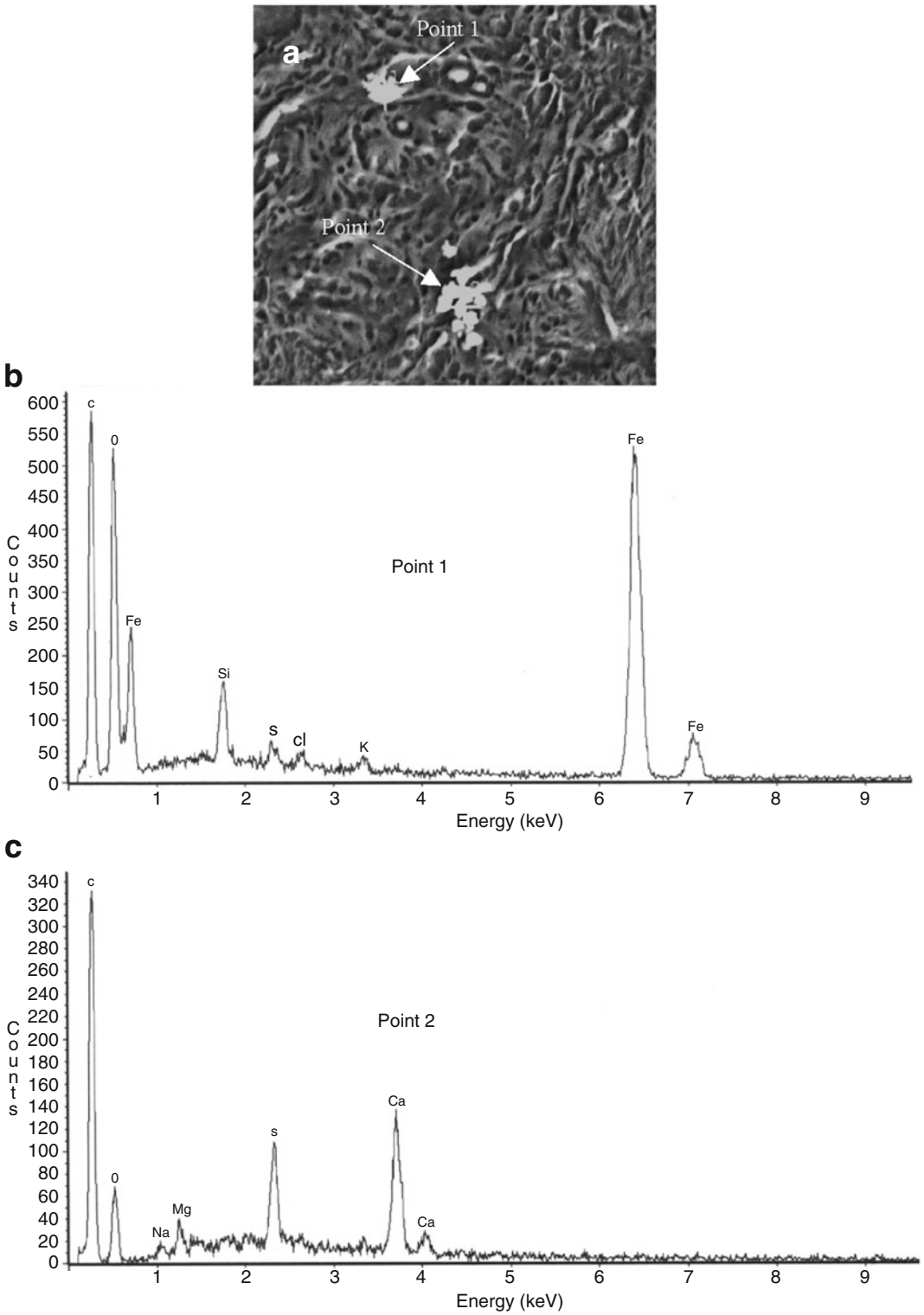


Fig. 32.6 Scanning electron photomicrograph showing several particles in lung tissue (**a**). The EDXA spectrum demonstrates the presence of exogenous silicon (**b**), and endogeneous sulfur and calcium elements were found within the lung tissue (**c**). This case is from a 37-year-old male with a history of occupational exposure as a coal miner

Figure 32.6a is an SEM image of foreign material in lung tissues in which two clusters of particles can be seen. The EDXA spectrum is shown for clusters 1 and 2. Although the two particle clusters are similar in appearance in the SEM image, their elemental compositions as obtained by EDXA are different. Cluster 1 shows the presence of silicon, in addition to iron, sulfur, chlorine, potassium, oxygen, and carbon (Fig. 32.6b). The silicon, potassium, and iron suggest

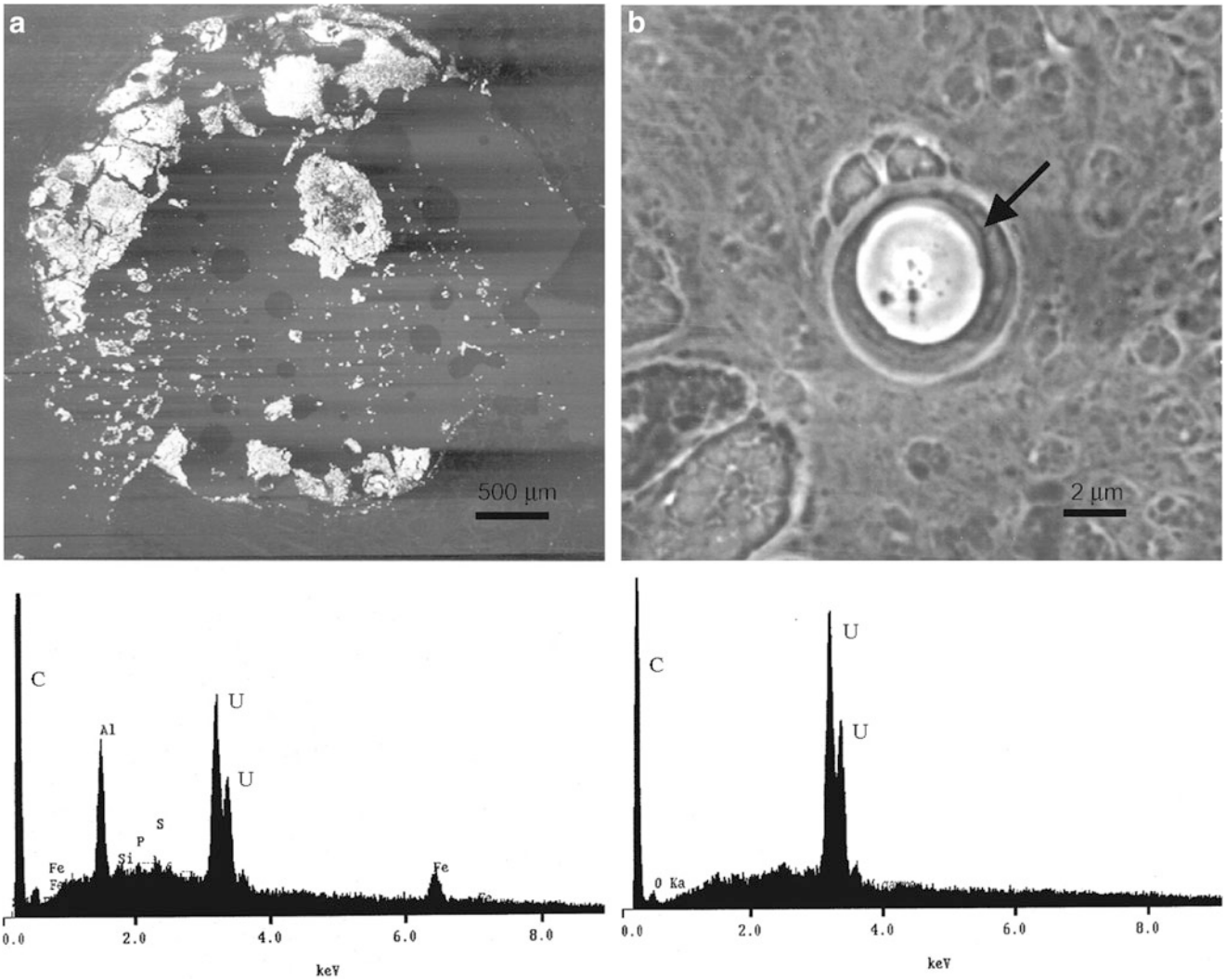


Fig. 32.7 (**a**) Scanning electron photomicrograph of uranium in a subcutaneous soft tissue biopsy. (**b**) The EDXA spectrum of uranium at an energy of approximately 3.8 keV, as well as other elements, including aluminum and iron (magnification 20×). When using a higher magnification (200×), only the presence of uranium was detected within the subcutaneous soft tissue section

the possibility of the presence of silicates, which are exogenous materials (*i.e.*, they are not produced by living organisms). Cluster 2 is composed of calcium, sulfur, sodium, magnesium, oxygen, and carbon. All of these elements participate in a variety of biological processes or act as building blocks of tissue; thus cluster 2 is composed of endogenous material (*i.e.*, it might have been produced by the living tissue). Another example is shown in Fig. 32.7a, which demonstrates the inclusion of uranium in a case of a subcutaneous soft tissue. The SEM–EDXA was obtained at two magnifications: 20× and 200×. At the lower magnification, the x-ray microanalysis demonstrated the presence of not only uranium but also aluminum and iron. At higher magnification, only uranium was detected. Energy-dispersive x-ray microanalysis allows us to examine the elemental composition of samples and to identify whether a material is of foreign or natural origin.

32.5 Conclusion

Each of the techniques described in this chapter—histochemical, Raman and infrared spectroscopies, and ultrastructural characterization with elemental analysis—has its advantages and limitations in chemical characterization of geochemical and medically relevant inclusions in tissue. Histochemical methods offer an inexpensive and simple approach, but they provide limited chemical information for complex biological systems. Light microspectroscopies have proven to be an excellent source for the chemical identification of tissue samples and other materials with or without minimal preparation, but instrumentation is expensive and requires the attention of well-trained technicians. SEM with x-ray microanalysis is extremely useful in morphological and elemental analyses, but sample

preparation is extensive and instrumentation is expensive. No one technique, then, is highly superior when compared to the others; the choice of a method will depend on the particular sample, the instrumentation available, the type of information needed, and the expertise of the researcher.

See Also the Following Chapters. Chapter 14 (Water Hardness and Health Effects) • Chapter 31 (Inorganic and Organic Geochemistry Techniques)

Further Reading

Carden A, Morris MD (2000) Application of vibrational spectroscopy to the study of mineralized tissues (review). J Biomed Opt 5 (3):259–268

Centeno JA, Kalasinsky VF, Johnson FB, Vihn TN, O'Leary TJ (1992) Fourier transform infrared microspectroscopy identification of foreign materials in tissue sections. Lab Invest 292(2):624–628

Centeno JA, Ishak KG, Mullick FB, Gahl WA, O'Leary TJ (1994) Infrared microspectroscopy and laser Raman microprobe in the diagnosis of cystinosis. Appl Spectrosc 48(5):569–572

Centeno JA, Mullick FG, Panos RG, Miller FW, Valenzuela-Espinoza A (1999) Laser-Raman microprobe identification of inclusions in capsules associated with silicone gel breast implants. Mod Pathol 12(7):714–721

Centeno JA, Mullick FG, Martinez L, Page NP, Gibb H, Longfellow D, Thompson D, Ladich ER (2002) Pathology related to chronic arsenic exposure. Environ Health Perspect 110(5):883–886

Diem M (1999) Introduction to modern vibrational spectroscopy, 1st edn. Wiley, New York

Gandolfi G (1967) Discussion on methods to obtain X-ray powder patterns from a single crystal. Miner Petrogr Acta 13:67–74

Gillie JK, Hocklowski J, Arbuckle GA (2000) Infrared spectroscopy. Anal Chem 72(2):71R–79R

Goldstein JI, Newbury D, Joy D, Lyman C, Echlin P, Lifshin E, Sawyer L, Michael J (2002) Scanning electron microscopy and X-ray microanalysis, 1st edn. Plenum Press, New York

Hanlon EB, Manoharan R, Koo T-W, Shafer KE, Motz JT, Fitzmaurice M, Kramer JR, Dasari RR, Feld MS (2000) Prospects for *in vivo* Raman spectroscopy. Phys Med Biol 45:R1–R59

Johnson FB, Pani K (1962) Histochemical identification of calcium oxalate. Arch Pathol 74:347–350

Ling J, Weitman SD, Miller MA, Moore RV, Bovik AC (2002) Direct Raman imaging techniques for study of the subcellular distribution of a drug. Appl Opt 41(28):6006–6017

McCreery RL (2000) Raman spectroscopy for chemical analysis, 1st edn. Wiley, New York

Mulvaney SP, Keating CD (2000) Raman spectroscopy. Anal Chem 72(2):145R–157R

Pestaner JP, Mullick FG, Johnson FB, Centeno JA (1996) Calcium oxalate crystals in human pathology: molecular analysis with the laser Raman microprobe. Arch Pathol Lab Med 120:537–540

Prophet EB, Mills B, Arrington JB, Sobin JB, Sobin LH (eds) (1995) Armed forces institute of pathology special staining techniques. American Registry of Pathology, Washington, DC

Prospero JM (2001) African dust in America. Geotimes 46(11):24–27

Shafer-Peltier KE, Haka AS, Motz JT, Fitzmaurice M, Dasari RR, Feld MS (2002) Model-based biological Raman spectral imaging. J Cell Biochem Suppl 39:125–137

Uzunbajakava N, Lenferink A, Kraan Y, Willekens B, Vrensen G, Greve J, Otto C (2003) Nonresonant Raman imaging of protein distribution in single human cells. Biopolymers 72(1):1–9

Willard HH, Merritt LL, Dean JA, Settle FA (1988) Instrumental methods of analysis, 7th edn. Wadsworth Publishing, Belmont

Modeling Groundwater Flow and Quality

33

Leonard F. Konikow and Pierre D. Glynn

Contents

33.1 Introduction

In most areas, rocks in the subsurface are saturated with water at relatively shallow depths. The top of the saturated zone—the water table—typically occurs anywhere from just below land surface to hundreds of feet below the land surface. Groundwater generally fills all pore spaces below the water table and is part of a continuous dynamic flow system, in which the fluid is moving at velocities ranging from feet per millennia to feet per day (Fig. 33.1). While the water is in close contact with the surfaces of various minerals in the rock material, geochemical interactions between the water and the rock can affect the chemical quality of the water, including pH, dissolved solids composition, and trace-elements content. Thus, flowing groundwater is a major mechanism for the transport of chemicals from buried rocks to the accessible environment, as well as a major pathway from rocks to human exposure and consumption. Because the mineral composition of rocks is highly variable, as is the solubility of various minerals, the human-health effects of groundwater consumption will be highly variable.

Groundwater provides about one-third of the public water supply and 42% of the irrigation water in the United States (Kenny et al. 2009). Also, most of the rural population in the United States—more than 40 million people—supply their own drinking water from domestic wells (Alley et al. 1999). Consequently, groundwater is considered an important source of drinking water in every State (Fig. 33.2). Nearly all surface-water features (streams, lakes, reservoirs, wetlands, and estuaries) interact with groundwater. Groundwater is the source of base flow to streams and rivers (see Fig. 33.1) and often is the primary source of water that sustains a wetland habitat.

It is long recognized that the chemical content of drinking water can affect human health beneficially or adversely (Keller 1978). Although the potential side effects associated with some trace elements of natural origin (e.g. arsenic, selenium) or anthropogenic origin (e.g., hexavalent

L.F. Konikow (✉) • P.D. Glynn
United States Department of the Interior, United States Geological Survey, Reston, VA 20192, USA
e-mail: lkonikow@usgs.gov

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5_33,

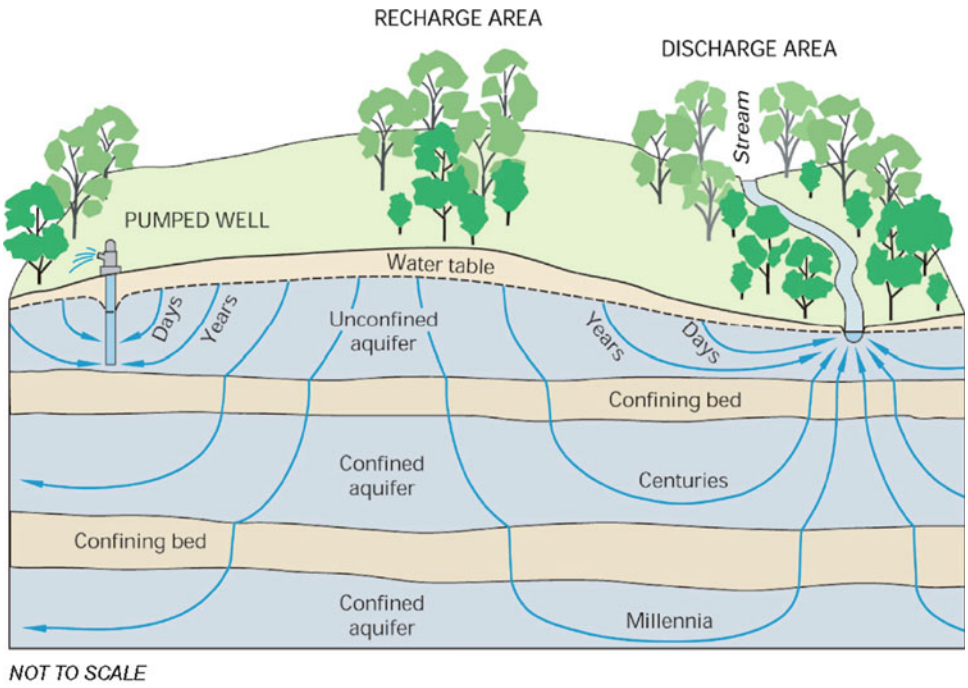


Fig. 33.1 Groundwater flow paths vary greatly in length, depth, and travel time from points of recharge to points of discharge in the groundwater system. Flow lines typically are perpendicular to lines (or surfaces) of equal hydraulic head (From Winter et al. 1998)

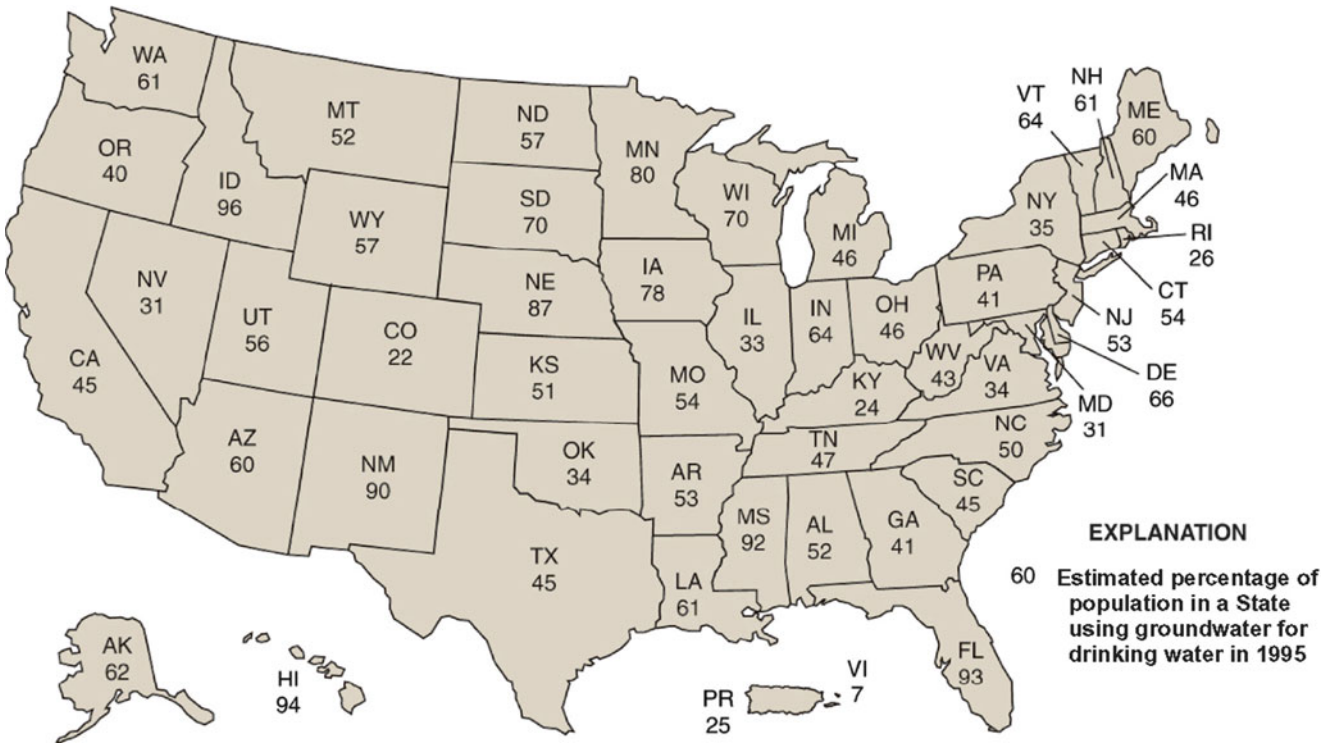


Fig. 33.2 Groundwater is an important source of drinking water for every State in the United States (From Alley et al. 1999)

chromium, organic compounds) present in concentrations exceeding public health standards for human consumption of drinking water has received the most attention in recent years, it is important to realize that many trace elements are greatly beneficial to human health (Hopps and Feder 1986).

It is clear that the chemical content of natural waters varies greatly from place to place. Geology has a stronger and more direct effect on the quality of groundwater than of surface water. Surface-water sources include a much greater component of direct precipitation and represent the rapid

integration of water derived from large and diverse source areas within a drainage basin than do groundwater sources; surface water also includes a smaller component reflecting the geochemical environment of the watershed. Furthermore, public-water supply systems based primarily on surface-water sources typically include a large distribution network and centralized treatment and monitoring facilities. Conversely, groundwater has much greater direct contact with the geochemical environment (e.g., mineral surfaces) during its slow migration and long residence time through the void spaces of the rocks that compose an aquifer system. Therefore, prior to its collection and distribution into a water supply system, the chemical content of groundwater will have been strongly affected by the geochemical environment in the rocks along the flow paths feeding wells or springs to which groundwater discharges. Many groundwater supply systems are small domestic systems designed to supply individual homes; these systems often are monitored on a minimal basis for chemical constituents in the water, especially for trace elements.

Understanding the pathway of dissolved minerals from the source rock to the environment or to human consumption is critical for evaluating and remediating possible toxic hazards. Evaluation and remediation, in turn, requires an understanding of the processes and parameters that control rock-water interactions and groundwater flow and solute transport. Conceptual knowledge of these processes and parameters can be quantified and incorporated into generic deterministic models, which can be applied to site-specific problems and be used to predict the fate and transport of dissolved chemicals.

The purpose of this chapter is to review the state-of-the-art in deterministic numerical modeling of groundwater flow, solute transport, and geochemical reaction processes. This chapter is intended to describe the types of models that are available and how they may be applied to complex field problems. However, as this chapter is only a review, it cannot offer comprehensive and in-depth coverage of this complex topic; instead, it guides the reader to references that provide more details. Other chapters in this book covering elements in groundwater are, for example, Chaps. 11, 12, 13, 14, and 16.

33.2 Physical Processes

33.2.1 Groundwater Flow

It generally is assumed that the process of groundwater flow is governed by the relation expressed in Darcy's Law, which was derived in 1856 on the basis of the results of laboratory experiments on the flow of water through a sand column. Darcy's Law states that the groundwater flow rate (or specific discharge) is proportional to the hydraulic gradient (related to pressure and elevation differences) and to hydraulic conductivity, a property that depends on the characteristics of the porous media (such as grain size distribution or fractures) and the fluid (such as density and viscosity) (see Bear 1979).

Darcy's Law, however, has limits on its range of applicability. It was derived from experiments on laminar flow of water through porous material. Flow probably is turbulent or in a transitional state from laminar to turbulent near the intakes of large-capacity wells. Turbulent flows also may occur in rocks as a result of the development of fractures, joints, or solution openings. What commonly is done in determining flow in such situations is to ignore local or small-scale turbulence and assume that flow behaves as if it were laminar flow through porous media on the regional scale, and, thus, that Darcy's Law applies at that scale.

In some field situations, fluid properties such as density and viscosity may vary appreciably in space or time. This variation may occur where water temperature or dissolved-solids concentration changes greatly. When the water properties are heterogeneous and (or) transient, the relations among water levels in monitoring wells, hydraulic heads, fluid pressures, and flow velocities are not straightforward. In such cases, the flow equation is written and solved in terms of fluid pressures, fluid densities, and the intrinsic permeability of the porous media.

33.2.2 Advective Transport

The migration and mixing of chemicals dissolved in groundwater obviously will be affected by the velocity of the flowing groundwater. That is, chemical constituents dissolved in flowing water will tend to migrate with the water—the faster the water moves, the faster and further the solutes will migrate. This entrainment of dissolved chemicals is called advective transport.

The specific discharge calculated from Darcy's Law represents a volumetric flux per unit cross-sectional area. But flow does not pass through the solid grains of the rock, only through the void spaces. Thus, to calculate the actual seepage velocity of groundwater, one must account for the actual cross-sectional area through which flow is occurring. The latter is done by dividing the specific discharge by the effective porosity of the porous medium. The effective porosity of fractured crystalline rocks can be less than 0.01, whereas for unconsolidated sands and gravels it can exceed 0.30.

33.2.3 Hydrodynamic Dispersion

Controlled laboratory and field experiments show that observed solute concentrations in a flow field cannot be predicted adequately just on the basis of seepage velocity, even for nonreactive constituents. Instead, it is observed that some solute will arrive at a given location sooner than

predicted by the mean seepage velocity, whereas some solute arrives later than the mean velocity would indicate. That is, there is a spreading about the mean arrival time. Similarly, solute distribution will spread spatially with time and travel distance. This spreading and mixing phenomenon is called hydrodynamic dispersion. It results from molecular and ionic diffusion, and from mechanical dispersion arising from small-scale variations in the velocity of flow that cause the paths of solutes to diverge or spread from the average direction of groundwater flow (Bear 1979). The outcome is a transient, irreversible, mixing (or dilution) process affecting the concentration distribution of a solute species in an aquifer.

The rate of solute flux caused by hydrodynamic dispersion is expressed in a form analogous to Fick's Law of diffusion. This Fickian model assumes that the driving force is the concentration gradient and that the dispersive flux occurs in a direction from higher concentrations towards lower concentrations at a rate related to a constant of proportionality—the coefficient of hydrodynamic dispersion. However, this assumption is not always consistent with field observations and is the subject of much ongoing research and field study (see, for example, Gelhar et al. 1992). The coefficient of hydrodynamic dispersion is defined as the sum of mechanical dispersion and molecular diffusion (Bear 1979). Mechanical dispersion is a function both of the intrinsic properties of the porous medium (expressed as a dispersivity coefficient, which is related to variability in hydraulic conductivity and porosity) and of the fluid flow (specifically, the fluid velocity). Molecular diffusion in a porous medium will differ from that in free water because of the effects of tortuous paths of fluid connectivity in porous media.

In most groundwater transport model applications, the dispersivity is defined in terms of just two unique constants—the longitudinal dispersivity of the medium and the transverse dispersivity of the medium. In practice, however, dispersivity values appear to be dependent on and proportional to the scale of the measurement. Field-scale dispersion (commonly called macrodispersion) results from large-scale spatial variations in hydraulic properties and seepage velocity. Consequently, the use of values of dispersivity determined for one scale of transport in a model designed to predict concentration changes over a different scale of travel probably is inappropriate. Overall, the more accurately and precisely a model can represent or simulate the true velocity distribution in space and time, the less of a problem will be the uncertainty concerning representation of dispersion processes.

33.2.4 Solute-Transport Equation

A generalized form of the solute-transport equation is presented by Bear (1979). The governing partial differential equation relates the change in concentration over time in a groundwater system to (1) hydrodynamic dispersion, (2) advective transport, (3) the effects of mixing with a source fluid that has a different concentration than the groundwater at the location of the recharge or injection, and (4) all of the physical, chemical, geochemical, and biological reactions that cause transfer of mass between the liquid and solid or air phases or conversion/decay of dissolved chemical species from one form to another. The chemical attenuation of inorganic chemicals can occur by sorption/desorption, precipitation/dissolution, or oxidation/reduction; organic chemicals can adsorb or degrade by microbiological processes and (or) volatilazation.

There has been considerable progress over the last 20 years in modeling reactive-transport processes; however, the complexity and computational requirements for solute-transport models and reaction models are each intense and, therefore, applications of coupled multi-species reactive-transport models are rare. Although some research tools are documented (e.g., see Lichtner et al. 1996, and Ibaraki and Therrien 2001), they still are at the leading edge of the state-of-the-art and usually require computational resources and input data that are beyond that available for most applications. Thus, for analysis of field problems, it is much more common to apply only groundwater flow and solute-transport models or only multispecies geochemical reaction models, or to apply both sequentially.

In summary, the mathematical solute-transport model requires at least two partial differential equations. One is the equation of flow, from which groundwater flow velocities are obtained, and the second is the solute-transport equation, which describes the chemical concentration in groundwater. If the properties of the water are affected significantly by changes in solute concentration, as in a saltwater intrusion problem, then the flow and transport equations should be solved simultaneously (or at least iteratively). If the properties of the water remain constant and uniform, then the flow and transport equations can be decoupled and solved sequentially, which is simpler numerically.

33.3 Geochemical Processes

33.3.1 Basic Concepts

Thermodynamic models describing aqueous chemical reactions or describing reactions between the aqueous phase and other phases (solid, gas, or surficial phases) use various common basic principles. Thermodynamic models assume that ion activities, which can be considered the "thermodynamically effective" concentrations determining the progress and direction of reactions, can be calculated from measured (or calculated) ion concentrations.

In addition to the calculation of ionic activities, thermodynamic models typically are based on the law of mass action. This law states that as the activity of the reactants is increased in a chemical reaction, the activity of the reaction products also will increase if equilibrium is maintained. For example, if the following chemical reaction is at equilibrium, $iA + jB \Leftrightarrow kC + lD$ where i, j, k, and l are stoichiometric coefficients, the law of mass action states that the following mathematical equation must hold:

$$K = \frac{a_C^k \, a_D^l}{a_A^i \, a_B^j} \tag{33.1}$$

where K is a constant (at a fixed temperature and pressure) and a_A, a_B, a_C, and a_D represent the activities of A, B, C, and D at thermodynamic equilibrium. The law of mass action applies to any kind of chemical reaction at thermodynamic equilibrium, not just to aqueous reactions.

33.3.2 Aqueous Speciation, Hydrolysis, and Oxidation/Reduction Reactions

The calculation of solution ionic strength, ionic activity coefficients, and ionic activities must be conducted iteratively with the solution of the mass-action equations that determine the speciation of the aqueous solution. Aqueous speciation is the partitioning of chemical constituents present in a solution (whose total concentrations are typically measured) into various aqueous species that represent the different molecular forms assumed by the constituents in the aqueous solution. Aqueous speciation reactions are homogeneous reactions—all reactants and products are aqueous species. Two examples of speciation reactions are:

$$\mathrm{Al^{3+}} + 3\mathrm{H_2O} \;\Leftrightarrow\; \mathrm{Al(OH)_2^+} + 2\mathrm{H^+} + \mathrm{H_2O}\ (\text{hydrolysis of aluminum}) \tag{33.2a}$$

$$6\mathrm{CN^-} + \mathrm{Fe^{2+}} \Leftrightarrow \mathrm{Fe(CN)_6^{4-}}\ (\text{cyanide complexation by ferrous iron}) \tag{33.2b}$$

Simple mass-action equations describing the relative aqueous activities of products and reactants can be written for any of the above reactions. Aqueous speciation reactions control the concentrations of individual aqueous species, and, thereby, may affect appreciably the toxicity of a solution. Some aqueous species can be much less toxic than others. For example, cyanide present in the complexed $\mathrm{Fe(CN)_6^{4-}}$ form is much less toxic than cyanide in the $\mathrm{CN^-}$ or $\mathrm{HCN^0}$ forms. Aqueous speciation reactions also affect the total concentration of constituents in solution, through their control of general solution characteristics such as ionic strength, acidic nature (pH), and redox potential (pe) of the solution, and through their control of individual aqueous species involved in mass-transfer reactions (e.g. mineral precipitation/dissolution, surface sorption/desorption, ingassing/exsolution reactions).

The pH and pe of an aqueous solution are sometimes described as the *master variables* controlling the speciation of aqueous solutions. The pH of a solution simply relates to the acidic nature of a solution, and more specifically to the activity of protons ($\mathrm{H^+}$) or equivalently of hydronium ions ($\mathrm{H_3O^+}$ or $\mathrm{H_9O_4^+}$) in the solution, as $pH = -\log a_{H+}$. The pe of a solution is related to the ratio of the activity of an aqueous species present in oxidized form to that present in reduced form. Mathematically, the definition of the pe of a solution is analogous to that of the pH variable, except that it is defined in terms of the activity of free electrons in the solution, as $\mathrm{pe} = -\log a_{e-}$. For all practical purposes, however, any electrons produced by an oxidation reaction always must be consumed by a reduction reaction. Nevertheless, the redox potential of a solution does have practical relevance in describing the degree to which aqueous species are in oxidized or reduced form.

The Eh of a solution is the redox potential measured in the field. It is directly related to pe by the relation:

$$\mathrm{pe} = \mathrm{Eh}\frac{F}{2.303RT} \tag{33.3}$$

where F is the Faraday constant, R is the gas constant, and T is the temperature in Kelvin. At 25 ° Celsius, the relation is $\mathrm{pe} = 16.904\ \mathrm{Eh}$, where Eh is expressed in volts.

Field measurements of Eh are problematic. First, the redox-active species present in a water often are not in redox equilibrium and, therefore, the measurement may be meaningless; that is, more than one redox potential may be present for the solution depending on which redox couple (e.g. $\mathrm{Fe^{3+}/Fe^{2+}}$) is chosen, and the measurement may at best represent some sort of "mixed" potential. Second, the only redox-active species to which platinum electrodes (typically used for Eh measurements) have been demonstrated to respond quickly, and, therefore, reflect their electrochemical equilibrium, are $\mathrm{Fe^{2+}}$, $\mathrm{Fe^{3+}}$, and $\mathrm{S^{2-}}$, and only when these species are present at concentrations of 10^{-5} moles/L or greater (Nordstrom and Munoz 1994). In general, redox equilibria and disequilibria in groundwater are best assessed through the measurement of individual redox couples, rather than through measurements of Eh.

Oxidation/reduction reactions can occur either in the aqueous phase only (homogeneous reactions) or between the aqueous phase and other phases (heterogeneous reactions). In most groundwaters, the presence of organic carbon commonly drives a sequence of redox reactions as

water migrates from the unsaturated zone and water table to greater depths. Typically, organic carbon reduces dissolved oxygen in the water, then reduces nitrate to nitrogen gas (and sometimes to ammonia). Dissolved organic carbon also may react with manganese oxide minerals producing Mn^{2+} in the water; at slightly lower oxidation potentials the organic carbon will react with ferric-iron minerals (typically oxides) and generate dissolved ferrous iron in the water. At greater depths, water becomes sulfidic as the carbon starts to reduce dissolved sulfate to sulfide. Finally, when no further electron acceptors (such as SO_4, O_2, and NO_3) are present, water typically becomes methanic, i.e. remaining organic matter decomposes through a process of fermentation to methane and carbon dioxide. In this process, hydrogen generally is produced as an intermediate product. Despite strong thermodynamic potentials for their occurrence, rates of redox reactions are often slow unless microbially catalyzed. Most redox reactions in natural and contaminated environments are catalyzed microbially.

Redox reactions are important in medical geology, whether for natural or contaminated environments, because they affect the relative toxicity of various dissolved constituents in water. For example, Cr(IV) is a suspected carcinogen, whereas Cr(III) is an essential trace element for humans. Redox reactions also affect the solubility of various compounds (e.g. metal sulfides present in the rock materials).

33.3.3 Geochemical Mass-Transfer Processes

Mineral dissolution and precipitation processes are important in controlling the chemical evolution of groundwater. These processes strongly affect the overall chemical characteristics of the water through their effect on pH and pe conditions, ionic strength, and complexant concentrations (dissolved carbonate, sulfate, chloride, etc.). For example, the pH of natural waters often is buffered by the dissolution of calcite and described as $CaCO_{3,s} + H^+ \Leftrightarrow Ca^{2+} + HCO_3^-$.

Mineral precipitation processes also commonly limit the concentrations of many constituents in water. For example, Ba and Al concentrations in water often are limited by the precipitation of barite and aluminum hydroxide:

$$Ba^{2+} + SO_4^{2-} \Leftrightarrow BaSO_{4,s} \tag{33.4a}$$

$$Al(OH)^{2+} + 2H_2O \Leftrightarrow Al(OH)_{3,s} + 2H^+. \tag{33.4b}$$

Although the above reactions are written for pure minerals, mineral phases invariably contain foreign ions and impurities, which were entrained as occlusion pockets during the formation of the minerals or are substituting as an integral part of the mineral lattice. In either case, minerals can take up and (or) release these impurities through recrystallization processes. The thermodynamic theory describing the uptake and release of substitutional impurities in minerals (also known as solid solutions) is fairly complex and remains an area of active research (Glynn 2000), but is being increasingly implemented in geochemical modeling codes. Examples of solid-solutions reactions that may control trace-element concentrations include (1) the uptake of Cu, Ni, Co, and Zn by precipitating manganese oxides, (2) the uptake of chromate by barite recrystallization in contaminated waters, and (3) the control of fluoride concentrations through dissolution and recrystallization uptake of apatites. Biogenic apatites, such as found in fossil bones and teeth, commonly are initially rich in hydroxylapatite and slowly recrystallize upon contact with groundwater to fluoroapatite. In certain groundwater systems, however, the reverse process also has been demonstrated to occur. For example, Zack (1980) has shown that the exchange of hydroxide ions for fluoride present in fossil shark teeth is responsible for anomalously high fluoride concentrations in the Atlantic Coastal Plain aquifers of South Carolina (and is discussed in more detail below).

Sorption reactions often are important in controlling the concentrations of constituents in groundwater and often may even affect the observed pH. A typical porous medium aquifer (with a porosity of 0.2 and a cation exchange capacity of 5 meq/100 g) has about 500 meq of cation exchange capacity per liter of water (Drever 1997). This value is more than two orders of magnitude greater than the concentration of dissolved ions in dilute groundwater, and consequently can be expected to have a large effect on the chemistry of the water.

Sorption reactions generally are described either through an ion-exchange model (primarily affecting cations) or through a surface-complexation model. Ion-exchange models typically apply to mineral surfaces and interlayers with constant surface charge (e.g. clays and zeolites), and usually only consider cation exchange reactions such as $[Na - clay + K^+ \Leftrightarrow Na^+ + K - clay]$.

Surface-complexation models commonly are used to describe the sorption of aqueous species on surfaces with variable charge (e.g. iron and manganese oxides, silica, organic matter, and clay edges). These surfaces become more negatively charged with increasing pH, and, therefore, their cation sorption capacity increases and their anionic sorption capacity decreases. At any pH, the surfaces are considered to contain a mix of positively charged, negatively charged, and neutral sites. This mix and the ensuing chemical reactions among the various sites and aqueous ions and complexes are fully described by a speciation of the surface, in a manner analogous to that for an aqueous solution. This

speciation describes the surface as a series of various surface-complexes and bare-surface sites of different charges.

Gas dissolution/exsolution/volatilization reactions can affect the concentrations of organic (aromatic and light aliphatic compounds) and inorganic (O_2, CO_2, N_2, noble gases) constituents in groundwater near the water table. The reactions can strongly affect the general chemistry of water through their effect on pH, redox potentials, and ionic strength; in some cases (e.g. volatilization of HCN and light organic compounds), the reactions also can directly affect the concentrations of contaminants in water.

33.3.4 Biodegradation/Biotransformation

Microorganisms are important in the chemical evolution of waters, and, for all practical purposes, can be considered present in almost every groundwater environment, even under extreme conditions. Microorganisms have been found underground at depths of more than a kilometer (Pedersen 1993), at temperatures as high as 110°C (Stetter 1998), and in waters with up to 30% salinity (Grant et al. 1998). Most microorganisms are heterotrophic and use organic carbon as a primary energy source. However, chemolithotrophic organisms can use reduced inorganic substrates, such as NH_4, H_2, H_2S, and CH_4, to derive energy in both aerobic and anaerobic environments.

Microbes also are essential in the degradation of organic molecules, generally of complex molecules to simpler ones, ultimately to inorganic compounds and forms of C, N, H, S, Cl, and other elements. Complete transformation to inorganic compounds (mineralization) involves multiple, successive, biologically mediated reactions, which may proceed at different rates. Although the initial degradation rate of an organic molecule may be fast, degradation of some of its metabolites may be slow, which can be a problem if the metabolite is associated with a health risk. Although microbial "remediation" is an important issue being considered in investigations of anthropogenic groundwater contaminants (such as pesticides, herbicides, and petroleum products), microbes also are likely to affect the "natural contamination" of groundwater through their catalysis of coal-water interactions and their consequent mobilization of soluble polar aromatic and polynuclear aromatic hydrocarbons. These compounds are thought to be an important factor in the observed incidences of Balkan endemic nephropathy (Feder et al. 1991).

Microbes also are known to catalyze many inorganic reduction reactions—generally using organic carbon as a reducing agent. Examples from naturally occurring subsurface constituents, which when modified by microorganisms may become more mobile, include: solid Fe(III)-oxyhydroxides to dissolved Fe(II), nitrate to N_2 (denitrification) or further to ammonia, sulfate to sulfide, As (V) to As(III) (Dowdle et al. 1996), Se(VI) to Se(IV) and Se (0) (Switzer Blum et al. 2001), and U(VI) to U(IV) (Lovley et al. 1991). Microbes also are known to catalyze many of the reverse oxidation reactions in the list above. Redox reactions usually are associated with large changes in free energy, and therefore provide microbes with an energy source.

Although the occurrence of these redox reactions can be predicted from thermodynamic considerations, in practice the kinetics of the reactions would be orders of magnitude slower if they were not mediated by microbes. Numerical modeling of biodegradation/biotransformation reactions reduces the complexity of the multiple chemical, enzymatic, biological, and ecological processes that are mediating the transformation of a constituent of interest to a simple mathematical description of the overall transformation kinetics. The mathematical model chosen often considers not just the degradation or transformation of a particular compound, but also keeps track of the effect of the transformation on the size and productivity of the microbial community responsible for the catalysis of the transformation. "Monod" and "Michaelis-Menten" kinetic models are used to describe microbial utilization of chemical substrates and microbial growth kinetics (Schwarzenbach et al. 1993). The computer codes BIOMOC (Essaid and Bekins 1997) and RT3D (Clement 1997; Johnson et al. 2006) are examples of groundwater flow and transport codes that allow the simulation of biodegradation and transformation reactions using a variety of kinetic model formulations.

A promising new computational approach has also recently been used (e.g. Fang et al. 2011; Scheibe et al. 2009) that couples groundwater reactive transport models with constraint-based genome-scale microbial *in-silico* (i.e. computational) models and avoids the fitting of empirical microbial kinetic models. The microbial *in-silico* models are derived from genomic information and data on multiple-pathway reaction networks. They can simulate the metabolisms of well studied microorganisms and identify rates of growth, substrate uptake, and byproduct formation under a diversity of environmental conditions. The potential metabolic reactions are mapped based on organism genome sequences and available physiological data. Subsequently, an optimization objective is chosen (e.g. growth maximization, minimization of resource utilization) for given metabolic functions. Lastly, the system is constrained by calculating for a given environmental condition the optimal fluxes of all metabolites through the possible metabolic reactions. This approach has already been shown useful, for example for modeling field studies of uranium bioremediation (Scheibe et al. 2009); the approach benefits from the rapidly increasing number of genome sequences (and other genomic and proteomic information) available for groundwater environments.

33.4 Models

33.4.1 Overview

The word *model* has many definitions. A model perhaps is most simply defined as a representation of a real system or process. A *conceptual model* is a hypothesis for how a system or process operates. This hypothesis can be expressed quantitatively as a mathematical model. *Mathematical models* are abstractions that represent processes as equations, physical properties as constants or coefficients in the equations, and measures of state or potential in the system as variables.

Most groundwater models in use today are deterministic mathematical models. *Deterministic models* are based on conservation of mass, momentum, and energy and describe cause and effect relations. The underlying assumption is that, given a high degree of understanding of the processes by which stresses on a system produce subsequent responses in that system, the system's response to any set of stresses can be predetermined, even if the magnitude of the new stresses falls outside of the range of historically observed stresses.

Deterministic groundwater models generally require the solution of partial differential equations. Exact solutions often can be obtained analytically, but *analytical models* require that the parameters and boundaries be highly idealized. Some deterministic models treat the properties of porous media as lumped parameters (essentially, as a black box), but this precludes the representation of heterogeneous hydraulic properties in the model. Heterogeneity, or variability in aquifer properties, is characteristic of all geologic systems and now is recognized as critical in affecting groundwater flow and solute transport. Thus, it often is preferable to apply distributed-parameter models, which allow the representation of more realistic distributions of system properties. Numerical methods yield approximate solutions to the governing equation (or equations) through the *discretization* of space and time. Within the discretized problem domain, the variable internal properties, boundaries, and stresses of the system are approximated. Deterministic, distributed-parameter, *numerical models* can relax the rigid idealized conditions of analytical models or lumped-parameter models, and, therefore, they can be more realistic and flexible for simulating field conditions (if applied properly).

The number and types of equations to be solved are determined by the concepts of the dominant governing processes. The coefficients of the equations are the parameters that are measures of the properties, boundaries, and stresses of the system; the dependent variables of the equations are the measures of the state of the system and are determined mathematically by the solution of the equations. When a numerical algorithm is implemented in a computer code to solve one or more partial differential equations, the resulting computer code can be considered a *generic model*. When the grid dimensions, boundary conditions, and other parameters (such as hydraulic conductivity and storativity), are specified in an application of a generic model to represent a particular geographical area, the resulting computer program is a *site-specific model*. The capability of generic models to solve the governing equations accurately typically is demonstrated by example applications to simplified problems. This does not guarantee a similar level of accuracy when the model is applied to a complex field problem.

33.4.2 Numerical Methods

The partial differential equations describing groundwater flow and transport can be solved mathematically by using either analytical solutions or numerical solutions. In general, obtaining the exact analytical solution to the partial differential equation requires that the properties and boundaries of the flow system be highly and perhaps unrealistically idealized. Many of the limitations of applying analytical methods to complex field problems can be overcome by using analytical element methods, which apply analytical methods to subareas of the problem domain (see Haitjema 1995).

Alternatively, for problems where the simplified analytical models no longer describe the physics of the situation, the partial differential equations can be approximated numerically. In numerical approaches, the continuous variables are replaced with discrete variables that are defined at grid blocks or nodes. Thus, the continuous differential equation, which defines hydraulic head or solute concentration everywhere in the system, is replaced by a finite number of algebraic equations that defines the hydraulic head or concentration at specific points. This system of algebraic equations generally is solved using matrix techniques. This approach constitutes a numerical model.

Two major classes of numerical methods have come to be well accepted for solving the groundwater flow equation. These are finite-difference methods and finite-element methods. Each of these two major classes of numerical methods includes a variety of subclasses and implementation alternatives. An overview of the application of these numerical methods to groundwater problems is presented by Wang and Anderson (1982). Both of these numerical approaches require that the area of interest be subdivided by a grid into a number of smaller subareas (cells or elements) that are associated with nodal points (either at the centers or peripheries of the subareas).

Finite-difference methods approximate the first derivatives in the partial differential equations as difference quotients (the differences between values of the independent

variable at adjacent nodes with respect to the distance between the nodes, and at two successive time levels with respect to the duration of the time-step increment). Finite-element methods use assumed functions of the dependent variable and parameters to evaluate equivalent integral formulations of the partial differential equations. Huyakorn and Pinder (1983) present a comprehensive analysis of the application of finite-element methods to groundwater problems. In both numerical approaches, the discretization of the space and time dimensions allows the continuous boundary-value problem for the solution of the partial differential equation to be reduced to the simultaneous solution of a set of algebraic equations. These equations then can be solved using either iterative or direct matrix methods.

Each approach has advantages and disadvantages, but there are few groundwater problems for which one approach clearly is superior. In general, the finite-difference methods are simpler conceptually and mathematically, and are easier to program. They typically are keyed to a relatively simple, rectangular grid, which also eases data entry. Finite-element methods generally require the use of more sophisticated mathematics but, for some problems, may be more accurate numerically than standard finite-difference methods. A major advantage of the finite-element methods is the flexibility of the finite-element grid, which allows a close spatial approximation of irregular boundaries of the aquifer and (or) of parameter zones within the aquifer when they are considered. However, the construction and specification of an input data set are much more difficult for an irregular finite-element grid than for a regular rectangular finite-difference grid. Thus, the use of a graphical model preprocessor that includes a mesh generator should be considered. A hypothetical aquifer system with impermeable boundaries and a well field (Fig. 33.3a) has been discretized using finite-difference (Fig. 33.3b) and finite-element (Fig. 33.3c) grids. Grids can be adjusted to use a finer mesh spacing in selected areas of interest. The rectangular finite-difference grid approximates the aquifer boundaries in a step-wise manner, resulting in some nodes or cells outside of the aquifer, whereas sides of the triangular elements of the finite-element grid can follow closely the outer boundary using a minimal number of nodes.

33.4.3 Groundwater Flow Models

A major revolution in the quantitative analysis of groundwater flow systems came in the early 1970s with the introduction and documentation of two-dimensional, deterministic, distributed-parameter, digital computer simulation models. These models represented a major improvement over analytical methods because they allowed the representation of heterogeneous properties, complex boundary conditions, and time-varying stresses. As improved numerical methods were developed and more powerful computers became widely available, three-dimensional modeling became standard practice by the early 1990s. Practical aspects of applying groundwater models are reviewed by Anderson and Woessner (1992).

Groundwater flow models solve a governing partial differential equation. The solution defines the hydraulic-head distribution at every point within the boundaries of the problem domain. When this is accomplished using numerical methods, the solution inherently also provides the fluid fluxes throughout the discretization grid. Solving the flow equation requires the specification of the properties of the groundwater system (and their spatial variability), the boundary conditions, and, for transient problems, the initial conditions.

The knowledge of the heads (or water levels or fluid pressures) and the direction and rate of flow provides much insight into the nature of the groundwater flow system, and allows inferences to be made about (1) potential source areas for toxic substances detected in groundwater, and (2) potential discharge areas or receptors for flow and transport of dissolved toxic constituents away from known sources of soluble toxic substances. When problems are detected and analyzed using a simulation model, the model can be used as a management tool to help evaluate alternative decisions for reducing risks to public health or the environment.

The major difficulty in groundwater modeling is defining accurately the properties of the system and the boundary conditions for the problem domain. The subsurface environment is a complex, heterogeneous, three-dimensional framework. To determine uniquely the parameter distribution for a field problem, so much expensive field testing would be required that it is seldom feasible either economically or technically. Therefore, the model typically represents an attempt, in effect, to solve a large set of simultaneous equations with more unknowns than equations. It inherently is impossible to obtain a unique solution to such a problem. Therefore, limited sampling and understanding of the geological heterogeneity causes uncertainty in the model input data (aquifer properties, sources and sinks, and boundary and initial conditions). This uncertainty leads to nonuniqueness in the model solution.

Uncertainty in parameters logically leads to a lack of confidence in the interpretations and predictions that are based on a model analysis, unless the model can be demonstrated to be a reasonably accurate representation of the real system. To demonstrate that a deterministic groundwater simulation model is realistic, usually field observations of aquifer responses (such as changes in water levels for flow problems or changes in concentration for transport problems) are compared to corresponding model-calculated values. The objective of this calibration procedure is to minimize differences between the observed data and

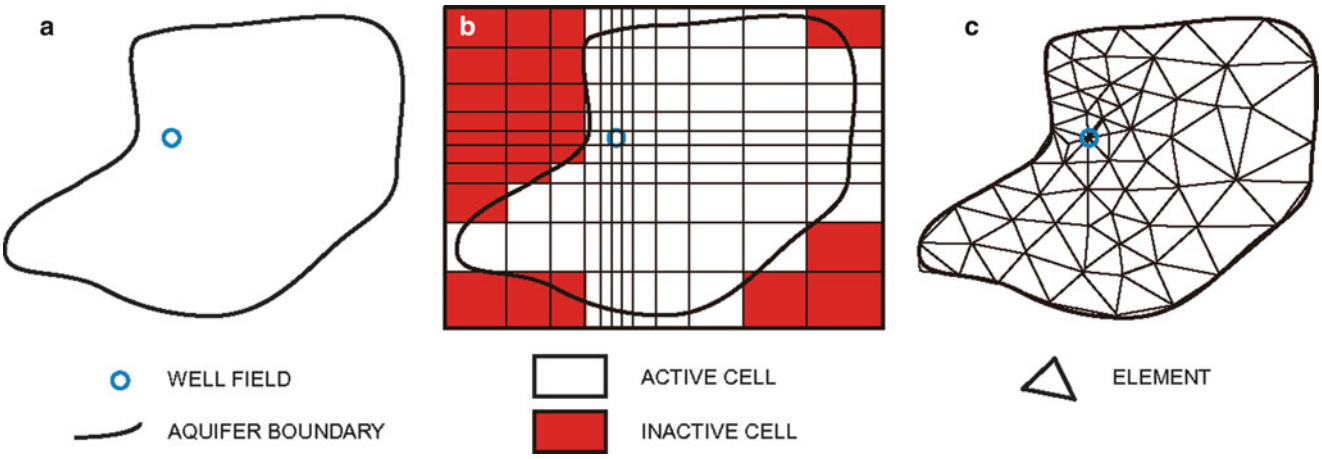


Fig. 33.3 Application of a numerical model to simple hypothetical problem, showing (**a**) an irregularly bounded aquifer discretized using (**b**) a finite-difference grid and (**c**) a finite-element grid (From Konikow and Reilly 1998)

calculated values. The minimization is accomplished by adjusting parameter values within their ranges of uncertainty until a best fit is achieved between the calculated values of dependent variables and the corresponding observations. Thus, model calibration often is considered a parameter-estimation procedure. Usually, the model is considered calibrated when it reproduces historical data within some acceptable level of accuracy. The level of acceptability is, of course, determined subjectively. Although a poor match provides evidence of model errors, a good match does not necessarily prove the validity or adequacy of the model (Konikow and Bredehoeft 1992).

The calibration of a deterministic groundwater model often is accomplished through a trial-and-error adjustment of the model input data to modify model output. Because a large number of interrelated factors affect the output, trial-and-error adjustment may be a highly subjective and inefficient procedure. Advances in automatic parameter-estimation procedures help to eliminate some of the subjectivity inherent in model calibration. The newer procedures, such as PEST (Doherty 2010), generally treat model calibration as a statistical procedure using multiple regression approaches. Parameter-estimation procedures allow simultaneous model construction, application, and calibration using uncertain data, so that the uncertainties in model parameters and in predictions and assessments can be quantified.

Automated parameter-estimation techniques improve the efficiency of model calibration and have two general components—one that calculates the best fit (sometimes called automatic history matching) and a second that evaluates the statistical properties of the fit. These techniques also are called *inverse models*, as they treat the system parameters as unknowns. The minimization procedure uses sensitivity coefficients that are based on the change in calculated value divided by the change in the parameter (for example, the change in head with changing transmissivity). The sensitivity coefficients may be useful in the consideration of additional data collection. Hill and Tiedeman (2007) provides an overview of methods and guidelines for effective model calibration using inverse modeling.

One of the most popular and comprehensive deterministic groundwater models available today is the U.S. Geological Survey's (USGS) MODFLOW code (McDonald and Harbaugh 1988; Harbaugh et al. 2000; Harbaugh 2005). This model actually is an integrated family of compatible codes that centers on an implicit finite-difference solution to the three-dimensional flow equation. The basic model uses a block-centered finite-difference grid that allows variable spacing of the grid in three dimensions. Flow can be steady or transient. Aquifer properties can vary spatially and hydraulic conductivity (or transmissivity) can be anisotropic. Flow associated with external stresses, such as wells, areally distributed recharge, evapotranspiration, drains, lakes, and streams, can also be simulated through the use of specified head, specified flux, or head-dependent flux boundary conditions. The implicit finite-difference equations can be solved using any one of several solution algorithms. Although the input and output systems of the program were designed to permit maximum flexibility, usability and ease of interpretation of model results can be enhanced by using one of several commercially available preprocessing and postprocessing packages; some of these operate independently of MODFLOW, whereas others are directly integrated into reprogrammed and (or) recompiled versions of the MODFLOW code.

A variety of other MODFLOW accessory codes, packages, and features are available. Most of these were developed by the USGS; examples include coupled surface-water and groundwater flow, aquifer compaction,

transient leakage from confining units, rewetting of dry cells, horizontal flow barriers, multi-node wells, graphical user interfaces, post-processing visualization, and a program that calculates water budgets. Other packages have been developed by non-USGS sources to work with MODFLOW; one example is the advective-dispersive solute-transport model MT3DMS (Zheng and Wang 1999; Zheng 2006).

The utility of groundwater flow modeling is illustrated by its application to a selenium problem in California, where more than two million acres of agricultural land is irrigated in the western San Joaquin Valley. Since 1967, imported surface water has been the primary source for irrigation; hence, groundwater pumping simultaneously declined (Belitz and Phillips 1992). This combination caused increased recharge to the underlying aquifers and subsequent water-table rises. By the early 1990s, the water table was high (within 10 ft of the land surface) over more than half of the area. Because such areas are prone to soil salinization and other problems, subsurface tile drains have been used to keep the water table deep enough to minimize these problems. However, the agricultural drainage water was high in selenium and eventually flowed into the Kesterson Wildlife Refuge—leading to deaths and deformities of waterfowl and aquatic biota (Deverel et al. 1984; Presser and Barnes 1985). These problems led to the closure of drains contributing selenium, which left considerable concern about how to manage the groundwater flow system in a manner that maintained agricultural productivity yet precluded selenium transport.

According to Presser et al. (1990), the source of the selenium is from weathering of pyritic marine shales in the Coast Ranges just west of the San Joaquin Valley. They believe that selenium is mobilized by oxidative weathering in an acidic environment, which concentrates the soluble selenate form of selenium. Selenate is transported readily in flowing groundwater and surface runoff.

A transient, three-dimensional, finite-difference model of the regional groundwater flow system was developed to assess water-table responses to alternative management that would affect groundwater recharge and discharge (Belitz et al. 1993). The model was calibrated using hydrologic data collected during 1972–1988. The model results indicate good agreement between measured and simulated depths to water (Fig. 33.4).

The calibrated model was used to evaluate the possible effects of various management practices on the depth of the water table (see Fig. 33.5 for representative results). The number of cells (each having an area of one square mile) subject to bare-soil evaporation is an indicator of the depth to water because only cells in which the water table is less than 7 ft deep will fall into this category. Higher water-table elevations also yield greater discharge to drains. If present practices are maintained, the area underlain by a high water table will continue to increase, as will discharge to drains. Reducing recharge (by increasing irrigation efficiency), increasing pumping, and removing land from agricultural production all will help to mitigate the problem. Thus, the groundwater flow model provides a powerful tool to help water managers mitigate the selenium problem while considering cost-benefit ratios.

33.4.4 Groundwater Pathline Models

Pathline models simulate the process of advective transport. They use calculated velocities to compute where and how fast water and nonreactive dissolved chemicals migrate. This requires the specification of an additional physical parameter—the porosity of the groundwater system (and its spatial variability). Also, the hydraulic-head gradients must be known, typically from the output of a groundwater flow model. It is useful for estimating where fluid and dissolved solutes are moving, how fast they are moving, and their source. They also can be useful for cross-checking age dates estimated from isotopic analyses. Pathline models, however, cannot calculate solute concentrations because dilution or reaction mechanisms are not included in these models. Pathline models usually are much more efficient to run than transport models.

A widely used pathline program is MODPATH (Pollock 1989), which uses MODFLOW model output and determines paths and travel times of water movement under steady-state and transient conditions. The semianalytical particle-tracking method assumes that each directional velocity component varies linearly within a grid cell in its own coordinate direction. For example, MODPATH was used to delineate source areas contributing recharge to a well in Minnesota (Fig. 33.6). The results depict the complicated and discontinuous spatial patterns of contributing recharge areas to wells in highly developed aquifer systems and would be extremely difficult to derive without the aid of such a model. If toxic constituents are detected in a particular well (such as well 11 in Fig. 33.6), the pathline model would help delineate the volume of rock material with which groundwater was in contact at earlier times—an invaluable aid in a search for the source of the toxic constituents.

33.4.5 Advection-Dispersion Models

The purpose of a model that simulates solute transport in groundwater is to compute the concentration of a dissolved chemical species in an aquifer at any specified time and place. The theoretical basis for the equation describing solute transport has been well documented in the literature

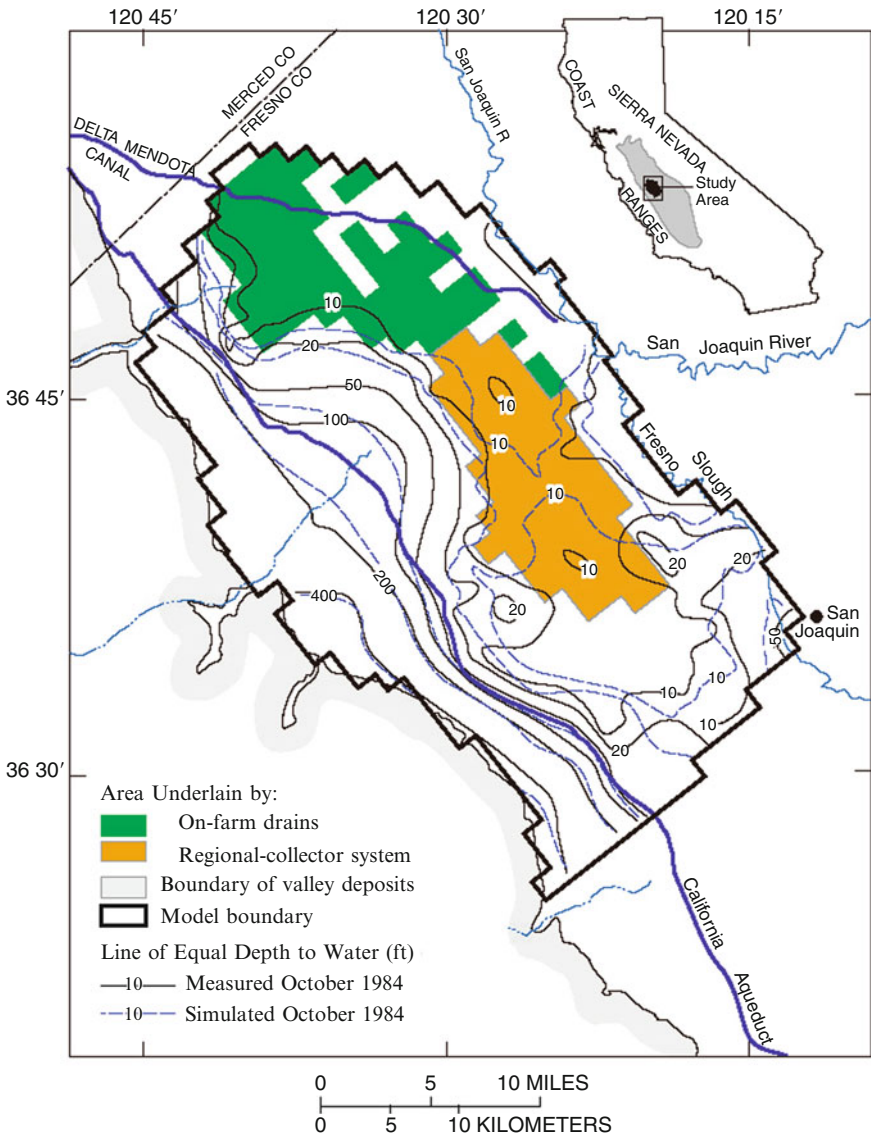


Fig. 33.4 Measured and simulated depths to water in the central part of the western San Joaquin Valley, California, October 1984 (Modified from Belitz et al. 1993)

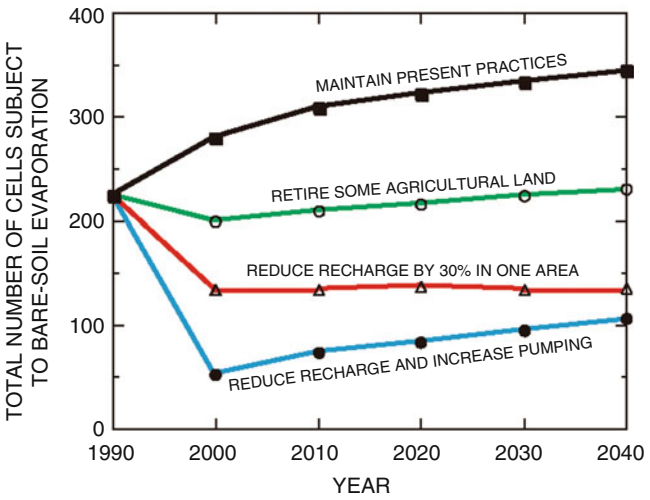


Fig. 33.5 Predicted effects of selected water management alternatives on area of high water table (Modified from Belitz and Phillips 1992)

(e.g., Bear 1979). Zheng and Bennett (2002) provide a conceptual framework for analyzing and modeling solute-transport processes in groundwater, as well as guidelines and examples for applications to field problems.

The mathematical solute-transport model requires at least two partial differential equations. One is the equation of flow, from which groundwater flow velocities are obtained, and the second is the solute-transport equation, whose solution gives the chemical concentration in groundwater. If the properties of the water are affected significantly by changes in solute concentration, as in a saltwater-intrusion problem, then the flow and transport equations should be solved simultaneously (or at least iteratively). If the properties of the water remain constant, then the flow and transport equations can be decoupled and solved sequentially, which is simpler numerically.

The solute-transport equation is more difficult to solve numerically than the groundwater flow equation, largely

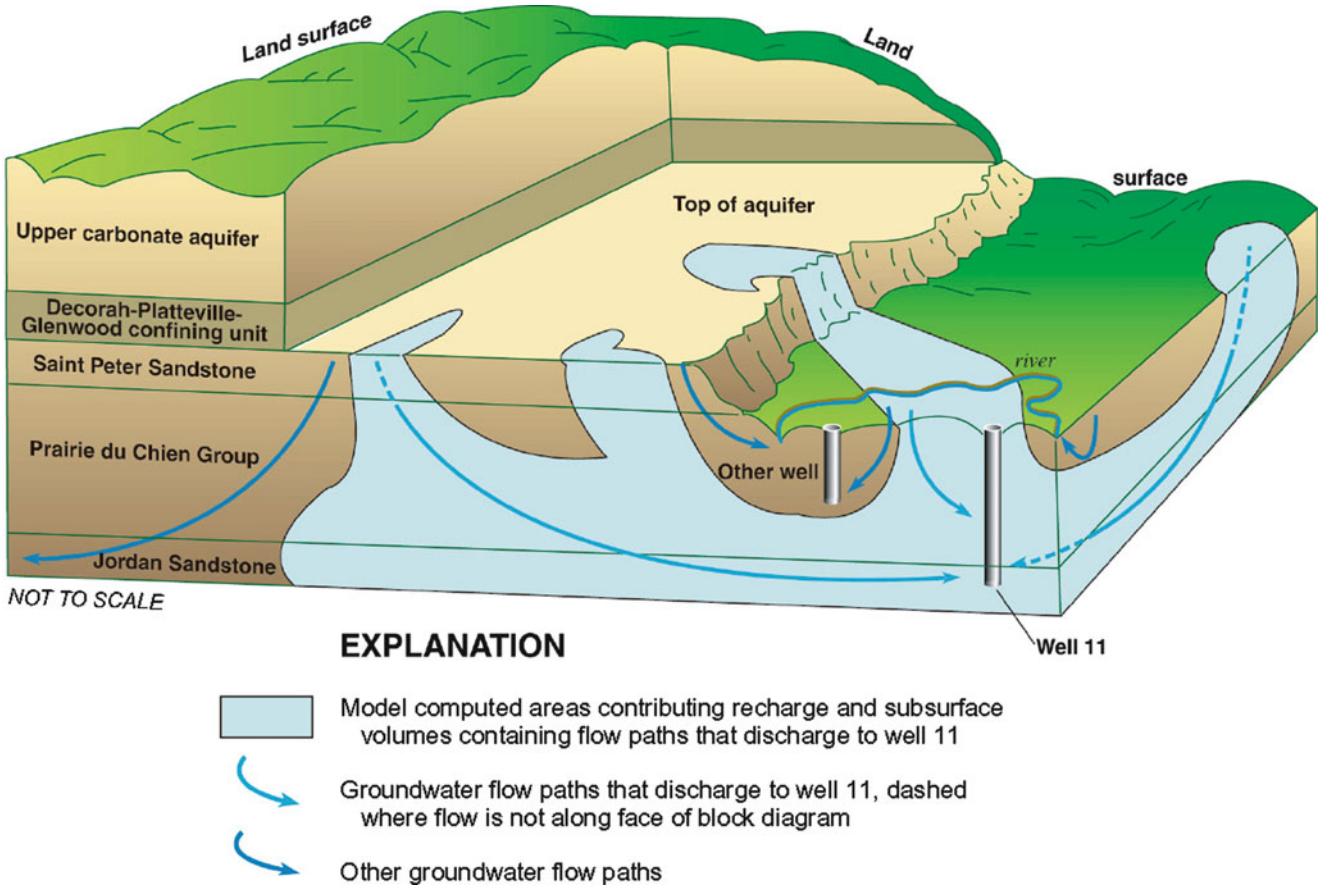


Fig. 33.6 Long-term (steady-state) contributing recharge areas for well 11 near Rochester, Minnesota, calculated using pathline model. Complex three-dimensional patterns of groundwater flow yield irregularly shaped volumes in the subsurface containing the many flow paths that originate at the water table and discharge at well 11 (From Franke et al. 1998)

because the mathematical properties of the transport equation vary depending upon which terms in the equation are dominant in a particular situation. When solute transport is dominated by advection, as is common in many field problems, then the governing equation approximates a hyperbolic type of equation (similar to equations describing the propagation of a wave or of a shock front). But if transport is dominated by dispersive fluxes, such as might occur where fluid velocities are low and (or) hydrodynamic dispersion is relatively high, then the governing equation becomes more parabolic in nature (similar to the transient groundwater flow equation).

The numerical methods that work best for parabolic partial differential equations are not best for solving hyperbolic equations, and *vice versa*. Thus, no one numerical method or simulation model will be ideal for the entire spectrum of groundwater transport problems likely to be encountered in the field. Further compounding this difficulty is that the seepage velocity of groundwater in the field is highly variable, even if aquifer properties are homogeneous, because of the effects of complex boundary conditions. Thus, in low permeability zones or near stagnation points, the velocity may be close to zero and the transport processes will be dominated by dispersion processes; in high permeability zones or near stress points (such as pumping wells), the velocity may be up to several meters per day and the transport processes will be advection dominated. In other words, the governing equation may be more hyperbolic in one area (or at one time) and more parabolic in another area (or at another time). Therefore, regardless of which numerical method is chosen as the basis for a simulation model, it will not be accurate or optimal over the entire domain of the problem. The transport modeling effort must recognize this inherent difficulty and strive to minimize and control the numerical errors. Konikow (2011) discusses some of the conceptual and mathematical difficulties in developing accurate solute-transport models of complex field problems.

Additional complications arise when the solutes of interest are reactive. Simple reaction terms do not necessarily

represent the true complexities of many reactions (see, for example, Glynn 2003). Also, particularly difficult numerical problems arise when reaction terms are highly nonlinear, or if the concentration of the solute of interest is strongly dependent on the concentration of other chemical constituents. For field problems where reactions appreciably affect solute concentrations, simulation accuracy is less limited by mathematical constraints than by data constraints. That is, the types and rates of reactions for the specific solutes and minerals in the particular groundwater system of interest rarely are known and require an extensive amount of data to assess accurately.

Finite-difference and finite-element methods also can be applied to solve the transport equation, particularly when dispersive transport is large compared to advective transport. However, numerical errors, such as numerical dispersion and oscillations, may be large for some problems. The numerical errors generally can be reduced by using a finer discretization (either shorter time steps or finer spatial grid), but this discretization will increase the computational work load. An example of a documented three-dimensional, transient, finite-difference model that simultaneously solves the fluid pressure, energy-transport, and solute-transport equations for nonhomogeneous miscible fluids is HST3D (Kipp 1997). An example of a finite-element transport model is SUTRA (Voss and Provost 2002).

Although finite-difference and finite-element models commonly are applied to transport problems, none of the standard numerical methods is ideal for a wide range of transport problems and conditions. Thus, there currently is much research on developing better mixed or adaptive methods that aim to minimize numerical errors and combine the best features of alternative standard numerical approaches. Examples of other types of numerical methods that also have been applied to transport problems include method of characteristics, random walk, Eulerian-Lagrangian methods, and adaptive grid methods. All these numerical methods have the capability to track sharp fronts accurately with a minimum of numerical dispersion. Documented models based on variants of these approaches include Konikow et al. (1996) and Zheng and Wang (1999).

As an example, the public domain MOC3D model (Konikow et al. 1996) is integrated fully with the MODFLOW-2000 code. The model computes changes in concentration over time caused by the processes of advective transport, hydrodynamic dispersion, mixing or dilution from fluid sources, matrix diffusion, a first-order irreversible-rate reaction, such as radioactive decay, and reversible equilibrium-controlled sorption with a linear isotherm. The model couples the groundwater flow equation with the solute-transport equation. MOC3D uses the method of characteristics to solve the solute-transport equation, by which a particle-tracking procedure represents advective transport and a finite-difference procedure calculates concentration changes caused by hydrodynamic dispersion.

There are many examples in the literature illustrating the application of solute-transport models to problems involving contaminant plumes emanating from point sources of contamination, but few dealing with natural sources. Some of the few involve the application of MODFLOW and MODPATH or MT3D to the arsenic problem in Bangladesh, wherein the sustainability of groundwater development is evaluated on the basis of constraining increases in arsenic concentrations in supply wells calculated by the coupled models (Cuthbert et al. 2002; Michael and Voss 2008; Burgess et al. 2010; Radloff et al. 2011).

33.4.6 Aqueous Speciation Modeling

Geochemical speciation models, such as WATEQF (Plummer et al. 1976) and WATEQ4F (Ball and Nordstrom 1991), calculate the distribution of chemical elements among different aqueous species (bare ions, complexes, and ion pairs) at a given temperature and pressure, determine whether the aqueous solution is supersaturated or undersaturated with respect to various solid mineral phases, and calculate the partial pressure of gases that would be in equilibrium with the calculated solution composition. Speciation models also calculate the total dissolved inorganic concentration (TDIC) of a solution given its measured alkalinity, or conversely calculate its alkalinity given a measured TDIC concentration.

Speciation codes solve a set of algebraic equations that are basically of two types: mass-balance equations and mass-action equations. Mass-balance equations relate the total dissolved concentration (a user-provided measured quantity) of given elements or components to the sum of the concentrations of their aqueous species multiplied by the stoichiometric coefficient of the element/component in each species. Mass-action equations provide thermodynamic relations describing the dependence of the activity (i.e. the thermodynamically effective concentration) of a given aqueous species on the activities of other aqueous species, pH, and redox potential.

To solve the equations described above, speciation codes require that the user provide a complete chemical analysis of the water, including not only the total dissolved concentrations of major and minor elements, but also the pH of the solution and some indication of its redox potential. The redox potential either can be indicated by an Eh or pe value, or alternatively by one or more redox couples. If only one redox couple (e.g. the Fe(II)/Fe(III) couple) is entered, it typically will be used to define the redox potential for all redox-active elements in the solution. More advanced codes, such as PHREEQC (Parkhurst 1995, and Parkhurst and

Appelo 1999), allow the specification of more than one redox couple and allow the user to apply each couple to control the redox distribution of specific redox-active elements.

In addition to the solution-specific data that must be entered by the user, speciation codes also require a thermodynamic database that provides equilibrium constants, as a function of temperature and pressure, for the various aqueous-speciation and complexation reactions considered, and for potential mineral and gas dissolution and precipitation/exsolution reactions. The quality of a speciation code's output will, in large part, be determined both by the quality of the user-entered data and by the quality of the thermodynamic database associated with the code. Ideally, thermodynamic databases should be internally consistent, should consider all major aqueous species, and should be based on accurate measurements. Thermodynamic consistency has many meanings (Nordstrom and Munoz 1994), for example: (1) the data are consistent with basic thermodynamic relations, (2) common scales are used for temperature, energy, atomic mass, and fundamental physical constants, (3) the same mathematical and chemical models were used to fit different data sets, (4) conflicts among measurements were resolved, and (5) appropriate choices of standard states were made and used for all similar substances. In practice, there is considerable uncertainty in thermodynamic data and judging the extent of the uncertainty for different elements, conditions, and calculated results requires geochemical expertise and experience in using the database.

Thermodynamic databases typically consider few organic species, even though organic species are important constituents in both natural and contaminated waters. Most codes and associated thermodynamic databases also are limited to modeling the speciation of relatively dilute waters, with ionic strengths (or salinity) lower than seawater. The few codes that are available to model the speciation of saline waters and brines usually have little or no data available to model the speciation of minor elements, metals, or radionuclides or redox states. Finally, most speciation codes assume that the aqueous species present are at equilibrium with each other. Although most "homogeneous" aqueous-speciation reactions are fast, this is not always the case for reactions involving redox-active species and elements, and/or strong aqueous complexes and polymerized species. The kinetics of formation/dissociation of those species can be slow and the kinetics of redox reactions often depend on microbial catalysis.

Speciation models help in understanding the speciation of an aqueous solution and its thermodynamic state, particularly with respect to the potential dissolution/precipitation/exsolution of various minerals and gases, and to the potential for the exchange or sorption of ions and aqueous species on mineral surfaces. Also, the results of speciation codes can provide valuable insight into the potential toxicity of a natural or contaminated water. For example, dissolved Cr species are more toxic in the +6 oxidation state than in the +3 oxidation state, and strongly complexed cyanide species such as ferro- and ferri-cyanides also are less harmful than CN^- and HCN^0 species. The speciation of a water reveals immediately the predominant forms of potentially toxic elements in a water (assuming that proper thermodynamic data are available) in addition to total concentrations. Finally, aqueous speciation codes often form the core of other geochemical modeling codes, such as "inverse" geochemical modeling codes, sometimes also confusingly referred to as mass-balance models, and also "forward" modeling codes (mass-transfer codes and mass-transport codes).

33.4.7 Inverse Geochemical Modeling

Inverse geochemical modeling uses available chemical and isotopic analyses, which are assumed to be representative of the chemical and isotopic evolution of groundwater along a given flow path, and attempts both to identify and quantify the heterogeneous reactions that may have been responsible for that chemical and isotopic evolution. A speciation code typically is run, as part of the inverse modeling process, to help the user determine the set of reactions that thermodynamically are feasible, to convert alkalinity measurements into TDIC concentrations, and to calculate the redox state (not the redox potential) of the waters considered. Establishing the redox state of the waters is a convention-based process and simply allows the user to ensure that an electron mass balance is maintained, and that no free electrons are created or destroyed as a result of the reactions considered. Nevertheless, apart from the above considerations, the inverse modeling approach does not require that reactions proceed to thermodynamic equilibrium and indeed some of the early inverse modeling codes did not contain a speciation code or a thermodynamic database.

Inverse modeling codes essentially solve a set of algebraic mass-balance equations describing the changes in chemistry and isotopic composition between two waters (or more in the case of "mixtures") and relate those changes to lists of potential reaction sets and reaction amounts. "Initial" waters represent source waters prior to mixing and reactions considered by the model. "Final" water represents the calculated outcome of the mixing and reaction processes. Typically, the user specifies a list of plausible reactions (sometimes called "phases") and also provides a list of components (chemical or isotopic) that will be used to set up and solve the set of mass-balance equations. The inverse modeling code calculates one or more possible "models" (i.e. reaction sets and amounts) that obey the specified mass

balances. Glynn and Brown (1996, 2012) provide a detailed description of inverse geochemical modeling, its requirements and limitations, and the relative capabilities of the two most commonly used codes, NETPATH (Plummer et al. 1994) and PHREEQC (Parkhurst and Appelo 1999).

The PHREEQC code has an advantage over NETPATH in that it accounts for the uncertainties in the analyses provided and, therefore, avoids consideration of reactions with small mass transfers that instead could be explained by uncertainties in the basic data. To do this analysis, PHREEQC assumes that the charge balance error on each given aqueous solution is caused by errors in the analytical data provided, and attempts to "adjust" the analytical data to correct for the charge balance error without exceeding uncertainty limits provided by the user for each analytical datum (total concentrations of each element, pH, isotopic data). PHREEQC also does a more complete accounting of redox balances than NETPATH, allowing redox balances to be maintained not only among the overall redox states of the different waters and reactions, but also among specified redox states for individual elements. Finally, PHREEQC also solves a water-balance equation, an alkalinity balance equation, and a mass balance on inorganic carbon.

NETPATH and its derivative codes NetpathXL (Parkhurst and Charlton 2008) and NETPATH-WIN (El-Kadi et al. 2011) have some capabilities that are not matched by the inverse modeling capabilities of PHREEQC. Foremost, NETPATH incorporates ^{14}C dating capabilities using various literature-based models, or alternatively and preferably, using its own reaction-based models. Additionally, NETPATH incorporates isotopic fractionation factors to calculate the ^{13}C, ^{34}S, and ^{15}N compositions of the "final" water. In its isotopic calculations, NETPATH also solves differential equations, which account for the progressive isotopic evolution of a water as various phases dissolve into it and various phases precipitate or exsolve from it with differing and evolving isotopic compositions. In contrast to NETPATH, PHREEQC inverse modeling only considers isotopic mass-balance constraints, as posed by the user. The PHREEQC user is required to provide the measured isotopic compositions and their uncertainties for the initial and final waters, for dissolving phases, and for precipitating/exsolving phases.

Inverse geochemical modeling is used to **explain** and help understand the observed chemical and isotopic evolution of natural (or contaminated) waters, rather than to **predict** future compositions (as is done by "forward" geochemical modeling). A minimum amount of data is required to use an inverse geochemical modeling code, namely the compositions of at least one "initial" water and a "final" water. Inverse geochemical modeling is best used early in the data-acquisition process because it forces the user to think and evaluate the nature and extent of knowledge gaps and uncertainties. Therefore, inverse geochemical modeling can be used to guide the field-data acquisition process. As should be the case for most hydrological and geochemical modeling, inverse geochemical modeling should be used as part of a continuous iterative cycle between data acquisition and data interpretation and modeling, until some desired level of detail is obtained in understanding the system investigated.

Although inverse geochemical codes at a minimum require two chemical analyses, one for each water, the modeling process requires appreciable knowledge and expertise. The user has to postulate a list of possible reactions that may be responsible for the observed evolution, and, therefore, needs to have a mineralogical knowledge of the system to be able to make reasonable guesses as to what minerals and gases might be dissolving, precipitating, or exsolving. The user also needs to consult the speciation results to determine which reactions thermodynamically are feasible. For example, if both the initial and final waters are undersaturated with respect to a given mineral, it is unlikely that a reaction model that requires precipitation of the mineral would be valid.

The user needs to have some understanding of the relative kinetics of various reaction processes, so as to be able to judge whether a given reaction process is likely to occur to the extent calculated for a given reaction model, given the estimated travel and evolution time of the water. Establishing a plausible hydrologic relation between the initial and the final waters, and estimating a likely travel time between sampling points, requires hydrological knowledge of the system and may involve application of a groundwater flow model. Conversely, the inverse geochemical modeling process may result in an improved, or sometimes radically modified, hydrologic understanding of the groundwater system. For example, if all available models predict that a chloride containing phase needs to precipitate, a thermodynamically unrealistic conclusion in most cases, it is likely instead that either (1) the initial and the final waters are not hydrologically related, or (2) the inverse geochemical modeling process perhaps should consider the diluting effect of an additional initial water so as to explain the lower chloride concentration of the final water.

Inverse geochemical models can account for the possibility of having more than one initial water being responsible for the evolution to a final water composition. Inverse geochemical codes do not consider the various possible mechanisms responsible for the "mixing" of the various initial waters: hydrodynamic dispersion, solute diffusion, mixing of various waters as a result of the sampling process (long screens, temporal variations in water chemistry), and other possibilities. It is the responsibility of the model user to assess the hydrological situation and consider the likelihood of the various processes that might cause this "mixing."

A primary value of inverse geochemical modeling is to force the model user to put all available hydrological, chemical, isotopic, and mineralogical data within a conceptual framework. This action should (1) result in an improved understanding of the chemical and isotopic reactions that may be responsible for the observed evolution of the waters, (2) help refine and improve the user's hydrological understanding of the system, and, most importantly, (3) help assess the nature of some of the remaining uncertainties in the constructed conceptual framework.

33.4.8 Forward Geochemical Modeling: Overview

Forward geochemical modeling differs conceptually from inverse geochemical modeling. Inverse modeling uses available aqueous-solution data and calculates the mass-transfer amounts of various reactions suspected of accounting for the evolution of an initial water to a final water. Inverse modeling is most useful when abundant chemical, isotopic, mineralogical, and hydrologic data are available, and when the user's objective is to explain the past chemical evolution of a groundwater system.

In contrast, forward modeling attempts to *predict* the future chemical composition of an aqueous solution given an initial solution and given certain postulated reactions, some of which usually are considered to go to thermodynamic equilibrium. Forward modeling is most useful when the amount of chemical and isotopic data available for a given groundwater system is limited and when the modeler's objective is to predict the future evolution of the system.

33.4.9 Forward Modeling: Mass-Transfer Codes

Mass-transfer geochemical codes are used to predict the possible evolution of a water as it contacts, forms, and (or) reacts with other phases such as minerals, gases, surface phases, organic matter, and non-aqueous-phase-liquids (NAPLs). Most currently available geochemical codes consider only interactions with minerals, gases, and surfaces. A mass-transfer code essentially is an extension of a speciation code. The main difference is that a mass-transfer code uses thermodynamic (and sometimes kinetic) information to calculate not only the speciation of the aqueous solution (i.e. the aqueous-phase reactions), but also to calculate the effect of heterogeneous reactions (reactions between the aqueous phase and other phases) on the composition and speciation of the aqueous phase and on the composition of contacting phases. Many possible reactions and processes can be simulated, including mineral dissolution and precipitation, gas dissolution and exsolution, gas bubble formation, ion exchange on fixed-charge surfaces, ion sorption on variable charge surfaces, evaporation, dilution and mixing of aqueous solutions, precipitation and dissolution of solid-solution phases, boiling, temperature and pressure changes, radioactive decay, and biodegradation reactions. Most commonly, the user makes the assumption that the processes go to full (or partial) thermodynamic equilibrium, but the most recent codes, such as PHREEQC (Parkhurst and Appelo 1999) and EQ3/6 (Wolery 1992) also can consider reaction kinetics (given appropriate rate law and kinetic constant information from the user) and can calculate changes in composition as a function of time.

Similar to a speciation code, a mass-transfer numerical model solves a set of algebraic mass-balance and mass-action equations. The mass-balance equations impose conservation of mass for the various components of the system across all phases. The mass-action equations provide for specification of thermodynamic equilibrium for both homogeneous (aqueous-only) and heterogeneous (mass-transfer) reactions. If reaction kinetics are simulated, a set of one or more ordinary differential equations also is solved. Mass-transfer codes have all the limitations of speciation codes (uncertainties, errors, and gaps in thermodynamic and analytical data). In addition, numerical convergence problems tend to occur more frequently in mass-transfer codes than in speciation codes. These problems usually are caused by the extreme changes in the concentrations of individual species that can result from even minor heterogeneous-reaction-driven changes in the pH or pe conditions.

Mass-transfer geochemical codes are useful tools in understanding and predicting the effects of reaction processes in groundwater systems. They can be used to predict the minimum and maximum concentrations that may be expected, as a function of varying physico-chemical conditions, for various chemical elements and constituents that may be either toxic or essential to human health. The accuracy of the predictions will be much greater for major constituents (Ca, Na, Mg, K, Cl, SO_4, C, SiO_2) than for minor and trace elements, which often are of concern in water-quality studies. Multiple competing processes often control the concentrations of minor and trace elements; these elements usually are associated and heterogeneously distributed across many different mineral phases and surfaces. Considerable uncertainties and gaps exist in the available thermodynamic data for minor and trace-element processes. Finally, many of the processes are kinetically controlled and are not adequately described by the assumption of thermodynamic equilibrium.

Despite the above uncertainties, geochemical mass-transfer codes have the potential to improve the understanding of minor and trace-element geochemistry. These codes are essential in determining and predicting the effects of the major reaction processes that are responsible for the

evolution of pH, pe, and major element and complexant concentrations. Understanding and predicting the dominant chemical characteristics of the groundwater system is key to understanding and predicting the effects of reaction processes that control the concentrations minor and trace elements.

Inverse and forward geochemical modeling codes can be complementary, as illustrated by an example describing a fluoride water-quality problem. Groundwater from the Black Creek aquifer in the coastal region of South Carolina has elevated concentrations of fluoride. The general geochemistry of the groundwater, the occurrence and causes of elevated fluoride concentrations, and the public-health aspects of the fluoride problem have been discussed by Zack (1980) and Zack and Roberts (1988). Fluoride concentrations in groundwater in this region generally range from 0.5 mg/L in shallow upgradient (younger) waters to 5.5 mg/L in downgradient (older) waters (compared to a recommended limit of 2.0 mg/L). Dentists in the area have noted a high occurrence of dental fluorosis (mottling of dental enamel) among people who have lived since childhood in the area. The problem seems to be mainly cosmetic; epidemiological studies have indicated no significant long-term health risk for fluoride concentrations of 10 mg/L or less.

Zack (1980) and Zack and Roberts (1988) gave a thorough and reasonable explanation of the sources and factors affecting fluoride concentrations and the general geochemistry of the groundwater in the region. Our numerical modeling analysis supports their conclusions and provides further insight, which may allow a better understanding of fluoride geochemistry in groundwaters of the Atlantic Coastal Plain and elsewhere.

Geochemical speciation of a typical high fluoride water (Table 33.1; Well Geo-117) indicates that the water is undersaturated with respect to fluorite (CaF_2), a mineral with fast reaction kinetics. If fluorite were the source of the dissolved fluoride, speciation of the high-fluoride groundwater likely would be close to thermodynamic equilibrium with respect to fluorite. Although some phosphate nodules found in the Black Creek aquifer also contain fluoride, tests have shown that water drawn from phosphatic deposits in eastern North Carolina contain relatively little fluoride (0.4 mg/L), and, consequently, are not thought to be a source of fluoride in the Black Creek aquifer. Instead, Zack (1980) and Zack and Roberts (1988) suggest that fossil shark teeth found in the most hydraulically conductive layers of the aquifer are the source of fluoride. Shark teeth consist of almost pure cryptocrystalline fluoroapatite.

The Black Creek Formation consists of fine to very fine sands interbedded with laminated clays. The shark teeth were deposited during the Upper Cretaceous in a marine environment and now are abundant in thin, relatively continuous layers of calcite-cemented quartz sands. These layers are present in the most transmissive upper third of the aquifer. Zack (1980) notes that there is a strong linear correlation between dissolved fluoride concentration and alkalinity (mainly dissolved bicarbonate).

Table 33.1 Water chemical compositions used in inverse and forward geochemical modeling of ground waters from the Black Creek aquifer in South Carolina

	Water source		Seawater
Property	GEO-113	GEO-117	
Temp[a]	20	20	25
pH[b]	8.5	8.5	8.22
HCO_3	390	626	142
Ca	2	3.4	410
Mg	1.8	1.4	1,350
Na	170	320	10,500
K	7	0.9	390
Fe	0.02	0.3	0.003
Cl	51	83	19,000
SO_4	9.2	4.2	2,700
F	0.5	4.6	1.3
SiO_2	13	19	6.4
PO_4	0.09	0.15	0.28

Units are in mg/L, unless otherwise indicated. Seawater composition from Hem (1992)
[a]Temperature in degrees Celsius
[b]pH in standard units

Zack suggests that the geochemical evolution of the groundwater is controlled primarily by the reaction of dilute low-pH CO_2-rich recharge waters that, while flowing, dissolve calcite cement and exchange Ca for Na present in the initially Na-rich marine clays. The cation exchange reaction causes more dissolution of calcite than otherwise would occur. In turn, dissolution of calcite cement exposes surfaces of shark teeth. The high pH environment found at the calcite-dissolution interface causes substitutional exchange of hydroxide for fluoride in the apatite of the shark teeth, thereby resulting in high dissolved fluoride concentrations. Zack argues convincingly that substitutional exchange rather than fluoroapatite dissolution is the cause of the high fluoride concentrations.

Inverse geochemical simulations were used to identify and quantify reactions that could explain the geochemical evolution of a recharge water into a more saline high fluoride water. The simulations were conducted first with the NETPATH code, and, subsequently, with the PHREEQC code. PHREEQC has the advantage that it accounts for data uncertainties and keeps a mass balance on H, O, charge, and alkalinity. NETPATH, which gave similar results to PHREEQC, is easier to use in the initial exploration of possible reaction models. The simulation results (Table 33.2) largely confirm Zack's conceptual model, although it was found that proton exchange reactions occurring on disseminated organic materials could offer an additional

Table 33.2 Three inverse models determined with PHREEQC for the evolution of a low-fluoride water (Well Geo-113) to a high-fluoride water (Well Geo-117)

Geochemical model			
	CO_2 DISS	H^+ RELEASE	H^+ UPTAKE
CO_2	1.904	–	4.134
Calcite	2.23	4.135	–
HX	–	3.808	−4.461
CaX_2	−2.214	−4.118	0.0168
KX	−0.173	−0.173	−0.173
MgX_2	−0.110	−0.110	−0.110
NaX	4.819	4.819	4.819
CH_2O	0.223	0.223	0.223
Goethite	0.104	0.104	0.104
FeS(ppt)	−0.099	−0.099	−0.099
SiO_2	0.100	0.100	0.100
Fluoroapatite	0.216	0.216	0.216
Hydroxyapatite	−0.216	−0.216	−0.216

Minor mixing of seawater is included (seawater fraction = 0.0016 based on chloride.). Mass transfers are in millimoles per kg of H_2O. Positive numbers indicate mass transfer into the aqueous phase

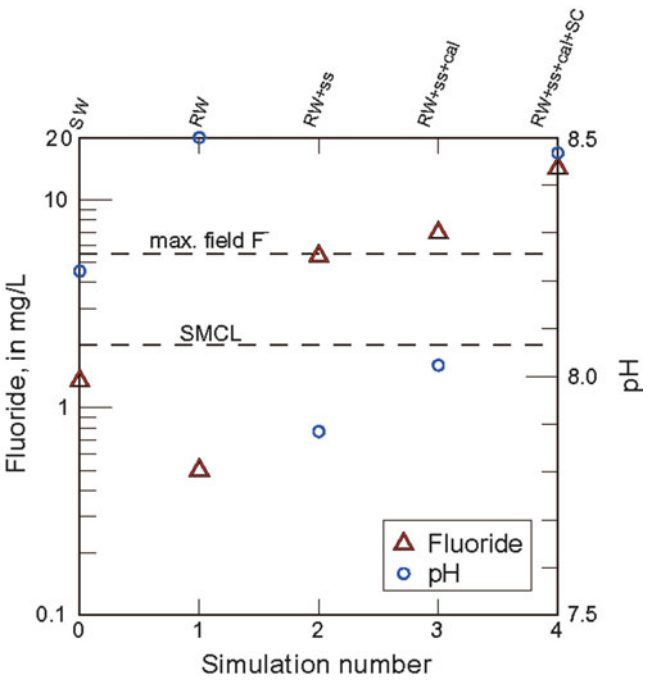


Fig. 33.7 Simulated pH values and fluoride concentrations resulting from reactions between a low-fluoride recharge water (RW) sampled from well Geo-113 in the Black Creek aquifer, and combinations of the following: a fluoroapatite solid-solution (*ss*), calcite (*cal*), and a proton buffering surface (*SC*) modeled using a goethite surface-complexation model. Seawater (*SW*), pH, and fluoride values also are provided for reference, as are fluoride values corresponding to the U.S. Environmental Protection Agency's Secondary Maximum Contaminant Level (SMCL) and the maximum observed in the aquifer (*dashed lines*)

control on the pH values of the groundwater without invoking a source of dissolved CO_2. Among reactions that were not considered in our preliminary modeling, silicate mineral weathering reactions could provide a sink for protons, whereas pyrite and marcasite oxidation could provide a source of protons. Additional information, such as isotopic data, would be required for further determination of the most likely reactions controlling the evolution of the groundwater.

Mass-transfer modeling (Fig. 33.7) with PHREEEQC indicates that thermodynamic consideration of reactions with a fluoroapatite-hydroxyapatite solid-solution series is essential in explaining the fluoride concentrations, pH values, and other geochemical characteristics observed in the groundwater. Reacting the low-fluoride water (well Geo-113) with calcite and a 99.9% fluoroapatite solid solution resulted in a fluoride concentration slightly above the maximum observed in the Black Creek aquifer, although the pH was lower than observed by about half a pH unit. Adding a proton buffering surface (previously equilibrated with seawater) increased the pH to the observed field value of 8.5, but also resulted in an increase in the fluoride concentration to 17 mg/L, three times the maximum observed in the field. Modifications of the number of surface-complexation sites and of the zero-point-of charge of the surface proton buffer could be attempted to obtain a better fit of the field data. Further modeling analyses (incorporating other reactions) should be conducted to provide a better understanding of the factors controlling fluoride concentrations in the Black Creek aquifer.

33.4.10 Forward Modeling: Mass-Transport Codes

Geochemical mass-transport codes are used to simulate (1) the movement of groundwater, (2) the transport of dissolved constituents, and (3) their reactions both within the water phase and with other phases. In addition to solving sets of algebraic mass-balance and mass-action equations, mass-transport codes also solve sets of partial differential equations describing, as a function of space and time, the distribution of groundwater potentials and velocities, and the advective and dispersive movement of solutes.

Geochemical mass-transport codes incorporate all the limitations and uncertainties associated with the use of (1) geochemical mass-transfer codes and (2) nonreactive solute-transport codes. Geochemical mass-transport codes commonly have convergence problems and other numerical problems (e.g. numerical oscillations, numerical dispersion) associated with the numerical solution of partial differential equations. Also, the description and simulation of physico-chemical processes in geochemical transport codes suffers

from a dichotomy of scale. Physical transport processes are described at a much larger scale than the molecular-level-based-scale applicable to chemical reactions. This dichotomy generates conceptual and numerical errors and uncertainties in the application and use of geochemical mass-transport codes.

Additionally, running geochemical transport codes can require large computer time and memory, even on today's computers. Increases in computer processing speeds have been matched by the increasing sophistication and simulation capabilities of geochemical transport codes. Possible increases in the "realism" offered by more sophisticated and complex codes, however, are counter-balanced by increased data requirements and associated increases in the uncertainties relating both to the data entered and to the mathematical representation of the simulated processes. Sensitivity analyses, where simulations are run multiple times to test the effects of the data and process uncertainties, are crucial in any intelligent use of geochemical transport codes, but commonly are hampered by computer time requirements.

Geochemical mass-transport codes can be used to predict "best-case" and "worst-case" scenarios of contaminant transport, but in most cases are not exact predictive tools. Both geochemical mass-transfer and mass-transport codes are useful tools that can be used to improve conceptual understanding and to gain an appreciation of the relative quantitative importance of processes controlling the chemical evolution (and transport) of natural or contaminated waters.

The use of geochemical transport modeling is illustrated by an application to an arsenic problem in Oklahoma. The Central Oklahoma aquifer underlies about 8,000 km^2 of central Oklahoma and is a major source of drinking water for the region. The aquifer is composed mostly of fine-grained sandstones interbedded with siltstone and mudstone. Schlottmann et al. (1998) describe its mineralogy and geochemistry, and they also recognize the occurrence of arsenic as a problem. Recharge to the aquifer occurs mainly in its unconfined eastern area and most streams are gaining (see Fig. 33.8). To the west, the aquifer is confined by low permeability rocks.

Concentrations of dissolved arsenic in the Central Oklahoma aquifer exceed the 1986 Federal drinking water limit of 50 μg/L in about 7% of 477 analyses and even more frequently exceed the more recent standard of 10 μg/L. The highest dissolved arsenic concentrations are found primarily in the western confined part of the aquifer (Fig. 33.8).

Mineralogical and sequential extraction analyses have shown that iron oxides (goethite and hematite) in the sandstones are the primary mineral sources of arsenic. Arsenic sorbs strongly to iron oxides, particularly at pH values below 8. Discrete arsenic mineral phases were not found, although some evidence was found of high arsenic concentrations in pyrite grains. Pyrite only is found in isolated, poorly conductive, low redox zones. Indeed, waters in the Central Oklahoma aquifer generally are oxic, with dissolved oxygen concentrations above 1 mg/L, and there is little organic matter or iron sulfides present. Iron oxide minerals predominate instead.

Extensive geochemical modeling of the Central Oklahoma aquifer has succeeded in elucidating the factors controlling dissolved arsenic concentrations and the general geochemical evolution of the waters (Parkhurst et al. 1993; Parkhurst written communication 2002). Both inverse and forward geochemical modeling was conducted, including a three-dimensional geochemical transport model using the USGS code PHAST, which is based on coupling HST3D with PHREEQC (Parkhurst et al. 2010). Parkhurst's geochemical model assumes that the aquifer initially is filled with sodium chloride brine equilibrated with calcite and dolomite minerals, a cation-exchanger (clays), and a hydrous iron oxide surface with complexed arsenic.

In Parkhurst's model, fresh recharge water, equilibrated with calcite, dolomite, and with carbon dioxide at a partial pressure close to 100 times atmospheric (typical soil CO_2 partial pressure) enters the unconfined part of the aquifer in the east. The recharge water reacts with the initially Na-rich exchanger clays and with the As-rich and proton-depleted hydrous ferric oxide surfaces. As the groundwater flows through the porous media, ion exchange gradually changes the calcium-magnesium bicarbonate recharge water into a sodium bicarbonate water. The initial dissolution of soil carbon dioxide keeps the pH of the recharge water relatively low (between 7.0 and 7.5). After loss of contact with the soil CO_2 reservoir, however, the pH of the recharging water gradually increases from the further dissolution of calcite and dolomite (because of uptake of Ca and Mg on exchange sites) and also from the protonation of the initially proton-depleted hydrous ferric oxide surface. The calculated pH of the resulting sodium bicarbonate water ranges from 8.5 to 9.2, which is close to the observed values. Under these higher pH conditions, sorption sites on the hydrous ferric oxide surface become predominantly negatively charged and, consequently, desorption of arsenic occurs, resulting in higher dissolved As concentrations.

In addition to simulating geochemical reactions, the PHAST code applied by Parkhurst to the Central Oklahoma aquifer also simulates groundwater flow and solute-transport processes. Boundary conditions for the simulation domain (Fig. 33.9) included specified pressures along the eastern part of the northern and southern boundaries to represent the hydrologic effects of adjacent rivers, which are the primary sinks for water discharge from the aquifer. A specified-flux boundary condition also was placed over the unconfined eastern part of the aquifer to simulate recharge.

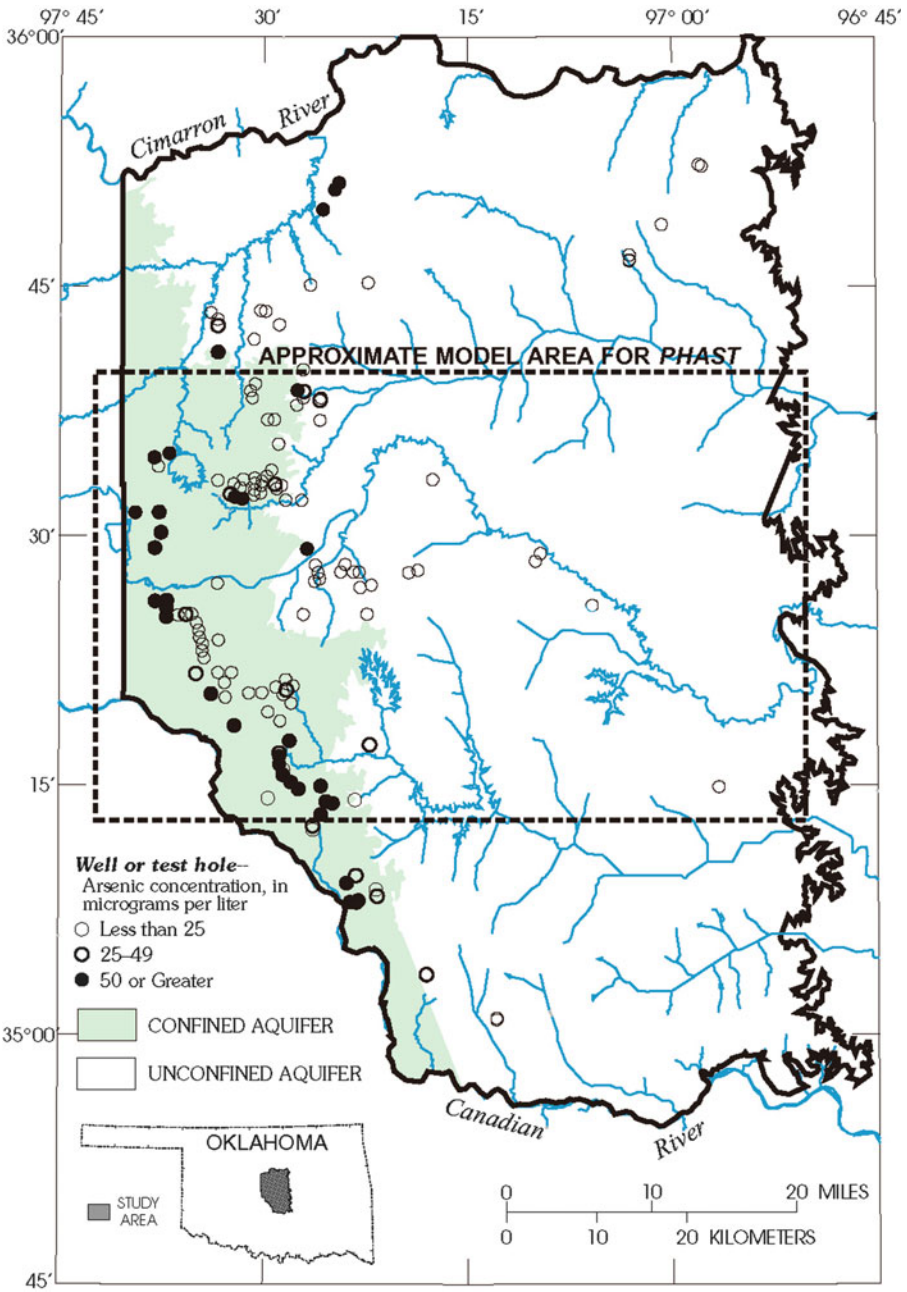


Fig. 33.8 Areal distribution of arsenic in water from deep (>100 m) wells and test holes from the Central Oklahoma aquifer, showing the area where the *PHAST* model was applied (Modified from Schlottmann et al. 1998 and Parkhurst, written communication 2002)

The use of the PHAST model enabled Parkhurst and his co-workers to analyze the magnitude and sensitivity of various factors affecting groundwater flow, solute transport, and geochemical evolution observed in the Central Oklahoma aquifer. Their integrated model was successful in matching general hydrological and geochemical observations and in explaining the occurrence of high arsenic concentrations in the western part of the aquifer.

33.5 Model Design and Application

33.5.1 Overview

The first step in model design and application is to define the nature of the problem and the purpose of the model (Fig. 33.10). This step is linked closely with the formulation

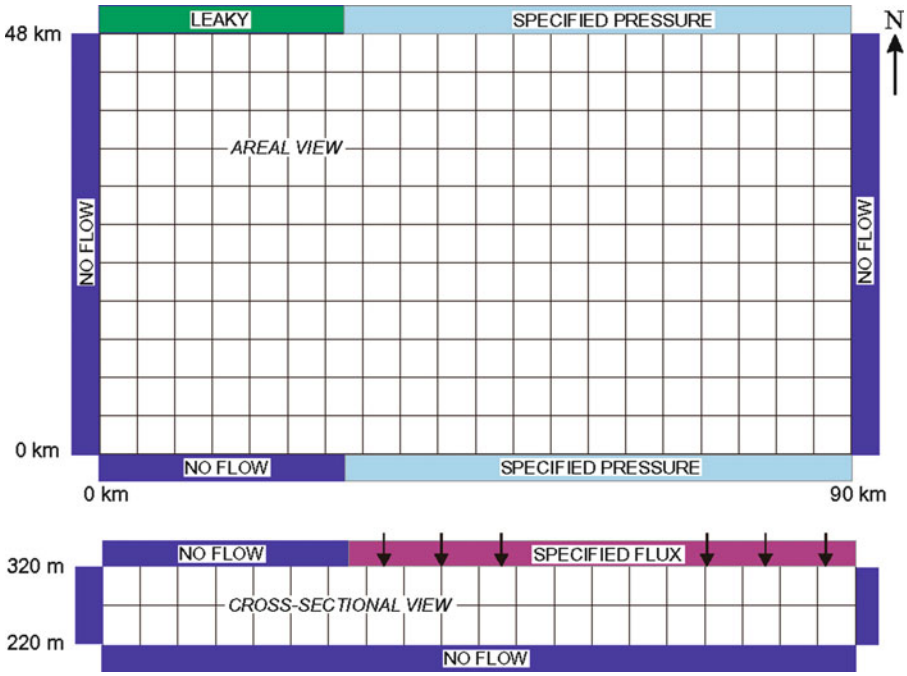


Fig. 33.9 Schematic representation of grid and boundary conditions for application of *PHAST* model to Central Oklahoma aquifer (From Parkhurst, written communication 2002)

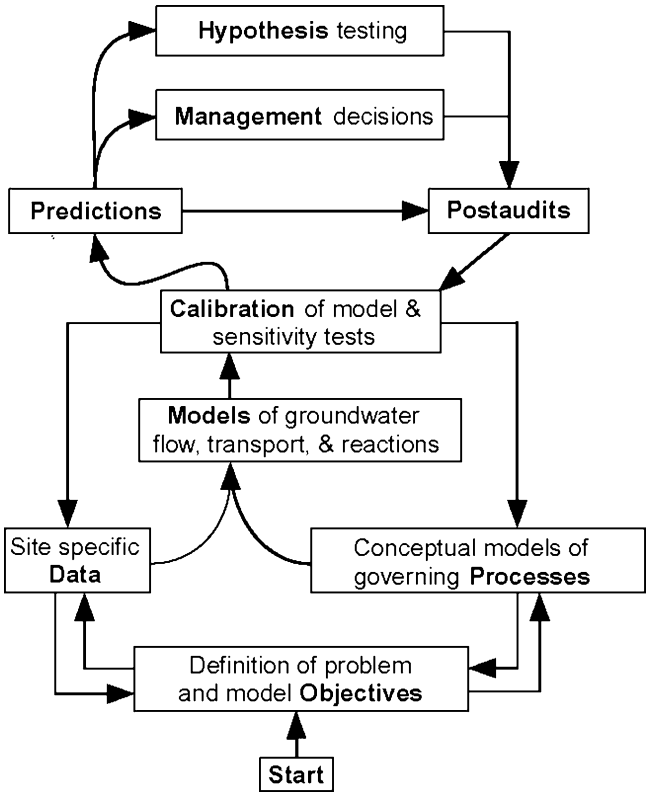


Fig. 33.10 The use of models in the analysis of groundwater systems (Modified from Konikow and Reilly 1998)

of a conceptual model, which is the foundation of model analysis (Bredehoeft 2005). In formulating a conceptual model, one must evaluate which processes are important for the particular problem at hand. Some processes may be important to consider at one scale of study, but negligible or irrelevant at another scale. Good judgment is required to evaluate and balance the tradeoffs between accuracy and cost, with respect to model development, model use, and data requirements. The key to efficiency and accuracy in modeling a system probably is more affected by the formulation of a proper and appropriate conceptual model than by the choice of a particular numerical method or code.

Once a decision to develop a model has been made, a code (or generic model) must be selected (or modified or constructed) that is appropriate for the given problem. Next, the generic code must be adapted to the specific site or region being simulated. Development of a numerical deterministic, distributed-parameter, simulation model involves selecting or designing spatial grids and time increments that will yield an accurate solution for the given system and problem. The analyst must then specify the properties of the system (and their distributions), boundary conditions, initial conditions (for transient problems), and geochemical processes/reactions. All of the parameter specifications and boundary conditions really are part of the overall conceptual model of the system.

Any model is a simplified approximation of a very complex reality, but the model should capture the essential features and processes relative to the problem at hand. The selection of the appropriate model and appropriate level of model complexity remains subjective and dependent on the judgment and experience of the analysts, the objectives of the study, the level of prior information available for the system of interest, and the complexity of the modeled system. The trade-off between model accuracy and model cost always will be difficult to resolve, but always will have to be

made and may affect model reliability. Reilly and Harbaugh (2004) offer general guidelines for evaluating groundwater flow models.

Because the groundwater seepage velocity is determined from the head distribution, and because both advective transport and hydrodynamic dispersion are functions of the seepage velocity, a model of groundwater flow typically is calibrated before a pathline, solute-transport, or geochemical reaction model is developed. In a field environment, perhaps the single most important key to understanding a transport or reaction problem is the development of an accurate definition (or model) of flow. In highly heterogeneous systems, the head distribution and flow directions often can be simulated fairly accurately, whereas the calculated velocity field still may be greatly in error, resulting in considerable errors in simulations of transport.

33.5.2 Grid Design

The dimensionality of a flow or transport model (i.e. one-, two-, or three-dimensions) should be selected during the formulation of the conceptual model. If a one- or two-dimensional model is selected, then it is important that the grid be aligned with the flow system so that there is no unaccounted flux into or out of the line or plane of the grid. For example, if a two-dimensional areal model is applied, then there should be no major vertical components of flow and any vertical leakage or flux must be accounted for by boundary conditions; if a two-dimensional profile model is applied, then the line of the cross section should be aligned with an areal streamline, and there should not be any major lateral flow into or out of the plane of the cross section.

To minimize a variety of sources of numerical errors, the model grid should be designed using the finest mesh spacing and time steps that are possible, given limitations on computer memory and computational time. The boundaries of the grid also should be aligned, to the extent possible, with natural hydrologic and geologic boundaries of the aquifer. Where it is impractical to extend the grid to a natural boundary, then an appropriate boundary condition should be imposed at the grid edge to represent the net effects of the continuation of the aquifer beyond the grid. These boundaries also should be placed as far as possible away from the area of interest and areas of stresses on the system, so as to minimize any effect of conceptual errors associated with these artificial boundary conditions.

In specifying boundary conditions for a particular problem and grid design, care must be taken not to overconstrain the solution. That is, if dependent values are fixed at too many boundary nodes, at either internal or external nodes of a grid, the model may have too little freedom to calculate a meaningful solution (Franke and Reilly 1987).

To optimize computational resources in a model, it sometimes is advisable to use an irregular (or variably-spaced) mesh in which the grid is finest in areas of point stresses, where gradients are steepest, where data are most dense, where the problem is most critical, and (or) where greatest numerical accuracy is desired. Similarly, time steps often can be increased geometrically during a transient simulation.

33.5.3 Model Calibration and Refinement

Model calibration may be viewed as an evolutionary process in which successive adjustments and modifications to the model are based on the results of previous simulations. Overviews on the philosophy of applying and testing groundwater flow and geochemical models are presented by Konikow and Bredehoeft (1992) and Nordstrom (1994). In general, it is best to start with a simple model and add complexity or refine the grid in small increments, as needed and justified.

In applying and evaluating a model, one must decide when sufficient adjustments have been made to the representation of parameters and processes and at some time accept the model as being adequately calibrated (or perhaps reject the model as being inadequate and seek alternative approaches). This decision often is based on a mix of subjective and objective criteria. The achievement of a best fit between values of observed and computed variables is a regression procedure and can be evaluated as such. That is, the residual errors should have a mean that approaches zero and the deviations should be minimized. There are various statistical measures that can be used to assess the reliability and "goodness of fit" of groundwater models. The accuracy tests should be applied to as many dependent variables as possible.

The use of deterministic models in the analysis of groundwater problems is illustrated, in a general sense, in Fig. 33.10. Perhaps the greatest value of the modeling approach is its capability to integrate site-specific data with equations describing the relevant processes as a quantitative basis for predicting changes or responses in a groundwater system. One objective of model calibration should be to improve the conceptual model of the system. The improvement in understanding of a system derived from a model application and calibration exercise for hypothesis testing often is of greater value than the predictive value for management purposes. Another objective should be to define inadequacies in the data base and help set priorities for the collection of additional data.

33.5.4 Model Error

Discrepancies between observed and calculated responses of a groundwater system are the manifestation of errors in the

conceptual or mathematical model. In applying groundwater models to field problems, there are three sources of error, and it may not be possible to distinguish among them (Konikow and Bredehoeft 1992). One source is conceptual errors—that is, misconceptions about the basic processes that are incorporated in the model. Conceptual errors include both neglecting relevant processes as well as inappropriate representation of processes. Examples of conceptual errors include the use of a two-dimensional model where significant flow or transport occurs in the third dimension, or the application of a model based upon Darcy's Law to media or environments where Darcy's Law is inappropriate. A second source of error involves numerical errors arising in the equation-solving algorithm, such as truncation errors, round-off errors, and numerical dispersion. A third source of error arises from measurement errors and from uncertainties and inadequacies in the input data that reflect our inability to describe comprehensively and uniquely the properties, stresses, and boundaries of the groundwater system. In most model applications, conceptualization problems and data uncertainty are the most common sources of error.

In solving advection-dominated transport problems in which a sharp front (or steep concentration gradient) is moving through a groundwater system, it is difficult numerically to preserve the sharpness of the front. Obviously, if the width of the front is narrower than the node spacing, then it is inherently impossible to calculate the correct values of concentration in the vicinity of the sharp front. Even in situations where a front is less sharp, the numerical solution technique can calculate a greater dispersive flux than would occur by physical dispersion alone or would be indicated by an exact solution of the governing equation. That part of the calculated dispersion (or spreading of solute about the center of mass) introduced solely by the numerical solution algorithm is called numerical dispersion. Numerical dispersion can be controlled most easily by reducing the grid spacing, although that will increase computational costs proportionately.

One measure of numerical accuracy is how well the model conserves mass. This can be measured by comparing the net fluxes calculated or specified in the model (e.g. inflow and sources minus outflow and sinks) with changes in storage (accumulation or depletion). Mass-balance calculations always should be performed and checked during the calibration procedure to help assess the numerical accuracy of the solution.

As part of the mass-balance calculations, the hydraulic and chemical fluxes contributed by each distinct hydrologic component of the flow and transport model should be itemized separately to form hydrologic and chemical budgets for the modeled system. The budgets are valuable assessment tools because they provide a measure of the relative importance of each component to the total budget.

33.5.5 Geochemical Model Design

The guidelines for the design and application of geochemical models are similar to those outlined above, especially concerning geochemical transport modeling, which depends on the establishment of flow and solute-transport models. Other types of geochemical models do not depend on the establishment of a spatial grid, and on the attendant issues of grid spacing and boundary conditions. The accuracy of a geochemical mass-transfer model may depend on the time steps if reaction kinetics are being simulated. Other types of geochemical calculations (equilibrium mass-transfer calculations, speciation calculations, inverse geochemical modeling) do not depend on the numerical value of any time-step increments, but can, in some cases, depend on the specification of an initial system state and also on the order in which different isolated geochemical processes or sets of processes are applied/simulated in the system. In all cases of geochemical modeling, the quality of results obtained strongly depend on the quality of the input chemical data and on the quality of any thermodynamic data used by the model.

Inverse geochemical modeling can be used as a first step in helping to construct a geochemical transport model (e.g. Glynn and Brown 1996, 2012). The idea is to use inverse geochemical modeling to determine all the possible sets of reaction processes that potentially could explain the observed chemical and isotopic evolution of one groundwater into another. Consequently, different sets of reaction processes, and different geochemical characteristics, can be considered in a suite of transport simulations, thereby allowing the modeler to assess: (1) the dependence of the movement of a particular contaminant front or concentration on the reaction processes considered, and (2) the need for additional specific field data that potentially could eliminate some of the uncertainties regarding the applicable geochemical processes. As stated by Glynn and Brown (1996), "Identifying knowledge gaps and critical data needs, preventing us from more accurately determining the identity and importance of the reactions …, was one of the most important results of the inverse and reactive transport modeling simulations conducted."

33.6 Obtaining Model Codes

A large number of generic deterministic groundwater models, based on a variety of numerical methods and a variety of conceptual models, are available. In selecting a model that is appropriate for a particular application, it is most important to choose one that incorporates the proper conceptual model; one must avoid force fitting an inappropriate model to a field situation solely because of model

convenience, availability, or familiarity to the user. Usability also is enhanced by the availability of graphical preprocessing and postprocessing programs or features, and by the availability of comprehensive yet understandable model documentation.

A large number of public and private organizations distribute public domain and (or) proprietary software for groundwater modeling. Some Internet sites allow computer codes to be downloaded at no cost while other sites provide catalog information, demonstrations, and pricing information. The Integrated Groundwater Modeling Center, Golden, Colorado (http://igwmc.mines.edu/Software.html) distributes groundwater simulation models. Many of the U.S. Geological Survey public domain codes are available from links on their software web site at http://water.usgs.gov/software/.

See Also the Following Chapters. Chapter 12 (Arsenic in Groundwater and the Environment) • Chapter 13 (Fluoride in Natural Waters) • Chapter 16 (Selenium Deficiency and Toxicity in the Environment)

References

Alley WM, Reilly TE, Franke OL (1999) Sustainability of groundwater resources, U.S. Geological Survey Circular 1186. U.S. Geological Survey, Reston

Anderson MP, Woessner WW (1992) Applied groundwater modeling: simulation of flow and advective transport. Academic, San Diego

Ball JW, Nordstrom DK (1991) User's manual for WATEQ4F, with revised thermodynamic database and test cases for calculating speciation of major, trace and redox elements in natural waters. U.S. Geological Survey, Reston, Open-File Report 91-183

Bear J (1979) Hydraulics of groundwater. McGraw-Hill, New York

Belitz K, Phillips SP (1992) Simulation of water-table response to management alternatives, central part of the western San Joaquin Valley, California. U.S. Geological Survey Water-Resources Investigations Report 91-4193

Belitz K, Phillips SP, Gronberg JM (1993) Numerical simulation of ground-water flow in the central part of the western San Joaquin Valley, California. U.S. Geological Survey Water-Supply Paper 2396

Bredehoeft JD (2005) The conceptualization model problem–surprise. Hydrogeol J 13:37–46

Burgess WG, Hoque MA, Michael HA, Voss CI, Breit GN, Ahmed KM (2010) Vulnerability of deep groundwater in the Bengal aquifer system to contamination by arsenic. Nat Geosci 3:83–87. doi:10.1038/ngeo750

Clement TP (1997) RT3D—A modular computer code for simulating reactive multi-species transport in 3-dimensional groundwater aquifers, Report PNNL-11720, Pacific Northwest National Laboratory, Richland, Washington

Cuthbert MO, Burgess WG, Connell L (2002) Constraints on sustainable development of arsenic-bearing aquifers in southern Bangladesh. Part 2: preliminary models of arsenic variability in pumped groundwater. In: Hiscock KM, Rivett MO, Davison RM (eds) Sustainable groundwater development, Geological Society Special Publication 193. The Geological Society of London, London

Deverel SJ, Gilliom RJ, Fujii R, Izbicki JA, Fields JC (1984) Areal distribution of selenium and other inorganic constituents in shallow ground water of the San Luis Drain service area, San Joaquin Valley, California—a preliminary study. U.S. Geological Survey Water-Resources Investigations Report 84-4319

Doherty J (2010) PEST, model-independent parameter estimation—user manual, 5th edn. Watermark Numerical Computing, Brisbane

Dowdle PR, Laverman AM, Oremland RS (1996) Bacterial dissimilatory reduction of arsenic (V) to arsenic (III) in anoxic sediments. Appl Environ Microbiol 62:1664–1669

Drever JI (1997) The geochemistry of natural waters: surface and groundwater environments, 3rd edn. Prentice Hall, Upper Saddle River

El-Kadi AI, Plummer LN, Aggarwal PK (2011) NETPATH-WIN: an interactive user version of NETPATH, a geochemical and isotopic mass-balance model. Ground Water 49(4):593–599

Essaid HI, Bekins BA (1997) BIOMOC, a multispecies solute-transport model with biodegradation. U.S. Geological Survey, Reston Water-Resources Investigations Report 97-4022

Fang Y, Scheibe TD, Mahadevan R, Garg S, Long PE, Lovley DR (2011) Direct coupling of a genome-scale microbial *in Silico* model and a groundwater reactive transport model. J Contam Hydrol 122:96–103. doi:10.1016/j.jconhyd.2010.11.007

Feder GL, Radovanovic Z, Finkelman RB (1991) Relationship between weathered coal deposits and the etiology of Balkan endemic nephropathy. Kidney Int 40(suppl 34):S9–S11

Franke OL, Reilly TE (1987) The effects of boundary conditions on the steady-state response of three hypothetical ground-water systems—results and implications of numerical experiments, U.S. Geological Survey Water-Supply Paper 2315. U.S. Geological Survey, Reston

Franke OL, Reilly TE, Pollock DW, LaBaugh JW (1998) Estimating areas contributing recharge to wells: lessons from previous studies, U.S. Geological Survey Circular 1174. U.S. Geological Survey, Reston

Gelhar LW, Welty C, Rehfeldt KR (1992) A critical review of data on field-scale dispersion in aquifers. Water Resour Res 28:1955–1974

Glynn PD (2000) Solid-solution solubilities and thermodynamics: sulfates, carbonates and halides. In: Alpers CN, Jambor JL, Nordstrom DK (eds) Sulfate minerals, crystallography, geochemistry and environmental significance. Rev Mineral Geochem, Mineralogical Society of America, Washington, D.C. 40:481–511, chap. 10

Glynn PD (2003) Modeling Np and Pu transport with a surface complexation model and spatially variant sorption capacities: implications for reactive transport modeling and performance assessments of nuclear waste disposal site. Comput Geosci 29:331–349

Glynn PD, Brown JG (1996) Reactive transport modeling of acidic metal-contaminated ground water at a site with sparse spatial information. In: Steefel CI, Lichtner P, Oelkers E (eds) Reactive transport in porous media, vol 34, Reviews in mineralogy. Mineralogical Society of America, Washington, DC, pp 377–438, chap. 9

Glynn PD, Brown JG (2012) Integrating field observations and inverse and forward modeling: application at a site with acidic, heavy-metal-contaminated groundwater. In: Bundschuh J, Zilberbrand M (eds) Geochemical modeling of groundwater, vadose and geothermal systems. CRC Press, Leiden, pp 181–234, chap. 8

Grant WD, Gemmell RT, McGenity TJ (1998) Halophiles. In: Horikoshi K, Grant WD (eds) Extremophiles: microbial life in extreme environments, Wiley series in ecological and applied microbiology. Wiley-Liss, New York

Haitjema HM (1995) Analytic element modeling of groundwater flow. Academic, San Diego

Harbaugh AW (2005) MODFLOW-2005, The U.S. Geological Survey modular ground-water model—the ground-water flow process. U.S. Geological Survey, Reston, Techniques and Methods 6-A16

Harbaugh AW, Banta ER, Hill MC, McDonald MG (2000) MODFLOW-2000, The U.S. Geological Survey modular ground-water model—user guide to modularization concepts and the ground-water flow process. U.S. Geological Survey, Reston, Open-File Report 00-92

Hem JD (1992) Study and interpretation of the chemical characteristics of natural water, 3rd edn, U.S. Geological Survey Water-Supply Paper 2254. U.S. Geological Survey, Reston

Hill MC, Tiedeman CR (2007) Effective groundwater model calibration, with analysis of data, sensitivities, predictions, and uncertainty. Wiley, New York

Hopps HC, Feder GL (1986) Chemical qualities of water that contribute to human health in a postive Way. Sci Total Environ 54:207–216

Huyakorn PS, Pinder GF (1983) Computational methods in subsurface flow. Academic, New York

Ibaraki M, Therrien R (eds) (2001) Practical applications of coupled process models in subsurface environments. J Contam Hydrol 52

Johnson CD, Truex MJ, Clement TP (2006) Natural and enhanced attenuation of chlorinated solvents using RT3D, Report PNNL-15937 Pacific Northwest National Laboratory, Richland, Washington

Keller WD (1978) Drinking water: a geochemical factor in human health. Geol Soc Am Bull 89:334–336

Kenny JF, Barber NL, Hutson SS, Linsey KS, Lovelace JK, Maupin MA (2009) Estimated use of water in the United States in 2005, U.S. Geological Survey Circular 1344. U.S. Geological Survey, Reston

Kipp KL Jr (1997) Guide to the revised heat and solute transport simulator, HST3D—version 2. U.S. Geological Survey, Reston, Water-Resources Investigations Report 97-4157

Konikow LF (2011) The secret to successful solute-transport modeling. Ground Water 49:144–159. doi:10.1111/j.1745-6584.2010.00764.x

Konikow LF, Bredehoeft JD (1992) Ground-water models cannot be validated. Adv Water Resour 15:75–83

Konikow LF, Reilly TE (1998) Groundwater modeling. In: Delleur JW (ed) The handbook of groundwater engineering. CRC Press, Boca Raton

Konikow LF, Goode DJ, Hornberger GZ (1996) A three-dimensional method of characteristics solute-transport model (MOC3D). U.S. Geological Survey, Reston, Water-Resources Investigations Report 96-4267

Lichtner PC, Steefel CI, Oelkers EH (eds) (1996) Reviews in mineralogy, vol 34. Mineralogical Society of America, Washington, DC

Lovley DR, Phillips EJP, Gorby Y, Landa ER (1991) Microbial reduction of uranium. Nature 35:413–416

McDonald MG, Harbaugh AW (1988) A modular three-dimensional finite-difference ground-water flow model. U.S. Geological Survey Techniques of Water-Resources Investigations, chap. A1

Michael HA, Voss CI (2008) Evaluation of the sustainability of deep groundwater as an arsenic-safe resource in the Bengal basin. Proc Natl Acad Sci USA 105:8531–8536

Nordstrom DK (1994) On the evaluation and application of geochemical models. In: Proceedings of the 5th CEC natural analogue working group and alligator rivers analogue project, Report EUR 15176EN, 375–385

Nordstrom DK, Munoz JL (1994) Geochemical thermodynamics, 2nd edn. Commission of the European Community, Brussels

Parkhurst DL (1995) User's guide to PHREEQC—a computer program for speciation, reaction-path, advective-transport, and inverse geochemical calculations. U.S. Geological Survey, Water-Resources Investigations Report 95-4227

Parkhurst DL, Appelo CAJ (1999) User's guide to PHREEQC (version 2)—a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. U.S. Geological Survey, Reston, Water-Resources Investigations Report 99-4259

Parkhurst DL, Charlton SR (2008) NetpathXL-An Excel® interface to the program NETPATH. U.S. Geological Survey Techniques and Methods 6-A26 (Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government)

Parkhurst DL, Christenson S, Breit GN (1993) Ground-water-quality assessment of the central Oklahoma Aquifer, Oklahoma: geochemical and geohydrologic investigations. U.S. Geological Survey, Reston, Open-File Report 92-642

Parkhurst DL, Kipp KL, Charlton SR (2010) PHAST version 2—a program for simulating groundwater flow, solute transport, and multicomponent geochemical reactions. U.S. Geological Survey Techniques and Methods 6–A35

Pedersen K (1993) The deep subterranean biosphere. Earth Sci Rev 34:243–260

Plummer LN, Jones BF, Truesdell AH (1976) WATEQF–A FORTRAN IV version of WATEQ, a computer program for calculating chemical equilibrium of natural waters, U.S. Geological Survey, Reston Water-Resources Investigations Report 76-13 (revised 1984)

Plummer LN, Prestemon EC, Parkhurst DL (1994) An interactive code (NETPATH) for modeling NET geochemical reactions along a flow PATH—version 2.0. U.S. Geological Survey, Reston, Water-Resources Investigations Report 94-4169

Pollock DW (1989) Documentation of computer programs to compute and display pathlines using results from the U.S. Geological Survey modular three-dimensional finite-difference ground-water flow model. U.S. Geological Survey, Reston, Open-File Report 89-381

Presser TS, Barnes I (1985) Dissolved constituents including selenium in waters in the vicinity of Kesterson National Wildlife Refuge and the west grassland, Fresno and Merced Counties, California. U.S. Geological Survey, Reston, Water-Resources Investigations Report 84-4220

Presser TS, Swain WC, Tidball RR, Severson RC (1990) Geologic sources, mobilization, and transport of selenium from the California coast ranges to the Western San Joaquin Valley: a reconnaissance study. U.S. Geological Survey, Reston, Water-Resources Investigations Report 90-4070

Radloff KA, Zheng Y, Michael HA, Stute M, Bostick BC, Mihajlov I, Bounds M, Huq MR, Choudhury I, Rahman MW, Schlosser P, Ahmed KM, van Geen A (2011) Arsenic migration to deep groundwater in Bangladesh influenced by adsorption and water demand. Nat Geosci. doi:10.1038/NGEO1283

Reilly TE, Harbaugh AW (2004) Guidelines for evaluating groundwater flow models. U.S. Geological Survey, Reston, Scientific Investigations Report 2004-5038

Scheibe TD, Mahadevan R, Fang Y, Garg S, Long PE, Lovley DR (2009) Coupling a genome-scale metabolic model with a reactive transport model to describe *in situ* uranium bioremediation. Microb Biotechnol 2(2):274–286. doi:10.1111/j.1751-7915.2009.00087.x

Schlottmann JL, Mosier EL, Breit GN (1998) Arsenic, chromium, selenium, and uranium in the central Oklahoma aquifer. In: Christenson S, Havens JS (eds) Ground-water-quality assessment of the central Oklahoma aquifer, Oklahoma: results of investigations, U.S. Geological Survey Water-Supply Paper 2357A. U.S. Geological Survey, Reston, pp 119–179

Schwarzenbach RP, Gschwend PM, Imboden DM (1993) Environmental organic chemistry. Wiley, New York

Stetter KO (1998) Hyperthermophiles: isolation, classification and properties. In: Horikoshi K, Grant WD (eds) Extremophiles: microbial life in extreme environments, Wiley series in ecological and applied microbiology. Wiley-Liss, New York

Switzer Blum J, Stolz JF, Oren A, Oremland RS (2001) *Selenihalanaerobacter shriftii* gen. nov., sp. nov., a halophilic anaerobe from dead sea sediments that respires selenate. Arch Microbiol 175:208–219

Voss CI, Provost AM (2002) SUTRA, a model for saturated-unsaturated variable-density ground-water flow with solute or energy transport. U.S. Geological Survey, Reston, Water-Resources Investigations Report 02-4231

Wang HF, Anderson MP (1982) Introduction to groundwater modeling. Freeman, San Francisco

Winter TC, Harvey JW, Franke OL, Alley WM (1998) Ground water and surface water: a single resource, U.S. Geological Survey Circular 1139. U.S. Geological Survey, Reston

Wolery TJ (1992) EQ3/6, a software package for geochemical modeling of aqueous systems: package overview and installation guide (Version 7.0), UCRL-MA-110662 pt. 1, Lawrence Livermore National Laboratory, Livermore, California

Zack AL (1980) Geochemistry of fluoride in the Black Creek aquifer system of Horry and Georgetown Counties, South Carolina–and its physiological implications, U.S. Geological Survey Water-Supply Paper 2067. U.S. Geological Survey, Washington, DC

Zack AL, Roberts I (1988) The geochemical evolution of aqueous sodium in the Black Creek aquifer, Horry and Georgetown Counties, South Carolina, U.S. Geological Survey Water-Supply Paper 2324. U.S. Geological Survey, Reston

Zheng C (2006) MT3DMS v5.2 supplemental user's guide: technical report to the U.S. army engineer research and development center. Department of Geological Sciences, University of Alabama, Tuscaloosa

Zheng C, Bennett GD (2002) Applied contaminant transport modeling, 2nd edn. Wiley-Interscience, New York

Zheng C, Wang PP (1999) MT3DMS: a modular three-dimensional multispecies model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems; documentation and user's guide. Report SERDP-99-1. U.S. Army Engineer Research and Development Center, Vicksburg, MS

Appendices

Appendix A: International Reference Values

(Indicator values 2004. The values are continuously changing so these give an indication on the levels)

International Reference Values: Soils

Peter Bobrowsky, Roger Paulen, Pauline Smedley, and Brian J. Alloway

Maximum Permissible Concentrations of Heavy Metals and Metalloids in Soils ($mgkg^{-1}$)

Element	UK (1)	UK (2)	Netherlands		USA	AUS	NZ	Europe
			Targ	*Int*				
Arsenic	20	50	29	55	–	20	10	–
Cadmium	1 (pH 6)	3 (pH >5)	0.8	12	20	1 (3 SA)	3	1–3
	2 (pH 7)							
	8 (pH 8)							
Chromium	130	(400prov)	100	380	1,500	100	600	–
Copper	–	80 (pH 5–5.5)	36	190	750	100 (200 SA)	140	50–140
		100 (pH 5.5–6)						
		135 (pH 6–7)						
		200 (pH >7)						
Mercury	8	1 (pH >5)	0.3	10	8	1	1	1–1.5
Nickel	50	50 (pH 5–5.5)	35	210	210	60	35	30–75
Lead	450	300	85	530	150	150 (200 SA)	300	50–300
Zinc	–	200 (pH 5–7)	720	140	1,400	200 (250 SA)	300	150–300
		300 (pH >7)						

Notes and references: UK (1)—Contaminated Land Exposure Assessment (CLEA) guidance values (to be used as part of a risk assessment for contaminated sites), Department for Environment, Food and Rural Affairs, R & D Publications SGV 1, 3, 4, 5, 7, and 10, Environment Agency, Bristol, 2002. UK (2)—for normal agricultural soils and values for zinc and copper in all types of soil. The Soil Code: Code of Good Agricultural Practice for the Protection of Soil PB0617, MAFF, London, 1998. Netherlands—Dutch Limits *Targ* target values (which it is intended that soil should reach) and *Int* intervention values (when site needs to be cleaned up). These values are for assessing the need for remediation of land suspected of being contaminated and apply to a—standard soil containing 10% organic matter and 25% clay. VROM (2000) circular on target values and intervention values for soil remediation. Ministry of Housing, Spatial Planning and Environment, Department of Soil Protection (VROM) The Hague, The Netherlands. DBO/1999226863. USA—Maximum concentrations for soils treated with biosolids (sewage sludge) McGrath et al. 1994. Land application of sewage sludge: scientific perspective if heavy metal loading limits in Europe and the United States. *Environmental Reviews*, 2, 108–118. Australia and New Zealand—Guidelines for controlling metal concentrations in soils for reuse of biosolids (*SA* values used in the state of South Australia). McLaughlin et al. (2000). Review: A bioavailability-based rationale for controlling metal and metalloid contamination of agricultural land in Australia and New Zealand. *Australian Journal of Soil Research,* 38, 1037–1086. Europe—for countries of the European Union for soils receiving sewage sludge (assumes soil pH 6–7), lower value is guideline value, upper value is the mandatory limit. Commission of the European Communities (1986) Council Directive (86/278/EEC) on the protection of the environment, and in particular of the soil, when sewage sludge is used in agriculture. *Official Journal of the European Communities*, 15, 69–81.

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5,

International Reference Values: Water

Peter Bobrowsky, Roger Paulen, Brian J. Alloway, and Pauline Smedley

Regulations and guidelines: inorganic trace constituents in drinking water, excluding radiological parameters. All Units in μg/L unless other stated

Country or Institution	Nature of standards	Comments	Date	Al	Ag	As	B	Ba	Be	Ca	Cd	Cl	Cr	Cu	DO
Australia	Guidelines	Health-based guidelines	2011		100	10	4,000	2,000	60		2		50vi	2,000	
	Guidelines	Aesthetic guidelines	2011	200								250,000		1,000	85% saturated
Canada	Guidelines	Health-based guidelines	2010			10	5,000	1,000			5		50		
	Guidelines	Aesthetic guidelines	2010	100/ 200a								250,000		1,000	
Japan	Regulations	Drinking water quality standards	2004	200		10	1,000			300,000	10	200,000	50vi	1,000	
	Guidelines	Complementary items for water management	2004							10,000–100,000					
EC (European Commission)	Regulations	Maximum permissible values	1998			10	1,000				5		50	2,000	
	Regulations	Indicator parameters	1998	200								250,000			
US EPA	Regulations	Primary standards (Maximum contaminant levels, MCLs)	2009			10		2,000	4		5		100	1,300~	
	Guidelines	Secondary standards	2009	50–200	100							250,000		1,000	
WHO	Guidelines	Guideline values	2011				2,400	700			3			2,000	
	Guidelines	Provisional guideline values	2011			10							50		

DO dissolved oxygen
* Applies only to water from treatment plants using Al-based coagulants (100 μg/L conventional/200 μg/L other types)
~ action level
# short-term exposure
vi as Cr (VI)
i for infants, 100mg/L for adults and children over 3 months
p provisional

F	Fe	Hardness	Hg	H_2S	I	Mg	Mn	Mo	Na	NH_3	Ni	NO_2	NO_3	pH	Pb	Sb	Se	Si	SO_4	Tl	U	V	Zn
		as $CaCO_3$		as H_2S						as NH_3		as NO_2	as NO_3					as SiO_2	as SO_4				
1,500			1		500		500	50			20	3,000	50000i		10	3	10				17		
	300	200,000		50			100		180,000	500				6.5–8.5				80,000	250,000				3,000
1,500													45,000			6	10				20		
	300			50			50							6.5–8.5					500,000				5,000
800	300		0.5			300,000	50		200,000			32,800	44,300	5.8–8.6	10		10						1,000
						10,000–100,000	10				10p	164p		ca. 7.5		15					2p		
1,500			1								20	500	50,000		10	5	10						
	200						50		200,000	500				6.5–9.5					250,000				
4,000			2									3,300	44,300		15~	6	50			2	30		
2,000	300						50							6.5–8.5					250,000				5,000
1,500			6								70	3,000#	50,000#			20							
															10		40				30		

Appendix B: Web Links and Other Relevant Information

Annotated URLs for Chapter-Related Web Sites of Interest

Chapter 20

Ania BJ, Asenjo M (2002) Mycetoma. eMedicine J http://www.emedicine.com/MED/topic30.htm. Accessed 2002

Cambridge University Schistosomiasis Research Group (2002) Helminth infections of man. University of Cambridge, Cambridge CB2 1TN, UK, http://www.path.cam.ac.uk/. Accessed 2002

Canadian Centre for Occupational Health and Safety (1999) What is farmer's lung. Canadian Centre for Occupational Health and Safety, Hamilton, Ontario, Canada, http://www.ccohs.ca/oshanswers/diseases/farmers_lung.html. Accessed 1999

Carey J, Motyl M, Perlman DC (2001) Catheter-related bacteremia due to *Streptomyces* in a patient receiving holistic infusions. In: Emerging infectious diseases, vol 6, no 6, Nov–Dec, 2001, Centers for Disease Control and Prevention, Atlanta, Georgia, http://www.cdc.gov/ncidod/eid/index.htm

CDC (2002) U.S. Centers for Disease Control and Prevention, Atlanta, Georgia, http://www.cdc.gov/. Accessed 2002

DoctorFungus (2002) DoctorFungus.org, http://www.doctorfungus.org/thefungi/index.php. Accessed 2002

Health Canada (2002) Material safety data sheets, Health Canada Population and Public Health Branch, Ottawa, Canada, http://www.hc-sc.gc.ca/pphb-dgspsp/msds-ftss/index.html#menu. Accessed 2002

Mack DR (2001) *Dientamoeba fragilis* infection, eMedicine World Medical Library, http://www.emedicine.com/

McGinnis MR (1998) Introduction to mycology. In: Baron S (ed) Medical microbiology, http://www.ncbi.nlm.nih.gov/books/NBK8471/

NASA (2001a) Soil science education home page, Goddard Space Flight Center Laboratory for Terrestrial Physics, NASA, Greenbelt, Maryland, http://soil.gsfc.nasa.gov/index.php?section=7

National Cattleman's Beef Association and Cattleman's Beef Board (2001) Transmissible Spongiform Encephalopathies (TSE): BSE Info Resource, http://bseinfo.org/. Accessed 14 Feb 2002

National Institutes of Health (2001) Parasitic roundworm diseases fact sheet, National Institute of Health, Washington, DC, http://www.niaid.nih.gov/factsheets/roundwor.htm

National Institute of Allergy and Infectious Diseases (2000) Antimicrobial resistance: fact sheet, office of communications and public liaison, U.S. Department of Health and Human Services, p 4, http://www.niaid.nih.gov/factsheets/antimicro.htm

NRCS (1999) Soil biology primer, National Resources Conservation Service, U.S. Department of Agriculture, Washington, DC, 53p, http://www.gsfc.nasa.gov/gsfc/earth/toms/microbes.htm

Pitetti RD (2001) Visceral larva migrans, eMedicine World Medical Library, http://www.emedicine.com/

Soil Science Society of America (1998) Internet glossary of soil science terms, Soil Science Society of America, Madison, Wisconsin, https://www.soils.org/publications/soils-glossary

Soil Survey Staff (1998) Keys to soil taxonomy, Natural Resource Conservation Service, U. S. Department of Agriculture, Washington, DC, 40 p, ftp://ftp-fc.sc.egov.usda.gov/NSSC/Soil_Taxonomy/keys/2010_Keys_to_Soil_-Taxonomy.pdf

Standing Medical Advisory Committee (1998) The path of least resistance: report of the sub-group on Antimicrobial Resistance, Department of Health, London, U.K, http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4009357

University of California (2002a) Bacteria: life history and ecology, Copyright 1994–2002 by The University of California Museum of Paleontology, Berkeley, and the Regents of the University of California, http://www.ucmp.berkeley.edu/bacteria/bacterialh.html. Accessed 2002

University of California (2002b) Schistosomiasis in China, University of California, Berkeley, California, the Regents of the University of California, http://ehs.sph.berkeley.edu/china2/

Valley Fever Center for Excellence (2002) What is valley fever?, University of Arizona and Southern Arizona Veteran's Administration Healthcare System, Tucson, Arizona, http://www.valleyfeversurvivor.com/faq.html. Accessed 2002

World Health Organization (1998) Emerging and re-emerging infectious diseases: fact sheet #97, World Health Organization, Geneva, http://www.who.int/inf-fs/en/fact097.html. Revised Aug 1998

World Health Organization (1999) Removing obstacle to healthy development, World Health Organization, Geneva, http://www.who.int/infectious-disease-report/index-rpt99.html

World Health Organization (2002) Antimicrobial resistance: fact sheet #194, World Health Organization, Geneva, http://www.who.int/inf-fs/en/fact194.htm. Revised Jan 2002

Chapter 28

Selected Online Earth Science/Geospatial Journals

- *Canadian Journal of Remote Sensing* Provides index to journal issues dating back to 1992.

- ESRI Digital Chart of the World & Data Quality Project Downloadable papers related to using ESRI's Digital Chart of the World data series.
- ESRI White Papers: Papers related to using ESRI products as well as GIS in general.
- *GeoInformatica*—An International Journal on Advances of Computer Science for Geographic Information Systems.
- geoinformatik online (Uni Münster).
- Geo-Informations-Systeme (Wichmann/Huethig).
- GIS World Magazine
- Grassclippings: The *Journal of Open Geographic Information Systems*.
- *International Journal of GIS* (Taylor & Francis, London).
- National Research Council, Board on Earth Sciences and Resources Online reports.
- *Photogrammetric Engineering & Remote Sensing* (American Society of Photogrammetry and Remote Sensing).
- The Harlow Report: Geographic Information Systems Newsletter covering current GIS-related topics.
- Biomedical/Health Information
- Selected Biomedical/Health Data Resources
  - National Library of Medicine Developed by the U. S. National Library of Medicine, this program offers access to most of the MEDLARS databases including: MEDLINE, HealthSTAR, AIDSLINE, BIOETHICSLINE, HISTLINE (History of Medicine).
  - NCBI PubMed Sponsored by the National Center for Biotechnology Information, this search interface covers all citations covered in MEDLINE and PreMEDLINE.
  - NCI CANCERLIT Produced by the National Cancer Institute's International Cancer Information Center, CANCERLIT indexes over 1.3 million citations and abstracts from over 4,000 sources, including biomedical journals, books and doctoral theses.
  - Agency for Toxic Substances and Disease Registry (ATSDR): An agency of the U. S. Department of Health and Human Services. Monitors exposure to hazardous substances from waste sites, unplanned releases, and other sources of pollution present in the environment. Access to the *HazDat Database* and full-text to *Public Health Assessments*.
  - Centers for Disease Control and Prevention (CDC): Includes the latest health information and news, publications (such as the *Morbidity and Mortality Weekly Report*), statistics, funding information and public domain computer software for working with public health data.
  - National Center for Chronic Disease Prevention and Health Promotion: Clearinghouse for information on chronic disease prevention. Access to various full-text publications and reports.
  - National Center for HIV, STD, and TB Prevention: Comprehensive guide to the prevention, treatment, and elimination of HIV, STDs, and TB. Full-text of the *HIV/AIDS Surveillance Report* and the *STD Treatment Guidelines* as well as access to several databases covering news releases, funding opportunities, and health services.
  - National Center for Health Statistics (NCHS): The nation's principal health statistics agency. Includes several statistical publications for download, including several fact sheets, news releases, and reports.
  - National Center for Infectious Diseases: Develops programs to evaluate and promote prevention and control strategies for infectious diseases. Includes information about many infectious diseases as well as online access to the publications *Emerging Infectious Diseases* and *Health Information for International Travel*.
  - National Institute of Allergy and Infectious Disease (NIAID): Includes news for consumers and professionals, full-text newsletters, and consumer fact sheets and booklets, as well as a list of research activities and clinical trials.
  - National Institutes of Health (NIH): Gateway to clinical and consumer oriented resources including health information, funding opportunities, and scientific resources.
  - U.S. Census Bureau: Social, demographic and economic information. Includes full-text to the *Statistical Abstract of the United States*.
  - Center for International Health Information Provides timely, reliable, and accurate information on the Population, Health, and Nutrition (PHN) sector in developing countries assisted by USAID. Full-text of *Country Health Profile Reports*, and *Population, Health and Nutrition Indicators*.
  - World Health Organization (WHO): Promotes technical cooperation for health among nations, carries out programs to control and eradicate disease and strives to improve the quality of human life. Includes full text to the *World Health Reports* and the *Weekly Epidemiological Record*.
  - Agency for Toxic Substances and Disease Registry (ATSDR): An agency of the U. S. Department of Health and Human Services. Monitors exposure to hazardous substances from waste sites, unplanned releases, and other sources of pollution present in the environment. Access to the *HazDat Database* and full-text to *Public Health Assessments*.

Appendix C: Glossary

a-axis a vector direction defined by the space group and crystal structure for a particular crystalline form a term used in crystallography.

absorption the process by which a substance or a xenobiotic is brought into a body (human or animal) or incorporated into the structure of a mineral.

acanthosis increase in thickness of stratum spinosum (specific layer in epidermis/skin).

acid rain contamination of rain by artificial pollutants or natural emissions (such as sulfur dioxide from volcanic activity) which produces an acid composition.

activity the thermodynamically effective concentration of a chemical species or component.

acute myocardial infarction (AMI) gross necrosis of the heart muscle as a result of interruption of the blood supply to the area.

adsorption the binding of a chemical compound to a solid surface.

advection a transport process in which dissolved chemicals move with flowing groundwater.

albedo the percentage of the incoming solar radiation reflected back by different parts of the Earth's surface.

aldosterone a steroid hormone produced by the adrenal gland that participates in the regulation of water balance by causing sodium retention and potassium loss from cells.

aliquot a known amount of a homogeneous material, assumed to be taken with negligible sampling error. When a sample is "aliquoted", or otherwise subdivided, the portions may be called split samples.

alkali disease disease affecting animals that ingest feed with a high selenium concentration, characterized by dullness, lack of vitality, emaciation, rough coat, sloughing of the hooves, erosion of the joints and bones, anemia, lameness, liver cirrhosis, and reduced reproductive performance.

alkalinity the capacity of solutes in a solution to react with and neutralize acid determined by titration with a strong acid to an end point at which virtually all solutes contributing to the alkalinity have reacted. In general the alkalinity in water equates with the bicarbonate concentration.

allergy immunologic state induced in a susceptible subject by an antigen (allergen).

alluvial deposited by rivers.

alteration (Earth science) a process due to high-temperature fluids and gases that occurs within the Earth's crust and results in the formation of new mineral suites that are in equilibrium with their environment. Alteration can also occur at low temperatures.

aluminosilicate a mineral composed dominantly of aluminum, silicon, and oxygen, and lesser amounts of cations such as sodium, potassium, calcium, magnesium, and iron.

amorphous a lack of crystallinity or the regular extended three-dimensional order of the atoms in a solid.

anaerobic/aerobic environmental conditions in which oxygen is absent/present.

analyte any substance whose identity or concentration is being determined.

anemia any of several conditions in which the oxygen-carrying capacity of the blood is below normal due to reductions in the number of red blood cells (hypocytic) and/or the amount of hemoglobin per red blood cell (hypochromic).

aneuploidy cellular state where there is an abnormal number of chromosomes, not a multiple of the haploid number of chromosomes.

aneurysm localized ballooning of the aorta or an artery, potentially causing pressure on adjacent structures and liability to rupture.

angiotensin a vasoconstrictive hormone.

antisense nucleic acid that has a sequence exactly opposite an mRNA molecule made by the body binds to the mRNA molecule to prevent a protein from being made.

apo without, especially metalloproteins without the metal/metals.

apoptosis programmed cell death, in which a cell brings about its own death and lysis, signaled from outside or programmed in its genes, by systematically degrading its own macromolecules.

aqueous speciation the partitioning of chemical components between various aqueous species in a solution: free species (e.g., Ca^{2+}), ion pairs (e.g., $CaCO_3^0$), and complexes (e.g., $Fe(CN)_6^{3-}$).

aquifer a water-bearing rock formation.

aquitard a rock formation with poor permeability and hence a poor water-bearing unit.

archaea prokaryotes lacking a nucleus as bacteria, but they are as different from bacteria as are humans they

represent their own evolutionary pathway they live in extreme places with high temperatures.

arenosols sandy soils with >65% sand-sized (0.05–2 mm) particles these soils have low moisture and low concentrations of most elements and are highly prone to causing deficiencies of micronutrients in crops.

aridisol soils found in arid and semi-arid environments characterized by a light color, poorly developed soil horizons, high soluble salt content, little organic material, and a coarse texture.

arrhythmia irregularity of the heart beat.

arthroconidia fungal spores released by fragmentation or separation of the cells of a hypha.

asbestos a commonly used term for a group of fibrous silicate minerals that includes extremely fibrous serpentine (chrysotile) and the amphibole minerals crocidolite, amosite, tremolite, actinolite, and anthophyllite.

asbestosis degenerative fibrosis of the lung resulting from chronic inhalation of asbestos fibers.

ascariasis an infection caused by the parasitic worm *Ascaris lumbricoides* that is found throughout temperate and tropical regions. Intestinal infection may result in abdominal cramps and obstruction, while passage through the respiratory tract causes symptoms such as coughing and wheezing. In children, migration of the adult worms into the liver, gallbladder, or peritoneal cavity may cause death.

ascidian any minute marine invertebrate animal of the class Ascidiacea, such as the sea squirt.

ash fine particles of pulverized rock ejected from volcanoes.

asphyxiant gas which produces suffocation by replacing oxygen in the respiratory system.

ataxia lack of coordination of muscle for voluntary movement.

atelectasis absence of gas in lung tissue from nonexpansion.

atherosclerosis irregularly distributed intimal deposits of lipid.

atomization the dispersion of fluids into fine particles.

atrium the upper chamber of each half of the heart.

atrophy diminished cellular proliferation.

attribute information about geographic features contained within GIS data layers, or *themes*.

auger effect phenomenon occurring when an electron is released from one of the inner orbiting shells, thereby creating two electron vacancies of the residual atom and repeated as the new vacancies are filled or X-rays are emitted.

autosome a chromosome not involved in sex determination. The diploid human genome consists of 46 chromosomes, 22 pairs of autosomes and 1 pair of sex chromosomes (the X and Y chromosomes).

auxotroph a microorganism possessing a mutation in a gene that affects its ability to synthesize a crucial organic compound.

atypia reactive cellular state, which does not correspond to normal form.

background the property, as applied to a location or measurements from such locations, of being due to natural processes alone and unaffected by anthropogenic processes. In some instances the term natural background is used to reinforce the non-anthropogenic aspect. With the global atmospheric transport of anthropogenic contaminants, e.g., persistent organic pollutants (POPs), it is a moot point whether background sites exist for some substances.

basal cell carcinoma slow growing, locally invasive neoplasm derived from basal cells of epidermis or hair follicles.

baseline a measure of the natural background or ambient level of an element/substance. Some people also suggest that baseline is the current background which could include natural and anthropogenic components.

basolateral membrane part of the plasma membrane that includes the basal end and sides of the cell.

basophilic degeneration pathologic change in tissue noted by blue staining of connective tissue with hema- toxylin-eosin stain.

beneficiation process of concentrating ores.

benign usual or normal the opposite of cancerous when applied to cells or tumors.

bioaccumulation process by which an element is taken into an organism, possibly transformed into another chemical species, and retained so that the element's concentration in the biota is greater than its concentration in the media in which the biota is sustained.

bioapatite the name given to the complex calcium phosphate mineral that forms in biological tissues and is characterized by extremely small crystallite size maximum dimension is typically less that 20×10^{-9} m (200 Å). Generalized chemical formula: $(Ca,Na,Mg,\ldots[\,])_{10}(PO_4, HPO_4, CO_3, SO_4\ldots)_6(OH, F, Cl, CO_3, O, [\,])_2$

where. . .indicates the possible addition of other cations and [ ] indicates vacancies in the crystal structure at the cation or halogen sites.

bioavailability the property of a substance that makes its chemical uptake by biota possible.

bioessential/bioessentiality present in sufficient amounts to support essential biochemical processes imperative for sustaining life.

Biogeochemical cycle model encompassing the movement of elements (and some compounds) from the litho-sphere through the hydrosphere, atmosphere, and biosphere.

biosphere the sum of all organisms on Earth.

birefringence the ability of anisotropic (non-isometric) crystalline materials to split plane polarized light into two non-equal rays of distinct velocities depending on the direction of the transmission relative to the orientation of the atomic structure of the compound. When the two rays emerge from the crystal, one is retarded relative to the other. Precise measurements of the interference colors of the rays define the optical characteristics and identify the compound.

bisphosphonates a group of phosphorus- and carbon-containing compounds that have carbon connected to the phosphorus atom in place of one of the oxygen atoms of the tetrahedral phosphate (PO_4) groups.

blind staggers blind staggers occurs in cattle and sheep ingesting high concentrations of selenium and is characterized by impaired vision leading to blindness, anorexia, weakened legs, paralyzed tongue, labored respiration, abdominal pain, emaciation, and death.

bombs (volcanic) clots of lava that are ejected in a molten or semi-molten state and congeal before striking the ground.

bone a term applied to one of the many individual organs that make up vertebrate skeletons, or alternatively, to the fragments or the tissues that are found within these organs.

Bowen's disease an intraepidermal carcinoma characterized as a small, circumscribed elevation on the skin.

buffer a chemical compound that controls pH by binding to hydrogen ions.

bulk analysis chemical analysis of an entire body/substance of rock or soil or a subpart with little or no segregation of specific areas or components.

c-axis a vector direction defined by the space group and structure of a particular crystalline form. A crystallographic term.

calcisols soils with a high content of free calcium carbonate either developed on limestones, or which have become calcified by the deposition of calcium carbonate in pores and voids as a result of the evaporation of soil solution in arid environments. These soils generally have neutral or alkaline pHs and can adsorb some trace elements very strongly.

calcitonin hormone secreted by the thyroid gland important in the homeostatic regulation of serum calcium levels.

capillary electrophoresis electrophoretic separation technique performed in a small fused silica capillary.

carbon dioxide a colorless odorless gas high concentrations CO_2 acts as an inert asphyxiant in humans.

carbonatite an igneous rock composed of carbonate minerals.

carcinogen a substance that can directly or indirectly cause a cell to become malignant.

carcinogenesis the mechanism by which cancer is caused.

cardiomyopathy disease of the heart muscle (myocardium).

cardiovascular disease (CVD) disease pertaining to the heart and blood vessels, including, for example, both AMI and cerebrovascular disease (stroke).

catecholamines category of compounds including the neurotransmitters adrenaline and noradrenaline.

cation exchange exchange of cations between a solution and a negatively-charged solid phase (e.g., a clay mineral) in response to a change in solution conditions this is especially important in geochemistry for major cations such as calcium and sodium.

cation exchange capacity (CEC) the ability of a soil or soil constituent (e.g., clay mineral or humus) to adsorb cations on permanent, or pH-dependent, negatively charged sites on surfaces. Cations of different elements can replace each other as counter ions to the negative charges.

cDNA complementary DNA: a DNA molecule copied from an mRNA template by the enzyme reverse transcriptase.

cementum the thin tissue that forms the outer covering of a tooth below the gum line, similar in composition to dentine.

chaperones proteins that help in folding proteins correctly and that discourage incorrect folding. Metallochaperones assist in the delivery of metal ions to target proteins or compartments.

chelate the complex formed through the bonding of a metal ion with two or more polar groupings within a single molecule.

chitin a tough white to semi-transparent substance that forms the major structural component of arthropod exoskeletons and the cell walls of certain fungi.

chloroplast chlorophyll-containing photosynthetic organelle in some eurkaryotic cells.

choroid plexus a network of intersecting blood vessels of the cerebral ventricles that regulate intraventricular pressure.

chromatin the complex of DNA and proteins that make up eukaryotic chromosomes.

chromatography the separation of a mixture of compounds using solid, liquid, or gas phases based on affinity of molecules for the phase.

chromosome aberrations any deviation from the normal number or morphology of chromosomes.

clay minerals phyllosilicate minerals with a small grain size, commonly <4 μm but ranging down to colloidal dimensions. When mixed with a limited amount of water they develop plasticity. Clay minerals are formed by high-temperature hydrothermal alteration processes, e.g., kaolinite in altered granitic rocks or by low-temperature weathering processes, e.g., montmorillonite, smectite, chlorite, kaolinite, and illite.

clearance output of particles previously deposited in the respiratory tract.

coccidioidomycosis a respiratory disease of humans and animals caused by inhalation of arthroconidia of the soil-inhabiting fungus *Coccidioides immitis*. Fever, cough, weight loss, and joint pains characterize the disease, also called valley fever.

code (biological) the presentation of the content (of a molecule) in terms of symbols such as ATC and G for the DNA code where ATC and G are nucleotide bases.

codon the fundamental unit of the genetic code consisting of a triplet sequence of nucleotide bases which specifies the ribosomal binding of a specific amino-acid-bearing tRNA during protein synthesis or the termination of that process.

coenzyme a small molecule which binds to a protein to create a catalytic center.

collagen protein making up the white fibers (collagenous fibers) of skin, cartilage, and all connective tissue.

collimator a device for producing a beam of parallel rays.

compartment a separated solution volume of a cell by an enclosing membrane not at equilibrium with any other separated volume.

complex system natural or man-made system composed of many simple nonlinear agents that operate in parallel and interact locally with each other at many different scales. The behavior of the system cannot be directly deduced from the behavior of the component agents and the system sometimes produces behavior at another scale, which is called emergent behavior.

composite a mixture of several components or parts blended together to form a functional whole.

condensation polymer a polymer formed by loss of water molecules from monomers.

confined aquifer aquifer over- and underlain by impermeable or near-impermeable rock strata.

cooling the decrease of the activity of a radioactive material by nuclear decay.

coordination the association of one atom with another in three-dimensional arrays. The coordination number reflects the atomic size of an atom. Octahedral or sixfold coordination is typical of metal atoms with oxygen.

coronary heart disease (CHD) disease caused by deficiency of blood supply to the heart muscle due to obstruction or constriction of the coronary arteries.

cortical the tissue that forms the external portions of bones heavily mineralized with bioapatite-containing cells and exhibiting a variety of textures.

Cretaceous/Tertiary (K/T) boundary the Cretaceous period was the last in the Mesozoic era and was succeeded 64 million years ago by the Tertiary period of the Cenozoic era. It is marked by the sudden extinction of genera of living organisms, most famously the dinosaurs.

crust the outermost solid layer of a planet or moon.

crystallinity the three-dimensional regular array typical of solids with definite chemical composition and crystal structure.

crystalline basement solid igneous, sedimentary, or metamorphic rock may crop out at the ground surface or be overlain by superficial deposits (unconsolidated sediments or soils).

crystallite a general term applied to very small size materials, usually minerals, in which a crystal form or crystal faces may be observed, usually with magnification. The morphology of a crystallite suggests a material with a regular crystal structure and may be used to identify a specific compound or mineral species.

Cytochrome P-450 iron-containing proteins important in cell respiration as catalysts of oxidation-reduction reactions.

cytoplasm the central compartment of all cells that contains genes and DNA as well as synthetic systems.

database a structured set of persistent data, that in a GIS context, contains information about the spatial locations and shapes of geographic features, and their *attributes*.

decay (radioactive) the disintegration of the nucleus of an unstable atom by spontaneous fission or emission of an alpha particle or beta particle.

deconvolution a mathematical procedure used for separation of overlapping peaks.

definitive host the host in which a parasite reaches sexual maturity and reproduces.

dental calculus calcium phosphate mineral materials deposited around the teeth at and below the gumline, probably the result of bacterial action.

dental caries cavities in teeth arising from tooth decay.

dentine the tissue composed of greater than 70% bioapatite that forms the predominant segment of a tooth. This tissue is capped by enamel.

deposition fraction of particles in inspired air that are trapped in the lung and fail to exit with expired air. In geology it is the laying down of sediments.

derivatization the chemical modification of a naturally occurring compound so that it may be more volatile for gas chromatographic separation.

dermis inner aspects of skin that interdigitate with epidermis and contain blood and lymphatic vessels, nerves, glands, and hair follicles.

desorption release of a bound chemical compound from a solid surface (the opposite of adsorption).

detection limit minimum amount of the characteristic property of an element that can be detected with reasonable certainty under specific measuring conditions.

diagenesis changes to the original organic composition of a material caused by low-temperature processes, often involving bacterial action. It can occur in sediments where minerals are altered as well as organic matter. It changes the original chemistry of many minerals and bone when they are buried.

dioxygenase a class of oxidoreductases that catalyze the binding of diatomic oxygen to a product of the reaction.

DOC (dissolved organic compounds, or dissolved organic carbon) the soluble fraction of organic matter in soils and ground and surface waters comprising low molecular weight organic compounds which have the ability to complex many elements and render them more available to plants and more prone to leaching down the soil profile.

dose a general term for the quantity of radiation. The *absorbed dose* is the energy absorbed by a unit mass of tissue whereas the *dose equivalent* takes account of the relative potential for damage to living tissue of the different types of radiation. It is also the quantity of a substance taken in by the body in general.

dose response the relationship between an exposure dose and a measurable biological effect.

dowagers hump the abnormal concave bending of the upper or thoracic spine as a result of osteomalcia or osteoporosis often obvious in older women.

drift (analysis) a slow change in the response of an analytical instrument (geology) it is a superficial sediment.

dry matter (d.m.) remaining solid material after evaporation of all water. Often used to express concentration of minerals and trace elements to eliminate variation due to differences in water content of plant material.

ectodermal relating to ectoderm, the outer layer of cells in the embryo.

eco-district/eco-classification a relatively ecologically homogeneous area of the Earth's surface, an element of a classification based on climatic, biological, pedological, and geological criteria that becomes more specific from eco-zones, through eco-provinces and eco-regions to eco-districts.

effluent the material that is coming from a chromatographic separation. Can also be the waste outfall from industries and is also the term for sewage (sewage effluent).

eggshell calcification a thin calcified layer surrounding an intrathoracic lymph node.

elastosis degenerative changes of collagen fibers with altered staining properties.

electromagnetic spectrum the full range of frequencies, from radio waves to cosmic rays.

electrospray ionization (ESI) ionized molecules by application of a high voltage (approximately 5 kV) to the spray needle.

elimination how xenobiotics are removed from the bloodstream, either by metabolism or excretion.

emissions (volcanic) any liquid, solid, or gaseous material produced by volcanic activity.

enamel the tissue composed of greater than 96% bioapatite that forms the outer surface of teeth.

enantiomer one of two indistinguishable forms of a compound that differ only in the orientation in space a stereoisomer.

endemic Where a disease is confined to specific geographical areas.

endocytosis the process in which the plasma membrane engulfs extracellular material, forming membrane-bound sacs that enter the cytoplasm and thereby move material into the cell.

endosome a small vesicle resulting from the invagination of the plasma membrane transporting components of the surrounding medium deep into the cytoplasm.

endospore an asexual spore formed by some bacteria, algae, and fungi within a cell and released.

endothelium a tissue consisting of a single layer of cells that lines the blood, lymph vessels, heart, and some other cavities.

enterovirus group of viruses transient in the intestine which includes poliovirus, echovirus, and Coxsackievirus.

entisol entisols are soils that formed recently and are often found on floodplains, deltas, or steep slopes where soil development is inhibited. They are weakly developed and lack distinct soil horizons. Entisols have a wide geographic and climatic distribution.

enzootic a disease that affects animals in a specific area, locale, or region.

enzyme proteins that act as catalysts driving plant and animal metabolism.

eosinophils a specific type of white blood cell.

epidemiology the study of the prevalence and spread of disease in a community.

epidermis outer aspect of skin with multiple layers.

eruption (volcanic) the ejection of tephra, gas, lava, or other materials onto the Earth's surface as a result of volcanic or geothermal activity.

erythrocyte a mature red blood cell. Erythrocytes are the major cellular element of the circulating blood, and transport oxygen as their principal function. An increase in the number of cells normally occurs at altitudes greater than 3000 m.

erythron a collective term describing the erythrocytes and their predecessors in the bone marrow.

erythropoiesis the formation of erythrocytes in the bone marrow.

estrogen category of steroid hormones produced by ovarian and adipose tissues that can effect estrus and a number of secondary sexual characteristics and is involved in bone remodeling.

etiological the cause of a disease determined by etiology, the branch of medical science which studies the causes and origins of disease. The etiological agent of coccidioidomycosis is *Coccidioides immitis*.

etiology the process underlying development of a given disease.

eubacteria true bacteria so named to differentiate them from archaea (earlier known as Archaebacteria).

eukaryote cells of organisms of the domain Eukarya (kingdoms Protista, Fungi, Plantae, and Animalia). Eukaryotic cells have genetic material enclosed within a membrane-bound nucleus and contain other membrane-bound organelles.

eutrophication nutrient enrichment of waters that stimulates phytoplankton and plant growth and can lead to deterioration in water quality and ecosystems.

evapotranspiration transfer of water from the soil to the atmosphere by combined evaporation and plant transpiration. It results in a concentration of solutes in the remaining water.

excretion excretion is the mechanism whereby organisms get rid of waste products.

exon a DNA sequence that is ultimately translated into protein.

Exposure response the relationship between how much of a xenobiotic is presented to a person or animal and what happens in their body.

extracellular space in tissue that is outside of cells.

FAO/*Unesco* Soil classification system the soil classification system developed for the joint project by the UN Food and Agriculture Organization and UNESCO to produce the Soil Map of the World (1:5,000,000) published from 1974 onward.

felsic igneous rock rich in feldspar and siliceous minerals (typically light-colored).

ferralsols reddish iron oxide-rich soils characteristic of the tropical weathering and soil-forming environment (humid tropics). These soils generally have a low fertility with low CECs and nutrient contents. Also called oxisols (U. S. Soil Taxonomy), ferralitic, or lateritic soils.

ferritin a soluble protein storage form of iron containing as much as 23% iron.

ferromagnesian a silicate mineral dominated by iron, magnesium, sometimes with aluminum.

fibroblastic cells secretionary cells of connective tissue.

fibroblasts cells that produce collagen molecules.

fibrosis formation of fibrous tissue.

fluorapatite a mineral, ideal formula $Ca_5(PO_4)_3$ F. One of the members of the calcium apatite mineral group.

fluoride F^-, the dominant form of fluorine found in water.

fluorite the dominant fluorine mineral, CaF_2 occurs as an accessory mineral in some sediments and igneous rocks and in some hydrothermal mineral veins.

fluorosis disease affecting bones and teeth, caused at least in part by exposure to high doses of fluoride. Dental fluorosis causes weakening and possible loss of teeth, and skeletal fluorosis causes bone deformation and disability.

fluvial pertaining to rivers and streams.

forestomachs two or three sac-like dilations of the esophagus seen in ruminants and kangaroos. The physiological function of these structures is to serve as fermentation tanks to make cellulose and other carbohydrates in the feed available for absorbtion in the gastrointestinal tract of the animal.

fraction in this context, a term used in sedimentology, pedology, and other physical sciences to describe the mechanical size range of a material.

Fuzzy system a system that uses fuzzy sets and if-then rules to store, compress, and relate many pieces of information and/or data in order to build a model free estimator.

Gamma ray a distinct quantity of electromagnetic energy, without mass or charge, emitted by a radionuclide.

genome the DNA (or for some viruses, RNA) that contains one complete copy of all the genetic information of an organism or virus.

genotoxic the ability of a substance to cause damage to DNA.

geothermal pertaining to the internal heat of the Earth. Geothermal zones are areas of high heat flow, where hot water and/or steam issue at the Earth's surface. They are found close to tectonic plate boundaries or associated with volcanic systems within plates. Heat sources for geothermal systems may be from magmatism, metamorphism, or tectonic movements.

gleys soils under reducing conditions caused by permanent or intermittent waterlogging characterized by pale colors and low concentrations of iron oxides.

gliosis a chronic reactive process in neural tissue.

Glutathione peroxidase a detoxifying enzyme in humans and animals that eliminates hydrogen peroxide and organic peroxides it has a selenocysteine residue in its active site.

glycolysis the energy-yielding metabolic conversion of glucose to lactic acid in muscle and other tissues.

gneiss banded, usually coarse-grained metamorphic rock, having been modified from its original mineralogy and texture by high heat and pressure (high-grade regional metamorphism).

goitrogen a substance which causes or enhances the symptoms of iodine deficiency, e.g., goiter formation.

granite a coarse-grained igneous rock, composed mainly of quartz, alkali, feldspar, and mica. Accessory minerals may also include apatite, zircon, magnetite, and sphene. Granite characteristically has a high proportion of silica ($>70\%$ SiO_2) with high concentrations of sodium and potassium.

granitization a metamorphic process by which sedimentary and metamorphic rocks with a chemistry similar to granites (granitoids) are transformed mineralogically into rocks that look like the granites formed by igneous intrusive processes.

granulomatous inflammation inflammatory reaction where tissue cells of monocyte/macrophage cells predominate.

granulomatous reaction reaction leading to the formation of granuloma, or chronic inflammatory lesions.

grazing feeding behavior of cattle, sheep, and horses consumption of grass and other plants from the ground, mostly rather indiscriminately.

groundwater subsurface water in the zone of saturation in which all pore spaces are filled with liquid water (although sometimes the term groundwater is used inclusively for all water below the land surface, to distinguish it from surface water).

half-life the time in which one-half of the atoms of a particular radioactive substance decay to another nuclear form.

hardness water the content of metallic ions in water, predominantly calcium and magnesium, which react with sodium soaps to produce solid soaps or scummy residue and which react with negative ions to produce scale when heated in boilers.

haversian bone the tissue type found throughout the skeleton in humans that signifies sites of resorption and remodeling. Characterized in cross section by a circular outline and a lamellar distribution of cells and mineralized tissue around a central blood vessel, which is called the haversian canal.

heavy metal a metal with a density more than 4500kgm^{-3}.

helminth a multicellular worm, generally parasitic, often with a complex reproductive system and life cycle. Generally 50–2000 μm in length, but may be longer.

heme the protoporhyrin component of hemoglobin (in erythrocytes) and myoglobin (in myocytes), the proteinaceous chelation complexes with iron that facilitate transport and binding of molecular oxygen to and in cells.

hemolysis lysis of erythrocytes that potentially causes anemia.

hemorrhage profuse bleeding from ruptured blood vessels.

hemosiderin an insoluble iron-protein complex that comprises a storage form of iron mainly in the liver, spleen, and bone marrow.

hepatolenticular hepato, means belonging to the liver lenticular means lens shaped and refers to the basal ganglia of the brain.

herbivores animals normally feeding on plant material such as cattle, horses, sheep, antelope, deer, and elephants, but also rodents like mice, rabbits, and hares. As vertebrates lack enzymes in the gastrointestinal tract that can digest cellulose and other complex carbohydrates present in plants, they utilize microorganisms living in their gastrointestinal tract for this process. See also Ruminants and Large Intestine Fermenters.

hexagonal a description of a specific crystallographic form in which the c-axis is perpendicular to three axes, usually designated as a axes, which are 120 degrees relative to each other. Apatite crystals often show hexagonal prisms with a 60 degree angle measured between adjacent vertical or prism crystal faces.

histology science concerned with the minute structure of cells, tissue, and organs, utilizing light microscopy.

histomorphometry the study of the textures of tissues using sections of samples embedded in paraffin or epoxy. The sections cut from the embedded blocks may be stained to assist in the identification of specific tissue components, i.e., collagen or special components in the nucleus of a cell.

histones the family of five basic proteins that associate tightly with DNA in the chromosomes of eukaryotic DNA.

homeostasis the state of equilibrium in the body with respect to various functions and the chemical compositions of fluids and tissues, including such physiological processes as temperature, heart rate, blood pressure, water content, blood sugar, etc., and the maintenance of this equilibrium.

homeostatic control the ability or tendency of an organism or cell to maintain internal equilibrium by adjusting its physiological processes.

homologue a member of a chromosome pair in diploid organisms or a gene that has the same origin and functions in two or more species. To an organic chemist this is series of compounds that are similar in structure. For instance methanol, ethanol, and the other alcoholsrepresent a homologous series of compounds.

hormone a circulating molecule released by one type of cell or organ to control the activity of another over the long term, e.g., thyroxine.

host a human or animal in which another organism, such as a parasite, bacteria, or virus, lives.

humus the fraction of the soil organic matter produced by secondary synthesis through the action of soil microorganisms it comprises a series of moderately high molecular weight compounds that have a high adsorptive capacity for many metal ions.

Hydraulic conductivity the volume of water that will move in unit time under a unit hydraulic gradient through a unit cross-sectional area normal to the direction of flow.

Hydraulic gradient the change in static head (elevation head + pressure head) per unit distance in a given direction. It represents the driving force for flow under Darcy's law.

hydrodynamic dispersion the irreversible spreading of a solute caused by diffusion and mechanical dispersion (which, in turn, is caused by indeterminate advective transport related to variations in velocity about the mean).

hydroxylapatite name of the mineral, ideal chemical formula $Ca_5(PO_4)_3(OH)$, one of the members of the calcium apatite mineral group. Hydroxylapatite occurs naturally throughout the different types of rocks on the surface of the Earth and closely resembles the mineral deposits in normal and pathological tissues. See also Bioapatite.

hyperchromatic excessive dark staining.

hyperkeratosis hyperplasia of the stratum corneum (specific layer in epidermis/skin), the outermost layer in the epidermis.

hyperplasia an increase in the number of cells in tissue or an organ.

hypertension high blood pressure.

hyphae the branching threadlike filaments, generally 2–10 μm across, characteristic of the vegetative stage of most fungi.

hyphenated techniques generally, two analytical methods connected in series, e.g., a chromatographic technique directly connected to a spectroscopic technique.

hypoxia less than the physiologically normal amount of oxygen in organs/tissues.

idiopathic describing a disease of unknown cause.

igneous rocks formed from the cooling and solidification of molten rock originating from below the Earth's surface, includes volcanic rocks.

incidence quantifies the number of new cases/events that develop in a population at risk during a specified time interval.

inductively coupled plasma (ICP) an argon plasma with a temperature of approximately 7000–10,000 K, produced by coupling inductively electrical power to an Ar stream with a high-frequency generator (transmitter). Then plasma is used as an emission source (atomic emission spectrometry) or as an ionization source (mass spectrometry).

inselberg an isolated peak of hard rocks that has stubbornly resisted erosion most commonly found in the tropics.

integrin a membrane protein that conveys information in both directions across the plasma membrane.

internal dose amount of an agent penetrating the absorption barriers via physical or biological processes.

iodothyronine deiodinase selenoproteins responsible for the production and regulation of the active thyroid hormone from thyroxine.

ischemia ischemia occurs due to the disruption of the supply of blood and oxygen to organs and cells.

isoform the descriptor for a specific form of a protein that exists in multiple molecular forms also, for enzymes, isozyme.

isotachophoresis separation mode in capillary electrophoresis, separating according to analyte conductivity.

isotope one of two or more atoms with the same atomic number but with different atomic weights.

Kashin-Beck disease an endemic osteoarthropathy (stunting of feet and hands) causing deformity of the affected joints occurs in Siberia, China, and North Korea.

keratinocytes cells of the epidermis that produce the protein keratin.

Keshan disease an endemic cardiomyopathy (heart disease) that mainly affects children and women of childbearing age in China.

kinase an enzyme catalyzing the conversion of a proenzyme, or zymogen, to its metabolically active form, frequently via phosphorylation or proteolytic cleavage.

K_m the Michaelis constant in enzyme kinetics.

lahar a hot or cold flow of water-saturated volcanic debris flowing down a volcanic slope.

lamellar bone the tissue that shows sequential layers of mineralized matrix, cells, and the blood system required to maintain its viability. This tissue probably represents a second stage after the initial deposition of woven bone.

large intestine fermenters different animal species utilizing bacteria and protozoa in their large intestine (cecum or colon) to digest cellulose and starch in plants eaten so the nutrients can be absorbed in the gut of the animal. Horses, donkeys, zebras, rabbits, and hares are examples of animal species utilizing large intestine fermentation to facilitate digestion.

lattice an array with nodes repeating in a regular three-dimensional pattern. A crystal lattice is the array distinctive for the chemical and physical structure of the crystalline compound.

lava magma which erupts onto the Earth's surface lava may be emitted explosively, as lava fountains, or by oozing from the vent as lava flows.

leachate a liquid that carries dissolved compounds from a material through which it has percolated (e.g., water which carries adsorbed elements from settled volcanic ash into soil or water).

Lewis acid a chemical center which accepts electron pair donation from a donor base, e.g., M^{2+} is a Lewis acid in the complex $M^{2+} \underleftarrow{} OH_2$.

Lewy bodies intracytoplasmic inclusion seen in Parkinson's disease.

lichenoid accentuation of normal skin markings.

ligand a binding unit attached to a central metal ion.

limestone a sedimentary rock composed of calcium carbonate.

lithosphere the solid Earth.

lOAEL the lowest dose at which adverse effects are observed to occur in an experimental setting.

loess natural sedimentary formation made up of wind-lain mineral dust, mainly in the silt size range (1–60 μm), most of which accumulated, often in great thickness, during the Quaternary (the last about 2.6 million years).

lumen a cavity of passage in a tubular organ the lumen of the intestine.

lymph nodes small nodes along the bronchi that drain the tissues of lymph fluid.

lymphatic vascular channel that transports lymph, a clear fluid with predominantly lymphocytes.

lysis destruction of a cell's plasma membrane or of a bacterial cell wall, releasing the cellular contents and killing the cell.

macronutrient general term for dietary essential nutrients required in relatively large quantities (hundreds of milligrams to multiple grams) per day includes energy (calories), protein, calcium, phosphorus, magnesium, sodium, potassium, and chloride.

macrophage mononuclear phagocytes (large leukocytes) that travel in the blood and can leave the bloodstream and enter tissues protecting the body by digesting debris and foreign cells.

magma any hot mobile material within the Earth that has the capacity to move into or through the crust.

marine black shales sedimentary rocks formed from organic-rich muds which have developed under strongly reducing conditions and are generally enriched in a wide range of trace elements.

matrix the basis or collection of materials within which other materials develop. The organic matrix is the base in which mineral materials are deposited to form bone.

matrix effect the combined effect of all components of the sample other than analyte on the measurement of quantity.

melanin dark pigment that provides color to hair, skin, and the choroid of the eye.

mesothelioma a highly malignant type of cancer, usually arising from the pleura, which is the lining of the thoracic cavity, and characteristically associated with exposure to asbestos.

messenger (transmitter) a molecule or ion used to convey information rapidly in or between cells, e.g., Ca^{2+}.

metabolism the enzymatic chemical alteration of a substance. In toxicology, how xenobiotics are converted chemically in life sciences generally, the pathways of chemical reactions that occur in the body.

metabolome the small organic molecule composition in concentration units of a cell or compartment.

metadata data about data, typically containing information such as time and place of database creation, *field* and *record* identifier information (attributes), data development process, map projection, and person to contact regarding the database also known as data dictionary.

metalliferous rich in metals.

metalloid an element which behaves partly as a metal and partly as a non-metal, sometimes referred to as a "semi-metal."

metallome the element composition in concentration units of a whole or a part of a cell where the element may be in free or combined form.

metamorphic rocks rock formed from the alteration of existing rock material due to heat and/or pressure.

micellar electrokinetic chromatography separation mode in capillary electrophoresis, separating according to the ability of apolar analytes to enter the (apolar) core of surface charged micelles.

micronutrient general term for dietary essential nutrients required in relatively small amounts (less than multiple milligrams) per day includes the vitamins and trace elements.

microradiograph a picture produced using X-rays or rays from a radioactive source showing the minute internal textures of a planar thin section of a mineralized tissue sample.

mineral a naturally occurring compound with definite chemical composition and crystal structure, of which there exist over 4000 officially defined species.

mineral elements equal to elements. This term is used by nutritionists.

mineral group an aggregate of mineral species that shares structural and chemical affinities.

mineral nutrient a metal, non-metal, or radical that is needed for proper body function and maintenance of health also used in reference to plant nutrition.

mineralization the presence of ore and non-ore (gangue) minerals in host rocks, concentrated as veins, or as replacements of existing minerals or disseminated occurrences typically gives rise to rocks with high concentrations of some of the rarer elements.

mitochondrion subcellular organelle containing the electron transport chain of cytochromes and the enzymes of the tricarboxylic acid cycle and fatty acid oxidation and oxidative phosphorylation, thus, constituting the cell's primary source of energy.

mitogenic a factor that causes mitosis of cells.

mitosis the division of a cell into two daughters with identical complements of the nucleic material (chromosomes) characteristic of the species.

model a conceptual, physical, or mathematical representation of a real system or process.

monoclinic the description of a special crystallographic form for the structure of a compound in which the three axes are not mutually perpendicular.

monooxygenase a class of oxidoreductases that catalyze the dissociation of molecular (diatomic) oxygen such that single oxygen atoms are bound to different products of the reaction.

mT metallothionein.

mucosal cell cell of the mucous membranes of the gastrointestinal tract.

multichannel analyzer (MCA) an instrument that collects, stores, and analyzes time- or energy-correlated events.

multistage carcinogenesis model a mathematical model that assumes a sequential series of DNA-damaging events is necessary for a single cell to become malignant. The model also assumes linearity at low doses.

mycelium the vegetative part of a fungus (or in some cases bacteria), consisting of a mass of branching, threadlike hyphae.

mycorrhizae symbiotic fungi which colonize the outer layers of the roots of many plant species and whose external mycelium effectively increases the effective absorptive surface area of the roots.

myocyte a muscle cell.

myxedematous cretinism form of mental retardation caused by perinatal iodine deficiency.

natural background a term used to describe the geochemical variability and the range of data values due to natural processes, that characterize a particular geological or geochemical occurrence. See also Background and Baseline.

nebulizer interface at plasma detectors for aerosol production.

necrosis cell death.

nephrotoxin cytotoxin specific for cells of kidney.

neurotransmitter any of several compounds released by neurons to stimulate other neurons.

neutrophil a specific type of white blood cell.

nOAEL the highest dose at which no observed adverse effects occur in an experimental setting.

nuclide a general term applied to any atom with data on the number of protons and neutrons in its nucleus.

odds probability of disease divided by probability of no disease (p/1-p) within a study group (e.g., exposed individuals).

odds-ratio ratio between odds for exposed and odds for non-exposed ($odds_{+exp}/odds_{-exp}$).

oligonucleotide a DNA polymer composed of only a few nucleotides.

omnivores animals normally feeding on both plant and animal material. Species considered omnivores are humans, dogs, and swine.

oncogene a gene that controls growth and when aberrant or when activated inappropriately may permit cancer to develop.

operon a cluster of genes with related functions that are under the control of a single operator and promoter, thereby allowing transcription of these genes to be turned on and off.

organ systems part of body performing a specific function.

organelle a compartment found in eukaryotes derived from captured bacteria and with residual independent genes, e.g., *mitochondria* which create useful energy from oxidation of sugars and *chloroplasts* which create useful energy from light-generating oxygen.

organization a managed flow of material and energy in contrast with static order.

orthogonal (analytical) speciation concept analytical strategies which employ combinations of various separation and/or detection methods are called orthogonal analytical concepts.

ortholog a gene in two or more species that has evolved from a common ancestor.

osteoblasts a bone-forming cell function with bone-removing cells (osteoclasts) in the normal process of bone remodeling.

osteoclasts multinucleate cells that destroy bone tissue.

osteomalacia impaired mineralization of bone tissues resulting in areas where mineral is missing. One possible cause of osteomalacia is a deficiency of vitamin D, the hormone required for adequate calcium absorption and deposition as bioapatite in bone tissues.

osteon the bulls-eye pattern of concentric rings of lamellar bone around a vascular canal. This structure is detected in tissue sections that form as a result of bone tissue remodeling. See Haversian Bone.

osteoporosis a generalized term for the loss of bone tissues in bone organs. There are multiple possible causes of osteoporosis and the loss may occur at any age, but it is more prevalent in older individuals. The variations of osteoporosis remain active areas for investigation.

osteosclerosis disease characterized by abnormal hardening of bone due to excessive calcification.

oxalic acid a dicarboxylic acid (ethane dioic acid, $C_2H_2O_4$) found in some plants and produced by molds forms stable chelation complexes with divalent cations (Ca^{2+}, Mg^{2+}, Fe^{2+}, Zn^{2+}, Cu^{2+}) rendering them unavailable from the diet.

oxidation chemical process which can lead to the fixation of oxygen or the loss of hydrogen, or the loss of electrons the opposite of reduction.

oxidoreductase an enzyme that catalyzes an oxidation-reduction reaction.

p*53* gene a tumor-suppressor gene that codes for a transcription factor involved in preventing genetically damaged cells from proliferating.

Paget's disease a disorder in which the normal resorption and sculpting of bone is compromised and superfluous or more dense mineralized tissue is deposited.

parakeratosis retention of nuclei in the cells of the stratum corneum.

parasitimia the condition of having parasites within the bloodstream. Usually the parasite is a protozoan.

parathyroid hormone hormone secreted by the parathyroid gland important in the homeostatic regulation of serum calcium levels.

parent material the weathered rock material on which a soil is formed. Can be either fragments of the underlying solid geology or transported drift material overlying the solid geology.

parenteral administration of substance into organism not through gastrointestinal tract but through intramuscular, subcutaneous, or intravenous injection.

parkinsonism clinical syndrome characterized by diminished facial expression, slowness of voluntary movement, rigidity, tremor, and stooped posture.

pedogenesis the process of soil formation involving various physical and chemical processes which give rise to the formation of a soil profile. The nature of soil formed is determined by the interactions of the climate, vegetation, parent material, topography, and time.

periodic table a tabular classification of the chemical elements whereby they are organized into (vertical) groups based on progessive increases in numbers of electron shells surrounding the atomic nucleus and (horizontal) rows based on changes in the internal complexities of the electron shells. Elements within any group have similar chemical properties.

periplasm a secondary enclosed compartment of a prokaryote outside the cytoplasm and surrounding it.

permafrost permanently ice-bearing frozen ground, found in the Arctic, Antarctic, and some high-altitude regions.

pH a measure of the acidic (or alkaline) nature of an aqueous solution, expressed as the negative base −10 logarithm of the activity of protons in the solution. Solutions with pH values below 7 are considered acidic values greater than 7 indicate basic (or alkaline) conditions.

phagocytosis a type of endocytosis in which extensions of a plasma membrane engulf extracellular particles and transport them into the interior of the cell.

pharmacognosy the study of the useful drug effects of natural products.

phase a volume of space, solid, liquid, or gas in equilibrium with other volumes and described by a boundary. A homogeneous, distinct portion of a chemical system.

phase diagram a graphical representation of the stability relationships between phases in a chemical/physical system usually representing states at equilibrium. The presentation usually depicts relationships based on changes in composition, temperature, or pressure.

phenotype the physical characteristics of an organism that can be defined as outward appearance (such as flower color), as behavior, or in molecular terms (such as glycoproteins on red blood cells).

phosphorite a sedimentary rock with a high percentage of phosphate materials, shell, or bone fragments that may be mined for use as fertilizer. Prominent textural features are often nodules and pellets of extremely fine-grained calcium phosphate.

photoelectron electron that is ejected from the surface when light falls on it.

phyllosilicate a group of aluminosilicate minerals that have a sheeted crystal structure which permits cations to be trapped between the sheets and around the sheet edges. Because of these properties some are capable of sequestering geochemically significant amounts of cations, metals.

phytic acid inositolhexaphosphoric acid ($C_6H_6O_6[H_2PO_3]_6$) found in plants forms stable chelation complexes with divalent cations (Ca^{2+}, Mg^{2+}, Fe^{2+}, Zn^{2+}, Cu^{2+}) rendering them unavailable from the diet.

phytoavailability a specific instance of bioavailability with reference to plants. In some instances it is useful to differentiate between phyto- and bioavailability along the food chain. Phytoavailability controls the transfer of a trace element from soil to a plant, and bioavailabilty controls the transfer of the trace element from the plant material to the receptor organism the transfer factors are unlikely to be the same.

phytosiderophores organic compounds released by the roots of some plants suffering from a deficiency of iron or certain other micronutrients. They mobilize iron and elements co-precipitated onto iron oxides and render them available for uptake by the plant.

phytotoxic toxic to plants.

pica a craving for unnatural articles of food. The name pica comes from the Latin for magpie, a bird that picks up a variety of things either to satisfy hunger or out of curiosity. Geophagy, the deliberate ingestion of soil, is a form of pica.

Placer deposits alluvial deposits which contain ore minerals (commonly native gold, platinum, diamond, cassiterite) in economic quantities these are heavy minerals which are concentrated by reworking of primary ore bodies. They typically concentrate in low-energy environments such as floodplains and deltas. Many important placer deposits occur also as beach placers where they have been concentrated by seawater movement.

platelet a non-nucleated, hemoglobin-free cellular component of blood that functions in clotting also called a thrombocyte.

platform a term used in geology to describe a large stable section of the Earth's crust that is unaffected by current mountain building. Commonly formed over long periods of time by the erosion of the Earth's surface to relatively low relief.

plaque the unwanted deposition of mineral materials in tissue areas such as in the vascular system or around teeth within the gum tissues.

pleiotropy a situation in which a single gene influences more than one phenotypic characteristic.

pleural plaques a fibrous thickening of the parietal pleura which is characteristically caused by inhalation of the fibers of asbestiform minerals.

pM standard: the PM (particulate matter) standard is based on the total mass of particles measuring 2.5 μm or less observed in a 24-h period.

pneumoconiosis a chronic fibrosing lung disease from contact with respirable mineral dusts examples include silicosis and asbestosis.

podsol a type of soil which can be found in cool, humid environments on freely drained parent materials usually under coniferous trees or ericaceous vegetation. Typically has an iron pan as a result of leaching. Also called spodosols in the USDA Soil Taxonomy classification.

polymorph a term applied in mineralogy to describe minerals with the same composition that can crystalize in multiple crystallographic forms. Possibly the most well-known polymorphic minerals are calcite and aragonite both have the chemical composition $CaCO_3$.

primary term used to describe position in the biogeochemical cycle refers to bedrock.

primitive cell a cell thought to have existed some 3–4 billion years ago, although a related form can be found in extreme anaerobic conditions today.

prions an infectious microscopic protein that lacks nucleic acid thought to be responsible for degenerative diseases of the nervous system called transmissible spongiform encephalopathies (TSE) transmissible within and between species.

progesterone the steroid hormone produced by the corpus luteum, adrenal cortex, and placenta that prepares the uterus for reception and development of the fertilized ovum.

progestins a general term for the natural or synthetic progestinal agents.

prokaryote cells of the domains Bacteria or Archaea. Prokaryotic cells have genetic material that is not enclosed in a membrane-bound nucleus they lack other membrane-bound organelles.

proteome the full complement of proteins produced (expressed) by a particular genome.

protista eukaryotic one-celled living organisms distinct from multicellular plants and animals: protozoa, slime molds, and eukaryotic algae.

protozoa comprise flagellates, ciliates, sporozoans, amoebas, and foraminifers.

pulmonary alveoli out-pouchings on the fine lung passages in which oxygen exchange between the alveoli and the bloodstream occurs.

pump (in the context of organisms) a mechanical protein-based device in a cell membrane for transferring material from one compartment to another.

Purkinje cells large nerve cells found in the cerebellum, a large portion of the posterior aspect of the brain.

pyrite iron sulfide (FeS_2), otherwise known as fool's gold occurs commonly in zones of ore mineralization and in sediments under strongly reducing conditions.

pyroclastic flow a fast-moving heated cloud of gas and volcanic particles produced by explosive eruptions or volcanic dome collapse.

Quaternary the most recent period of geological time, spanning 0–2 million years before Present divided into the earliest period, the Pleistocene (ending with the last glacial maximum), and the subsequent Holocene (the last 13,000 years).

quaternary structure the three-dimensional structure of a multisubunit protein particularly the manner in which the subunits fit together.

radioactivity atoms (known as radionuclides) which are unstable and will change naturally into atoms of another element accompanied by the emission of ionizing radiation. The change is called radioactive decay.

radionuclide a radioactive nuclide.

radon a colorless radioactive element comprises the isotope radon-222, a decay product of radium. ^{222}Rn (radon) is a gas. It occurs in the uranium-238 decay series and provides about 50% of the total radiation dose to the average person.

radon potential map a map showing the distribution of radon prone areas delineated by arbitrary grid squares, administrative or geological boundaries. The radon potential classification may be based on radon measurements in existing dwellings, measurements of radon in soil gas, or proxy indicators such as airborne radiometric measurements.

raman microprobe vibrational spectroscopic technique where light scatter allows for characteristic spectra of materials to be obtained.

raster a model of spatial data using an x,y coordinate system, rows and columns, and representing features as cells, or pixels, within.

reactive oxygen species general descriptor for the superoxide (O_2), singlet oxygen (O), and hydrogen peroxide (H_2O_2), each of which has a much greater chemical reactivity with intracellular nucleophiles (proteins, DNA) than molecular oxygen from which it is derived metabolically.

recessive a mode of inheritance in which a gene must be present from both parents for the trait to become manifest in an offspring.

recharge process by which water is added from the atmosphere or ground surface to the saturated zone of an aquifer, either directly into the aquifer, or via another formation.

record a unique entity, commonly in GIS a location, that possesses different values for its attributes in fields.

redox potential (pe or Eh) pe and Eh are related variables that express a measure of the ratio of the aqueous activity of an oxidized species (an electron acceptor, such as Fe^{3+}) to that of a reduced species (an electron donor, such as Fe^{2+}). The redox potential of a solution can provide a sense of the oxidizing or reducing nature of a solution or aqueous environment (oxic, suboxic, sulfidic, methanic).

redox reactions coupled chemical oxidation and reduction reactions involving the exchange of electrons many elements have changeable redox states in groundwater the most important redox reactions involve the oxidation or reduction of iron and manganese, introduction or consumption of nitrogen compounds (including nitrate), introduction or consumption of oxygen (including dissolved oxygen), and consumption of organic carbon.

reducing condition anaerobic condition, formed where nearly all of the oxygen has been consumed by reactions such as oxidation of organic matter or of sulfide reducing conditions commonly occur in confined aquifers.

reduction chemical process leading to the loss of oxygen or increase of electrons by a compound the opposite of oxidation.

reference nutrient intake (RNI) the daily dietary value of a nutrient above which the amount will almost certainly be adequate for everybody.

regolith a deposit of physically and/or chemically weathered rock material which has not developed into a soil due to the absence of biological activity and the presence of organic matter.

reitfield refinement a method of calculating the three-dimensional structure of compounds.

relational database database where data are organized according to the relationships between entities.

relative risk (RR) a risk is the number of occurrences out of the total. Relative risk is the risk given one condition versus the risk given another condition used in epidemiology.

repair (DNA) the action of biological machinery to fix damage, especially referring to maintenance of DNA integrity.

reservoir (biological) a host, carrier, or medium (such as soil), that harbors a pathogenic organism, without injury to itself in the case of carriers, and can directly or indirectly transmit that pathogen to individuals.

residence time period during which water, solutes, or particles remain within an aquifer or organisms as a component part of the hydrological cycle.

respiratory distress impairment of lung function, often resulting in uncomfortable respiratory symptoms, lowered oxygenation and/or elevated carbon dioxide levels in the blood.

retention time elution time of a compound in a chromatographic system depending on its interaction at the stationary phase.

rheumatoid indefinite term applied to conditions with symptoms related to the musculoskeletal system.

rhizosphere the zone around plant roots (2 mm thick) in which there is intense microbial activity due to root exudates and which has chemical properties different from the bulk of the soil.

ribozyme rNA molecule with catalytic activity.

rickets disease of children characterized by undermineralization of growing bone, leading to physical deformities of the weight-bearing bones most notably of the legs, wrists, and arms.

risk assessment a systematic way of estimating the probability of an adverse outcome based on the known properties of a hazard such as a chemical.

ruminants several groups of animal species utilizing bacteria, fungi, and protozoa in their forestomachs to digest cellulose and starch in plants eaten so the nutrients can be absorbed in the gut of the animal. Cattle, sheep, goats, antelope, deer, and camels are examples of ruminants.

saline intrusion phenomenon occurring when a body of salt water invades a body of fresh water it can occur either in surface water or groundwater bodies.

saprophyte an organism, often a fungus or bacterium, that obtains its nourishment from dead or decaying organic matter.

saprozoonoses zoonotic diseases where transmission requires a non-animal development site or reservoir. Soil can often serve as the reservoir.

sarcoidosis a systemic granulomatous disease of unknown cause.

sarcomatoid resembling a sarcoma, a neoplasm of soft tissue.

scanning electron microscope (SEM) a method employing an electron microscope and a finely-focused beam of electrons that is moved across a sample allowing the surficial textures to be examined at high resolution and the image displayed. By collecting the emitted electrons from a single spot (size 1–10 μm) chemical analysis of portions of the sample, i.e., a specific mineral species, can be made using energy dispersive X-ray analysis (SEM/EDXA).

screw axis a specific translational and rotational characteristic of a lattice direction (axis) defined as part of one of the known 230 space groups. The calcium apatite group has a screw axis designated as 6_3. The c-axis has sixfold-symmetry with a screw. The screw rotates 120 degrees around the sixfold-axis with each one-third translation along the axis, part of the space group designation of the apatite unit cell.

secondary terms used to describe position in the biogeochemical cycle refers to weathering products and processes resulting from, or acting on, primary rock material.

sedimentary rock rock formed by compression of material derived from the weathering or deposition of preexisting rock fragments, marine or other organic debris, or by chemical precipitation.

selenocysteine an unusual amino acid of proteins, the selenium analog of cysteine, in which a selenium atom replaces sulfur.

selenomethionine 2-amino-4-(methylseleno) butanoic acid.

selenosis selenium toxicity.

sesquioxide oxide mineral containing three atoms of oxygen and two atoms of another chemical substance. Iron and aluminum oxides are the most important in the natural environment.

shale a sedimentary rock composed of fine particles, mainly made up of clay.

silicate a mineral composed dominantly of silicon and oxygen, with or without other elements such as magnesium, iron, calcium, sodium, and potassium.

silicosis a form of pneumoconiosis produced by inhalation of fine silica particles.

smectite a group of clay minerals (phyllosilicates) that includes montmorillonite and minerals of similar chemical composition. They possess high cation exchange capacities, and are therefore capable of sequestering labile cations.

soil profile (solum) the vertical section of a soil from the surface to its underlying parent material. It comprises distinct layers (horizons) differing in appearance or texture and chemical properties. The soil profile is the basis of soil classification (soils with characteristic combinations of horizons).

soil texture the relative proportions of sand (0.05–2 mm), silt (0.002–0.05 mm), and clay (<0.002 mm) sized particles in a soil which affects both its physical and chemical properties.

solubility equilibrium concentration of a solute in water at a given temperature and pressure when the dissolving solid is in contact with the solution.

sorption the retention of ions on solid surfaces in the soil by a combination of mechanisms**:** ion exchange, specific adsorption, precipitation, and organic complexation.

space group a mathematical expression that uniquely defines the three-dimensional array typical of a crystalline material.

spallation splitting off, particularly applied to splitting off parts of the nucleus of an atom, resulting in the formation of a different element.

spherule a small spherical structure of the invasive phase of *Coccidioides immitis* that fills with endospores as it matures. The spherule ruptures at maturity releasing the infective endospores into the host.

spongiosis intercellular edema of epidermis.

spray chamber (chemical analysis) part of sample introduction system, connected to a nebulizer. Droplets from the aerosol that are too big are discarded.

squamous cell carcinoma malignant neoplasm derived from stratified squamous epithelium.

stable isotope isotope that does not undergo radioactive decay.

standardized mortality ratios a statistical method for comparing the mortalities of different population groups by separating data according to sex and then age band.

steatosis general term describing fatty degeneration t-RNA**:** transfer ribonucleic acid any of a number of such intracellular factors involved in protein synthesis by transferring in sequence individual amino acids to the ribosome.

stereoisomer one of two forms of a compound that is indistinguishable from the other outside of the orientation in space. An enantiomer.

stoichiometric a term applied when a phase or compound has the charge balance and chemical proportions expected in the ideal formula.

swayback neonatal ataxia, a clinical manifestation of copper deficiency in lambs. The condition is characterized by incoordination of movement and high mortality. The disease is known as lamkrius in South Africa, kipsiepsiep in Kenya, and enzootic ataxia in several other countries, including the former Soviet Union.

symbiosis the cohabiting of more than one organism which supply one another with vital material and energy.

synergy a positive interaction.

tachycardia rapid heart beat.

tachypnea rapid breathing.

tephra any solid material produced and made airborne by volcanic activity (including bombs, blocks, ash, and dust).

termite mounds a common source of geophagical material in the tropics. The edible part of a termite mound is the extremely mineraliferous, soft, protected interior comprising the queen's chamber, nursing galleries, and fungus gardens.

tetrahedral orthophosphate group the three-dimensional atomic array in which four oxygen atoms are distributed at the apices of the tetrahedron around the phosphorus atom.

theme (GIS) a GIS data layer, or coverage used in an overlay analysis with spatial referencing.

threshold in biology it is a dose level, below which, no adverse effect is expected. In Earth science it represents the upper or lower limit of background—above or below which is anomalous.

thylakoid a disk-shaped, membranous sac found in chloroplasts, the membranes of which contain the photosystems and ATP-synthesizing enzymes used in the light-dependent reactions of photosynthesis.

thyroxine also referred to as 3:5,3′:5′ tetra-iodothyronine (T_4) is the major hormone secreted by the thyroid gland. T_4 is involved in controlling the rate of met-abolic processes in the body and influencing physical development.

TNF tumor necrosis factor.

tomography a method employing transmission X-radiological analysis to visualize the bones or bony portions of the skeleton. The X-ray source moves relative to the patient.

tonsillar herniation physical displacement of cerebellar tonsil into foramen magnum, a large opening at base of the brain.

toxicity state of being poisonous and disturbing organ function.

toxicodynamics the mechanisms by which xenobiotics induce their effects in the body the mechanisms of the toxic response.

toxicokinetics the mechanisms by which xenobiotics are handled in the body, comprising the steps *absorption*, *distribution*, *metabolism*, and *excretion*.

toxicology originally, the study of poisons and now the general science of the handling by, and response of, the body to xenobiotics and the patterns of adverse effects that result.

toxocariasis also called visceral larva migrans (VLM), toxocariasis is caused through infection with the larvae of *Toxocara canis* or *T. cati* (the common roundworm of dogs and cats, respectively). After infection, the eggs hatch into larvae and are carried into the circulation and to various tissues. Respiratory symptoms develop, and there is a swelling of body organs such as the liver. A complication of VLM is epilepsy and ocular larva migrans, the latter caused by microscopic worms entering the eye.

toxoplasmosis a disease attributable to the ingestion of *Toxoplasma gondii*, one of the most common human parasites that infect 30–60% of the global population. Commonly caused by eating of undercooked meat with soil ingestion as secondary source. Recent research has suggested that human behavior can be adversely affected following *T. gondii* infection.

Trace elements (in medicine) general term for the nutritionally essential mineral elements that are required at levels of intake less than about $50 mg d^{-1}$ includes iron, copper, zinc, iodine, selenium, manganese, molybdenum, chromium, fluoride, and cobalt.

transcription the act of producing RNA from DNA leading to *translation*, protein production.

transfection the uptake and expression of a foreign DNA sequence by cultured eukaryotic cells or the introduction of foreign DNA into a host cell.

transposon a segment of DNA that can become integrated at many different sites along a chromosome (especially a segment of bacterial DNA that can be translocated as a whole).

trichiuriasis infestation with the roundworm *Trichuris trichiura* that may cause nausea, abdominal pain, diarrhea, and occasionally anemia and rectal prolapse.

triiodothyronine also referred to as 35,3′ triiodothyronine (T_3) produced in the thyroid gland and involved in controlling the rate of metabolic processes in the body and physical development.

trabecular the porous tissues forming the internal sectors of bones. The trabeculae are bone tissue spicules. This type of tissue is often adjacent to the hollow core or within the marrow cavity.

transmissible spongiform encephalopathies (TSE) rare forms of progressively degenerative diseases of the nervous system that affect both humans and animals. They are caused by agents called prions and generally produce spongiform changes in the brain. Examples include chronic wasting disease (CWD) in deer and elk, bovine spongiform encephalopathy (BSE) in cattle, and Creutzfeldt-Jakob disease (CJD) in humans.

type 1 collagen the special variety of the collagen molecule typically found in the matrix of tissues that will become mineralized as bone.

ultramafic rock igneous rock composed substantially of ferromagnesian silicate minerals and metallic oxides and sulfides, with <45% silica, and almost no quartz or feldspar.

ultrastructure morphometry of particles and cell structure based on electron microscopy.

unconfined aquifer aquifer containing unconfined groundwater, i.e., having a water table and an unsaturated zone.

unit cell the smallest geometric volume that uniquely defines the composition and precise structure of a crystalline compound. The basis for the repetitive pattern that completely characterizes a compound, its chemistry, and three-dimensional arrangements of all the constituent atoms.

USDA Soil Taxonomy the soil classification system devised by the United States Department of Agriculture (published in 1975).

vasodilation expansion of the blood vessels.

vadose zone also known as the "unsaturated zone" is the part of the Earth's surface extending down to the water table.

vector (GIS) model of spatial data using points, lines, and polygons to represent geospatial features and boundaries. Vector in the entomological sense, is typically an arthropod that transmits disease-causing pathogens to humans.

vector-borne disease disease that is transmitted from one vertebrate host to another by an invertebrate, usually an insect, tick, or snail.

viremia the existence of virus or viral particles in the bloodstream.

virulence the capacity of a microorganism for causing disease.

$\mathbf{v_{max}}$ the maximum velocity (never attained) in enzyme kinetics.

volatile fatty acids (VFA) common name for acetic acid, butyric acid, and propionic acid normally formed under

anaerobic conditions in the forestomachs and large intestine of herbivores. After absorption from the gastrointestinal tract, VFA can be further metabolized and used mainly for energy production. In ruminants, VFAs are the dominating energy source equivalent to glucose in the metabolism of other species.

volcanic gas gas produced by volcanic activity or geothermal processes. Steam is the most common gas those of relevance to health include the inert asphyxiants, irritant gases, or noxious asphyxiants.

volcanic monitoring geological and epidemiological testing and surveillance prior to, surrounding, and subsequent to an eruptive event or degassing episode includes the period of post-disaster recovery and rehabilitation.

volcano an opening in the crust from which gases, lava, and/or tephra are expelled.

voltammetry an electrochemical determination method based on the characteristic redox potential of the measured compound.

weathering a process at or near the Earth's surface caused by the interaction of water, oxygen, carbon dioxide, and organic acids with the minerals present includes hydrolysis and oxidation reactions. Weathering can result in the formation of new mineral suites that are in equilibrium with their environment. In Arctic and high mountainous regions chemical weathering may be limited, and weathering is largely limited to mechanical breakdown due to frost action that liberates fragments of the pre-existing minerals.

white muscle disease a complex medical condition, which is multifactorial in origin but linked to selenium deficiency, causes degeneration of the muscles in animal species. In lambs born with the disease, death can result after a few days. Later in life, animals have a stiff and stilted gait, arched back, are not inclined to move about, lose condition, and die.

world Reference Base for Soil Resources a classification system, database, and atlas produced by the working group RB International Society of Soil Science in 1998.

woven bone the first deposited bone tissue that may display a haphazard distribution of matrix, cells, vascular channels, and mineral and which is usually later reworked into lamellar or haversian bone over time.

xenobiotic a chemical substance foreign to the body or introduced to the body in higher quantities or by a different pathway than occurs in normal metabolism.

X-ray diffraction maxima the periodic coherent scattering of X-rays that arises from crystalline materials. These data are used to determine the coordinates from which the space group and unit cell of the compound can be determined.

X-ray/electron diffraction the method employed to examine the crystallinity and crystal structure of materials.

zoonotic/zoonosis a disease which has a natural reservoir in an animal or non-human species that can be transmitted to humans.

Index

O. Selinus et al. (eds.), *Essentials of Medical Geology: Revised Edition*,
DOI 10.1007/978-94-007-4375-5, © Springer Science+Business Media Dordrecht 2013