ERAS IN EPIDEMIOLOGY

The Evolution of Ideas

Merrye Susser
Zena Stein

Eras in Epidemiology

This page intentionally left blank

Eras in Epidemiology

The Evolution of Ideas

Mervyn Susser and Zena Stein

OXFORD
UNIVERSITY PRESS
2009

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that furthers
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016
www.oup.com

Library of Congress Cataloging-in-Publication Data
Susser, Mervyn.
Eras in epidemiology : the evolution of ideas / Mervyn Susser, Zena Stein.
p. ; cm.
Includes bibliographical references.
ISBN-13: 978-0-19-530066-6
1. Epidemiology—History. I. Stein, Zena. II. Title.
[DNLM: 1. Epidemiology—history. WA 11.1 S964e 2009]
RA649.S87 2009
614.4—dc22 2008042101

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

Preface

This work aims to follow the historical evolution of epidemiological ideas which are, specifically, ideas about health and disease in populations. We do so in terms of the successive material explanations that have shaped understanding about the causes of human disease. Our perceptions owe much to the fertile concept of scientific revolution that Ludwik Fleck advanced in his book (1935), a concept Thomas Kuhn (1962) further elaborated some three decades later. In that light, through the centuries, we seek to identify the shifts in the governing paradigms of collective thought bearing on disease in populations. Hence the title: *Eras in Epidemiology.*

For lack and deficiencies of data, we can give only a brief nod to the earliest accounts of thinking relevant to our interest. Ideas on health and disease seem first to have become apparent in the literature of Ayurvedic medicine of India. That literature provides scarcely a glimmer of ideas about health and disease *in populations.* From India, Ayurvedic ideas migrated to China perhaps half a millennium before the Christian era began. Thereafter, we trace such changes in key ideas discernible in our readings down to the end of the twentieth century. Where recorded knowledge and our own capabilities permit, this account seeks to identify those few historical events, the several exceptional individuals sufficiently gifted to influence rising generations,

and the key ideas and thought structures that wrought major shifts in the collective understanding of health and disease.

Following this reference to the earliest beliefs—in magical powers both friendly and malign, and in unaccountable, unidentifiable spirits as the mysterious forces governing human life—the story moves on through the medieval concepts of humors, and then to the pith of direct observation, to experiment, to analytic rigor, and so to the better understanding of the complexities of life and environment in all their forms. This historical path of change and development is sometimes slow and winding, sometimes steep and even precipitous. As each thought collective surmounts new pinnacles, so do many apparent peaks turn out simply to have obscured the summit beyond.

Our text thus has the central if ambitious aim of seeking, and recalling from recorded history, the governing concepts relevant to the epidemiological discipline. Such concepts may either advance or impede understanding. We shall try especially to identify and remark on those kernels that in some way bore fruit in collective thought.

To accomplish such an aim this account digs, if somewhat glancingly, into a history approaching 2,500 years. Given such a large and lengthy canvas, we can scarcely doubt that our adumbrations will reveal limitations recognizable in lapses of scholarship and unintended bias. Such weaknesses in scholarship as attend this work, we hope and reasonably expect, will be amended over time by the small but growing number of historians engaged with the history of science. Those dedicated historians excavating past history from contemporary records will surely dig deeper into what is relevant than we, lifetime practitioners of epidemiology, could do from the available literature we recognized as relevant.

Our own understanding of epidemiology evolved and accrued through the latter half of the twentieth century, during and after a medical schooling that began for both authors in South Africa in 1945 with the end of service in World War II. Outside that narrowly structured period of training, our friend Charlotte Kaplan introduced us to Sidney Kark, his wife Emily Kark, and their fellow workers: Guy Steuart, Harry Phillips, Eva Salber, Bert Gampel and Juliet Chester, Joseph Abramson, and John Cassel. On the Kark's' initiative this exceptional team had joined in the postwar development of government sponsored health centers with the enlightened primary aim of creating a national program of promotive health care: this was to be a program comprising both preventive and curative medicine, at the levels of individual, family, and community. From these friends and colleagues, as medical students we gained a first and long-lasting introduction to the theory and the practice of social medicine.

Upon graduating in 1950, the early seeds of our understanding were nurtured by our raw experience of practice among the roughly estimated 80,000 Africans massed in the square mile of the Township of Alexandra abutting on the northern edge of the City of Johannesburg. Together with Margaret Cormack and Michael Hathorn, another recently graduated couple, enthused and excited by the prospect, we were given joint charge of the Alexandra Health Center and University Clinic. Founded by medical students in 1940, the Health Center provided a formal segment of final year training for the medical students of Witwatersrand University in the City of Johannesburg. There, over the next five years, we experienced for ourselves the teaching and practice of epidemiology in the community. Given both our political contention with the repressive rigidity of the viciously race-based Apartheid regime, and our political activism in Alexandra Township in the face of that regime, unsurprisingly we found ourselves under police surveillance and government threat. In the face of persistent harassment and with the tenure of our posts at hazard, reluctantly, we departed our native land for England. A country with which South African history was closely intertwined, England was a natural refuge readily accessible to South Africans.

In 1956, at the outset of our English sojourn, we were as yet profoundly ignorant of the history of public health and its origins in revolutionary France of the late eighteenth century, and of its further development thereafter in nineteenth-century England, Germany, and elsewhere. Quite soon we had the good fortune to be sponsored by J.N. Morris, who features several times in this text, and Cicely Williams, the pediatrician famous for her description of Kwashorkor, the result of serious nutritional deficiencies in children, who had visited us in Alexandra. We also worked with Clement Toussaint, who introduced domiciliary care for tuberculosis in London as we had in Alexandra, then to be recruited to Manchester University's Department of Social and Preventive Medicine, headed by Professor C. Fraser Brockington. During the decade we spent there in teaching and research, we at once settled down to acquire an understanding of the history and the functions of public health in England, a history in which Brockington was especially well-versed. In parallel, we set about acquiring a deeper grasp of epidemiology from the works and with the help of all those engaged with academic social medicine. Promoted by a national government recovering from the drama and devastation of World War II, this was a discipline with epidemiology at its heart, although in our own view social science was a discipline scarcely less necessary for full understanding.

That gap was filled, as in pursuit of new research ventures with a content relevant to social and preventive medicine, we first worked with and were befriended by Sir Robert Platt (then chair of the Department of Medicine

at Manchester University and later, as Lord Platt, President of the venerable Royal College of Physicians in London), and Sir Douglas Black (First Medical Officer of Great Britain). Soon after our arrival in Manchester, we were made welcome by the anthropologists and sociologists in the Department of Social Anthropology headed by Professor Max Gluckman (himself a fellow South African sympathetic to us), and especially by William Watson, friend, mentor, and in time joint author of our text, *Sociology in Medicine* (1962). In this environment, our knowledge and understanding of both public health and social science grew apace. That learning gained impetus from participation in the academic life of the university generally, from our various researches and day-to-day work in health departments of surrounding towns, and from our reengagement with political life in a remarkably benign Britain by then shedding its empire.

In 1964, we received a joint invitation to spend the academic year of 1965 on sabbatical leave in the United States in New York City with a base at the Columbia University School of Public Health for Mervyn Susser and for Zena Stein likewise at the Association for the Aid of Crippled Children (AACC). In our travels that year across the United States, (sponsored by the Milbank Fund and the AACC), while both learning and lecturing we acquired a surprised new consciousness of the riches and productivity of postwar epidemiological research in the United States, the legacy from the early twentieth century of such pioneers as Joseph Goldberger, Edgar Sydenstricker, Wade Hampton Frost, and others.

Although comfortable in Manchester, we had nevertheless to consider invitations from the American schools we had visited to accept professorial chairs in epidemiology. Among them, the School of Public Health at Columbia University in New York City, by then not only familiar but seemingly well-suited to our interests, was congenial, and sited in a remarkable and lustrous city. It happened also that in Manchester, Fraser Brockington had chosen to retire, and the succession was in turmoil. In that light, there was little question that we should accept the Columbia offer. Now, happily in our fifth decade at Columbia and nearing the end of our careers—while as yet we retain our sense of affiliation with England and with a new post-Apartheid South Africa—we have only gratitude for a rewarding experience among the people of this university, this city, and this country.

Our understanding of the discipline of epidemiology had accrued step-by-step, first in the clinic and in the field, and thereafter in the classrooms, the journals, the workshops, the conferences, and the books of the period. Thus our own perceptions inhere largely in the jointly lived half-century and more that we have together devoted to the study and practice of epidemiology. Throughout, the demands of teaching, research, and practice gave direction to our thinking and writing. We are, of course, products of our own times

and thought collectives, with all the limitations that implies. Accordingly, with few exceptions, our personal story hardly ventures beyond our own era of the latter half of the twentieth century. In the year 2000, after all, the structure of the genome was not yet fully deciphered and published. Hence, while foretold, the impact on epidemiology of genomics and of epigenetics, too, was still to be fully realized. One current exemplar of these possibilities has been the advent of and the response to human immunodeficiency virus (HIV), as outlined in Chapter 23. Our exposition in the text that follows this preface aims to portray the history of the evolution of relevant ideas and the consequent range of understanding, biological as well as social, that an epidemiologist of our time and in the time to come might contribute. Not least, the planetary threat of global warming and climate change calls for a broad vision and appropriate action.

As with this book we end our efforts to penetrate and understand the evolution of epidemiology as a discipline, we might borrow an aphorism: the historian halts at the gateway of the future. It is for successors to enter in.

Mervyn Susser and Zena Stein

References

Fleck L (1935). *Genesis and Development of a Scientific Fact*. Chicago: University of Chicago Press (see 1979 edition with comment by Kuhn).

Kuhn TS (1962). *The Structure of Scientific Revolutions*. New York: Oxford University Press.

Susser MW and Watson W (1962). *Sociology in Medicine*. London: Oxford University Press.

This page intentionally left blank

Acknowledgments

Mervyn Susser had already begun the conceptualization and writing of this book when, about two years ago, our son Ezra Susser proposed that a collaboration with Zena Stein and himself would bring its development more fully and speedily to a timely conclusion. In the event, other demands deterred Ezra from full collaboration. Instead, he has given full and careful reading to the whole text, and has advised on general as well as specific issues. He has cheerfully agreed to the inclusion of the two published papers, originally co-authored with him, and now included as Chapters 24 and 25. With Dana March, he has contributed Chapter 26 as the concluding piece.

Ezra also introduced us to Kim Fader. Although initially she undertook the tracing and editing of the bibliography, Kim went well beyond that. She has carefully read every chapter, identified gaps, and taken charge of the bibliography. Manidipa Sengupta typed the final script, and earned our thanks as she dealt with the many emendations in draft after draft of the text. Benjamin Fader provided additional fact-checking help on more than half of the chapters.

Individual chapters have been read, criticized, and improved by knowledgeable and generous friends and colleagues. Mary-Claire King and Joseph Berg read Chapter 21; Stephen Morse read Chapter 22; Ida Susser read Chapters 20 and 23; Alan Berkman and Louise Kuhn read Chapter 23. As

we frankly admit, we are ourselves responsible for such errors of fact and judgment that remain.

Aside from crediting several books that have enlightened us with regard to specific and defined topics and periods, among many resources on which we drew are three admirable antecedents of our own work, although these were subject, as are we, to the limitations of their times. We refer in particular to C.E.A. Winslow, The Conquest of Epidemic Disease (1943); George Rosen, *A History of Public Health* (1958); and C. Fraser Brockington, *Public Health in the Nineteenth Century* (1965). Of course, we cite many more focused texts throughout our book. In particular, in the sequence of chapters dealing with the development of epidemiologic methods, it is a pleasure to acknowledge the relevance of *History of Epidemiologic Methods and Concepts*, written by our colleague and friend, Alfredo Morabia.

References

Brockington F (1965). *Public Health in the Nineteenth Century.* Edinburgh: E. & S. Livingstone.

Morabia A. *History of Epidemiologic Methods and Concepts.* Boston: Birkhauser, 2004.

Rosen G (1958). *A History of Public Health.* New York, NY: MD Publications.

Winslow CEA (1943). *The Conquest of Epidemic Disease.* Princeton, NJ: Princeton University Press.

Contents

Eras in Epidemiology

This page intentionally left blank

1

The Scope and Purposes of Epidemiology

In a current definition, epidemiology is the study of the distributions and determinants of states of health in human populations. The term seems consonant with its Greek origin: *epi*, upon and *demos*, the people (although the term *epidemic* seems originally to have described arrivals). The brief definition above has room for most present-day activities of epidemiologists. One might usefully add, however, that the ultimate purposes of such activity are the prevention, surveillance, and control of health disorders. The addition emphasizes a determinant of health that weighs heavily in public health and medicine, which is the conscious intervention that societies elect to undertake in matters that relate to health. The reach of epidemiology extends from describing the impact on health of environment and of the genetic make-up of populations to the many interactions between such factors, and beyond that to measuring the effects of efforts to control and evaluate both medical care systems and individual treatments (Susser 1978).

The distribution of a health condition in particular populations describes the frequency of its disposition among factors of potential relevance or interest. The crux of this definition is its significance for descriptive studies. In the words of an old saw in epidemiology, distributions tell where, when, and among whom given health conditions are to be found. Such studies constitute a starting point. For conditions of known genesis, they point to appropriate

action. For those of unknown genesis, they point to the need for further investigation and research.

A determinant can be any factor, whether event, characteristic, condition, or other definable entity, so long as it brings about change for better or worse in the occurrence or distribution of a health condition. Epidemiology as a discipline has thus evolved in intimate linkage with the needs of public health. The discipline is an essential scientific base for securing the health of societies, and thus demands of epidemiologists an understanding of causality. This is not to say that epidemiological methods cannot be studied for their own sake nor applied to population problems other than those of health. These latter purposes are far from the mainstream, however. In the ancient as well as the modern world and in populations large or small, advances in the study of health disorders stemmed almost always from need in the face of disordered health.

Many readers unfamiliar with the discipline will attach to epidemiology the connotations that arise from the narrow everyday usage of the word "epidemic," namely, an explosive outbreak of an infectious disease. This usage was entrenched in the late nineteenth century, an emphasis contingent on the then newly developed germ theory. Of course, epidemics existed long before the germ theory was invented to explain them, and so too did epidemiologists, and the term is not confined to transmissible infection. A number of books of the Hippocratic corpus, compiled as early as 400 BCE, bear the title Epidemics (see Hippocrates, Chadwick, and Mann 1950).

The Sydenham Society's Lexicon of Medicine and Allied Sciences (1870–1900), cited by the Oxford English Dictionary, defined epidemic (adjective) as: "Prevalent among a people or a community at a special time and produced by some special causes not generally present in the affected locality." Also given are some early metaphorical usages, unrelated to infectious disease and more appropriate to chronic and psychiatric disease: from Sir Walter Scott, "The epidemic terror of an imaginary danger"; in another figurative use from Edmund Burke, "An epidemic of despair"; from John Milton, "Epidemick Whordom."

The definition of the Sydenham Society was an apt one for its time. It indicated that the essence of an epidemic is a cluster of events: the events occur more frequently than they would by chance in several dimensions. Clustering conveys broadly the dimensions of modern usage in epidemiology. An epidemic, we may say, is a cluster of similar episodes of illness circumscribed in the simple long-standing formulation of time, place, and person. More precisely, we describe these three dimensions in terms of time period, spatial distribution, and affected population. This definition restores generality to the term epidemic. Narrowly constructed terms for when, where, and who, sum it up. It should be noted, however, that room is left

for looseness or maneuver in the designation of a cluster. Excepting mass occurrence, the character of the distribution of the disorder might be elicited, and the epidemic designation therefore assigned, only after close study and refined statistical test.

The substance of epidemiology is study of the occurrence, cause, and control of health disorders and illness. The means for studying their antonym, good health, are little advanced. The World Health Organization defines health as a state of physical, psychological, and social well-being. Such states have not been made accessible to reliable measurement and enumeration. The bounds of good health remain blurred, the signs of good health equivocal. Productive epidemiologic research demands firmly outlined definition of what is under study. The means for definition come closer to attainment as we discover molecular and genetic markers that help presage future risks. For the present, however, we are limited to describing and measuring health largely in terms of the absence of signs and precursors of ill health.

Thus we can generally speak with authority only of the obverse of good health. Disease, illness, and sickness convey dynamic states of ill health; impairment, disability, and handicap describe conditions that tend to be static. The distinctions made by these two sets of terms are noted for the sake of precision in defining the field of interest of the discipline; they also ensure that we compare like with like. Disease, illness, and sickness are not synonymous. They constitute a hierarchy of levels of organization. Disease is best thought of as a process that creates a state of physiological and psychological dysfunction confined to the individual. Illness is best thought of as a subjective state, a psychological awareness of dysfunction, also confined to the individual. Sickness is best thought of as a state of social dysfunction, a social role assumed by the individual that is defined by the expectations and response of society and that thereby affects the state of relations with others.

Impairment, disability, and handicap are not synonyms either (Susser and Watson 1971). They constitute an analogous hierarchy of levels. They refer to established, stable, and persisting states rather than to unstable evolving processes. Impairment refers to a persisting physical or psychological defect in the individual. Disability refers to persisting physical or psychological dysfunction, also confined to the individual. Handicap, like sickness, refers to persisting social dysfunction, a social role assumed by the impaired and disabled individual that is also defined by the expectations and response of society (Nagi 1965; World Health Organization 2001).

Epidemiology describes the occurrence and evolution of these disordered states of health and seeks to discover their causes and prevent them. For these purposes enumeration is an essential first step. The study of the relation of these states of disordered health to persons, society and habitat follows as

the logical next step. To establish these relationships, we must discover the distribution of the enumerated disorders, that is, the way they vary in the population and also the personal attributes and circumstances with which they are associated.

In order to discover the connections between the phenomena of health disorders, and particularly their causes and effects, our knowledge in medicine and public health is built from studies of subjects in four broad classes: *molecular and genetic* studies, the most refined and specific class; *clinical case studies*, using clinical observation and investigation in individuals with aid from laboratory data, individually or in series; *experiment*, through all levels from the inanimate to the animate, including human beings; and *population studies*, including the physical, biological, and social environment and using a variety of methods including ecological studies. (Such studies are by their nature complex and we defer discussion to later chapters.)

In the present day, extraordinary advances in molecular biology and genetics are so close and pervasive that all will be familiar with their outlines. From the epidemiological perspective, the particular importance of these advances is first in specifying the particular character—with an exactitude previously almost unimaginable—of pathological change in health disorders in individuals, and the presence or absence of markers in populations that indicate persons at risk, immunities and silent spread. The capacity for close definition opens the way to clarity and precision in the core epidemiological tasks: the description of the distributions of health disorders will be greatly refined; the determination of the precise locations and interactions underlying the disorders under observation will be much facilitated. In essence, this new knowledge brings to epidemiology the potential for a once unattainable degree of specificity and precision in defining states of health and disease processes.

Since the center of the medical frame of reference has been the case of the sick person, the case study was long at the heart of clinical understanding. In the case study, the observer collates and selects, in one or several cases, his direct observations. The clinician observes pathological process in relation to the symptoms and signs it yields in terms of anatomical structure and physiological function. Clues as to what is amiss depend often on what is unique or unusual as well on the recurring patterns of a given disease. Thomas Sydenham in the seventeenth century (1624–1689) first developed an approach that searched for common patterns in a systematic series.

A large qualitative element is inherent in any observation of patterns and even in the recognition of what is unique. It has been practical to devise and use reliable measures for a few only of the multitude of individual observations that the clinician weaves into integrated analyses. Summary measures and classifications abstracted from the whole may reflect only limited

aspects of the clinical picture. The same holds in even greater degree for all the observations that the social observer synthesizes in the analysis of the communal case, be it tribal village, city or suburb, factory floor, psychiatric ward, or hospital organization. Thus the advantages of intensive observation in the case study are balanced by the potential disadvantages of subjective observation. Wide variation characterizes individual cases: limitations of measurement can thus be severe both in general, and in the environmental and other dimensions antecedent to an outcome in particular.

Laboratory studies advance largely by means of experiment. Experiment defines a method of research by means of the prospective measurement of planned intervention. Galileo Galilei (1564–1642), in his explorations of natural laws, is commonly credited with originating systematic scientific experiment. Claims might also be made for isolated examples such as Archimedes in classical Greece and others. In the laboratory approach, finely tuned and specific intervention enables investigators to achieve precision in their measures of outcome. Rigorous control of the operations greatly strengthens interpretation of results. With the shift in focus from chemistry and physics to the study of increasingly complex living organisms, however, difficulties arise. The ever-changing interactions of living organisms and their capacity even for self-motivation progressively reduce specificity, precision, and certainty in interpretation.

The classic laboratory experiment, in the majority of instances, must be confined to animals other than human. The hope is either to simulate the human condition or to reach sound generalizations across species. Many reactions are species-specific; the same stimuli may give different results in different species. Generalization from animal experiment must thus be cautious, and the more so the more complex are the levels of function under study.

Psychological experiments are more beset with hazard in extrapolation to humans than are physiological experiments, and social experiments more so than psychological. Applied to human beings, the experimental method must still deal with severe limitations of range. However rigorous the design, clinical and prophylactic trials with human subjects cannot match the control attainable with laboratory animals. Ronald Fisher (1890–1962), in devising his agricultural studies, greatly improved the application of scientific design and statistical analysis in general, but in particular outside the laboratory (Box 1978). For epidemiologists, perhaps his most important innovations were in the design of experiment—notably the randomized controlled trial; in statistical analysis—notably analysis of variance (an original usage of a word now taken for granted in statistics and genetics), and multivariate statistical techniques. He was very much aware of the difficulties in arriving at undeniable truth, and maintained an analytical stance that insisted

on a doctrine of "rigorous uncertainty." The randomized trial allowed the researcher to estimate a numerical value for the degree of error that might inhere in a given study of specified size.

To all this, one may add the idea of polygenic inheritance. This approach to genetics allowed several factors to explain greater or lesser proportions of variance in a given effect. At the beginning of the twentieth century, protagonists of individual and population genetic theories had been bitterly at odds. On the one side they espoused Mendel's mid-nineteenth-century single gene theory; on the other, the later biometric approaches of the eugenicist Francis Galton (1822–1911), and his protégé Karl Pearson (1857–1936). In a paper of 1918, Fisher (1890–1962) resolved this disputation. His mathematically based theory reconciled Mendel's then dominant single gene theory (rediscovered at the beginning of the century) with the new Pearsonian biometrics (Fisher 1918; Gillham 2001). Later in the 1940s at the London School of Hygiene, Austin Bradford Hill contributed further to epidemiological rigor and design in his research and teaching. At that time, Hill was on friendly terms with the irascible Fisher. In epidemiology and medical research on human subjects, Hill was the major protagonist of randomization. Perhaps his greatest contribution was to apply, in human studies, Fisher's idea of controlling experimental and analytic bias by the "blind" and random allocation of subjects to exposed and unexposed groups. Hill's simple and masterly prose describing the general principles underlying epidemiological research in human populations counts as a major contribution in itself (Hill 1955).

Population studies are at the heart of epidemiology. They deploy the methods mainly of observation and, much less frequently, of experiment. The case study is in essence confined only to numerators. The population study gives numerators meaning by relating cases, as best we can, to the population from which they are drawn. This procedure creates a standard of comparison for the analysis of cases, without which observers can reach no conclusion about the abnormality or distinctiveness of any phenomenon. In a general way, epidemiology shares this procedure with other sciences that study populations, for instance, biology of all species, population genetics, and such social sciences as demography, economics, and sociology.

These population disciplines differ in the dependent variable that is the particular object of study. A variable is an abstract term useful for the devices of measurement. A *dependent variable* represents an effect or outcome, an *independent variable* a determinant or cause. Among the units under observation—in epidemiology units are commonly persons, or less often, groups—any variable represents a measure of a particular property. For instance, among persons such properties are age, weight, or health state; among groups, they are such characteristics as average income, social class, and education, or the populations of geographic areas defined by latitude,

altitude, or population density. Thus a variable represents a measure of a property specified, defined, and observed. It covers in its span the whole range of the units under study. If the variable is linear and quantifiable, it may be described by the differing numerical values on a continuous scale of a measurable characteristic. If the variable is integral, that is, an indivisible whole that is not quantifiable, it may be described by the differing qualities of some specified entity, say persons, habitats, or types of automation.

In essence, what population disciplines share and often exchange among themselves are the methods for field observations and interventions among populations, especially their enumeration, measurement, and analysis. In saying that these elaborations of method are the core of the methods of epidemiology, we affirm its common ground with the other disciplines involved in the study of society. States of health do not exist in a vacuum apart from people. People form societies, and any adequate study of the attributes of people is also in some degree a study of the manifestations of the form, the structure, and the processes of social forces.

However, epidemiology is segregated from other studies of society in the common ground it shares with other medical sciences by its choice of dependent variables, namely health and its disorders. It differs from other medical sciences, however, in that the standard unit of study is a group or population and not an individual. These units are most often chosen in terms of some criterion to suit specific research questions directed at states of health and their causes at the individual level. The criteria of choice are thus to some degree arbitrary; they can be, say, age, gender, or ethnic group, or a health disorder among patients, or a community defined by geographic or social borders.

Epidemiology, unlike the strictly clinical sciences, ranges beyond factors impinging directly on individuals to examine also those acting at higher levels of organization. Thus epidemiologists may seek out the effects on health states of memberships in social groups of all different kinds, or of the physical or the biological environment. The criterion for the choice of groups to be compared may then be in terms of social organization, as with communities defined in one of several ways, or of census tracts geographically defined, or of particular physical environments and, not least, with populations at risk of actual or potential epidemic spread.

With such units of observation, epidemiology brings into medical thought new levels of organization ranging beyond the typical clinical focus. The level of organization of populations entails the grouping of individuals. In whatever way they are assembled, from simple aggregates through a hierarchy of increasing complexity, grouping introduces sets of variables with new dimensions over and above those germane to the study of an individual. These variables are complex and numerous. Their introduction creates a

body of knowledge about states of health different from that confined to the clinic, the bedside or the laboratory, knowledge otherwise unavailable to studies of disease and health. The higher the level of organization, however, the more challenging are the tasks of definition, analysis, and explanation.

The study of disease among individuals alone as the unit of observation can of course shed light on the nature of the disordered state of functioning and its progress through time. Yet even in a series of affected cases, the study of individuals alone cannot ordinarily determine the limits of the disorder in relation to normality, nor can it securely predict its onset, progress, and outcome. The garnering of such knowledge calls for the numeracy inherent in epidemiological method. In the most common form of the method, epidemiologists have studied populations to discover characteristics and causes of disorders as these exist or appear across individuals in defined populations. Such populations comprise individuals assembled and aggregated by the definition of the researcher. The criteria are multifarious, for instance, occupation, or age-group, or sex, or location, or exposure to a suspected hazard, or the occurrence of a particular disorder, or several of these together. The basic requirement is that study populations constitute a defined group that can be enumerated; the terms of study include whatever the researcher chooses to examine that is within ethical limits.

Epidemiological method aims to gather data both sufficient in number, and assembled in a manner to permit meaningful comparison of relevant variables describing the units under study. Only such controlled comparisons make it possible to comprehend all the aspects of a disease process at different points in time. In the ideal, the spectrum to be comprehended is the life-course in whole or in part. It begins with antecedent factors that cause disease, moves back in time to the precursors that allow us to predict and prevent disease, and passes forward through the full span of clinical manifestations, course, and eventual outcome.

The ultimate aspiration of epidemiology, it could be said, is to understand and explain the health states of populations as they unfold over the life course and through history. Aspirations so ambitious may well seem grandiose to many researchers in biological fields and even, indeed, to epidemiologists. In truth, the immediate objectives of epidemiology can only be met piecemeal. Less grandly, then, its fundamental, immediate, and sufficient aim reduces to the elucidation of the causes and course of health disorders. In other words, epidemiology is the study in populations of the relation on the one hand between varying states of health considered as *effects* and, on the other hand, the possible *causes* influencing those states (not excluding purposeful medical and public health intervention). Of necessity, the epidemiological method is selective. Studies are conceived in terms of selected populations over a selected slice of time, whether periods of historical time or of ages

over the life-course. Whatever research design is used, in Olli Miettinen's term, this is the *study-base* (Miettinen 1968). The researcher approaches the task by seeking out the relations, in and across populations, between defined variables in numerical terms.

Observational studies often rely on the historical information that can be elicited and collated from existing sources (not only records, but people who can say how they feel or what happened). They may also create new sources in cross-sectional, retrospective, or forward-looking studies of various designs. Suitable mechanisms must be set up to collect the information. In a broad sense, observational studies can be made to include the secondary analysis of existing data previously collected for purposes other than the study in hand. To resort to such sources is nearly always an economy. For instance, the data may be culled from existing registers of vital statistics, from hospital records, and from censuses. They lose something in that the definitions of what is measured are not tailored to the specific objects of a given study. More commonly, observational study involves sorties designed in the form of field studies. These studies collect data within health care systems or among defined groups in the population at large and do so in a planned fashion designed for a specified object of study.

This brief introduction of the scope of epidemiology lays the ground in what follows, for addressing the developments in concepts.

References

Box JF (1978). *R.A. Fisher: The Life of a Scientist*. New York: Wiley.

Fisher RA (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Philos Trans R Soc Edinburgh* 52:399–433.

Gillham NW (2001). *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*. New York, NY: Oxford University Press.

Hill AB (1955). *Principles of Medical Statistics*. 6th ed. London: The Lancet.

Hippocrates (1950). Epidemics, Books I & III. In Chadwick J and Mann WN, transl. *The Medical Works of Hippocrates*. Oxford, England: Blackwell Scientific Publications.

Miettinen OS (1968). The matched pairs design in the case of all-or-none responses. *Biometrics* 24:339–352.

Nagi S (1965). Some conceptual issues in disability and rehabilitation. In Sussman, ed. *Sociology and Rehabilitation*. Washington: American Sociological Association.

Susser M (1978). Epidemiology. In Kruskal WH and Tanur J, eds. *International Encyclopedia of Statistics*, Vol I. New York: The Free Press and London: Collier MacMillan.

Susser M, Watson W (1971). *Sociology in Medicine*. 2nd ed. London: Oxford University Press.

World Health Organization (2001). *International Statistical Classification of Functioning, Disability and Health*.

2

The Relation of Concepts to Causes in Epidemiology

Like any other discipline, epidemiology did not come to be what it is through some set of academic and scientific rules that ordain its form and content. The multifarious areas of study ensconced in modern universities, an ever growing list, have varying histories. Beliefs have informed theories, theories guide investigation, and the results bolster or revise interpretation. Once entrenched and institutionalized, disciplines acquire a life of their own. They achieve a formal structure in the form of independent academic departments, a common culture of attitudes, morality, and behavior, and self-propagating doctrines that make for great persistence. Within these constraints, individual teachers and writers may promote or challenge reigning concepts more or less successfully. All this is guided and constrained by the economic, social, and political imperatives of the times.

The persistence of ideas thus engendered characterizes not only philosophies but the histories of the health sciences. Witness the long influence of the teachings of Galen, who wrote in Greek some 2000 years ago and became prominent in Rome. He was himself guided by the then 500-year-old Hippocratic theory of humors. Much of his doctrine pervaded medicine in Europe through 1500 years or more. Interpretation of the ancient Hippocratic corpus is hazardous even for meticulous scholars. The meaning of the original Greek text seems to be more accessible in some Hippocratic books than

in others. Galen did much to bring a degree of order out of the texts. They are often obscure and vary from one book to another among the several volumes. The translated prose takes account of various viewpoints. While plain and declarative, scholars have found that the translation can render the interconnections between one sentence and the next obscure; it can also arrive at conclusions that defy the understanding, not least, of these writers themselves.

A brief sample of one cautious modern translation of Hippocrates serves to illustrate:

> "The circumstances of the diseases, from which we used to form our judgment and to differentiate, learning from the common nature of all [persons and things] and from the individual nature of each [person and thing]; from the disease, from the patient, from the things administered, from the one who administers, (because this too contributes to an easier or more difficult development), from the state of the weather and of each country, both universally and with respect to the parts; from the custom, regimen, occupational activities, and age of each [patient]; through utterances, manners, silence, thoughts, sleep, insomnia, dreams—what kind? when?. . . . ." and so on for several more lines, until at last the denouement: "Starting from these things one has also to consider what happens under their effect." (Translation taken from Langholf 1990)

Ancient China provides a similar case of complex concepts, some earlier than the Greeks and some parallel in time. In Chinese texts, ideas about the causes of disease have been traced as far back as 1000 BCE. Some scholars consider that those Chinese theories of causation were certainly as cogent as those of Hippocratic writers centuries later (if to the modern eye no less flawed) (Unschuld 1985). Joseph Needham was a distinguished twentieth-century biologist who in his later years made a study of Chinese science. He found the Chinese to be far in advance of science in Europe at least until medieval times. By the time of the European Renaissance and the accompanying scientific efflorescence, however, scientific development in China had stalled and begun to wither. It did so in parallel with the tightening structure and formalizing society of that grand imperium (Needham 1956).

Some of those early theories are reported still to have currency in the contemporary practice of traditional Chinese medicine. No less remarkably, in the popular culture of the West of our own times similar survivals from the past persist. They are to be found especially among some of the many forms of complementary and alternative (CAM) medicine. A recent study in the United States reported that more than one third of respondents had recently used such a therapy (Institute of Medicine et al. 2005; Kronenberg et al. 2006). Doubtless the sick or discomfited draw some comfort and support thereby. Sometimes, one must recognize that traditionally derived remedies

can offer useful therapy. Skeptical one must be, but one needs also to remember the stories of digitalis, quinine, and a few other effective drugs that were adapted from folk remedies.

Through the works of Hippocrates in Greece and then Galen in Rome, the reign of Aristotle's founding doctrines persisted in the science of Europe beyond the following millennium and more (Singer 1979). Their persistence had owed much to the governance of the institutions of church and state. The flux of conquest and destruction ended the ideological dominance of the ancient civilizations of Greece, Rome, and Byzantium. First the nomadic tribes came from the steppes, followed by Arabs of the Middle East inspired by their prophet Mohammed. In Europe, there followed a long dark interval little troubled by scientific thought or practice. Greek scholars and texts found refuge in Alexandria in Egypt.

During the Islamic Golden Age, roughly from the eighth to the thirteenth centuries, however, Arabic civilization acquired and developed those same Alexandrian texts and passed them on to Italy and the rest of Europe. Thereafter, a mercantile Western Europe at last caught up. In Italy in the fourteenth century, European Renaissance scholars began to build knowledge and understanding. At first adopting the founding sciences of Greek, Roman, and the Islamic civilizations' additions, in the end they surpassed them.

In modern times cumulative innovation has led to once inconceivable growth in knowledge. The rate of advance and growth of science and technology in the twentieth century was greater than exponential. If the standard is taken as the time that indicators of change took to double, then for a wide range of indicators, doubling time became shorter and shorter. This acceleration held true for the number of scientific publications, including books and journals, as well as the abstracts that were invented to cope with growing numbers of books and journals; it is true for the numbers of scientists and scientific organizations, more numerous now than their sum in all history; it is true for the use of energy generators and computers, and for technical developments. Even though, in estimating the rate of change, we might separate quantity from quality of publications, technical from conceptual advances, marginal from fundamental, each of these are to some degree interdependent.

Some students of science as a social institution have predicted that such a rate of growth must reach a ceiling or that at the least there must be a slowing of its pace. The advent of slowing will bring no foreseeable reduction in the problems of change. As long as our civilization endures, scientists will still have to cope with a stunning flux of new knowledge. Change of this order requires that we learn both to put aside theories and even knowledge that has become useless or inept, and to use fresh knowledge that is apt and useful in the new contexts.

Concepts in Search of Fact

To acquire new facts and put old ones into storage might seem simple. To acquire new concepts and discard old ones is more difficult. In science generally, as in other areas of study, concepts play a major role—a role no less than facts, but different in determining modes of thought. Concepts are the organizing principles by which we bring our facts together and weave them into a coherent whole. Concepts thus constrain our unthinking explanations of the phenomena daily surrounding us. Not least, they guide the calculated and particular explanations we seek for the phenomena we choose as scientists to observe (Susser 1973).

A striking exposition of the role of concepts was advanced by Thomas Kuhn (1962). Kuhn recognized a significant antecedent of his theoretical position, which has since gained prominence and currency in the philosophy of science. This was the work of Ludwik Fleck (1935), a Polish venereologist.[1] In the 1930s, for his exposition of the process of scientific discovery, Fleck drew on the history of syphilis from its post-Columbus European debut in the 1490s and into the twentieth century. His was no mundane narrative history. Its underlying object was to advance a seminal new theory. Fleck's history of this single disease was directed precisely to advancing the idea that, in any era, contemporary perceptions of fact follow successive changes in theoretical concepts, which are also informed by sociocultural conditions. Fleck argued that the governing concepts of science, or in his words, "styles of thought," constitute a thought collective among scientists. In successive eras of thought and practice, revised theory displaces existing theory and reinterprets the facts in the light of the new theory. Fleck drew on the history of the medical approach to syphilis, although merely as the vehicle for documenting the changes through the centuries. In each historic period, what was perceived as fact about syphilis had evolved and changed. Each shift in the perceived facts, Fleck argued, arose from and accompanied a change in the ruling concepts of disease.

Thomas Kuhn developed Fleck's insight, then still hidden in obscurity, into a fuller general theory of scientific change (Kuhn 1962). This work was highly influential. In daily usage, Kuhn argued, scientists practice *normal*

[1] Though first published in German in 1935, Fleck's monograph did not appear in English until 1979. Kuhn acknowledged Fleck both in the preface of his own book as well as in the foreword to the English edition of *Genesis and Development of a Scientific Fact* (1962), though Fleck is not actually cited in Kuhn's book. Fleck's life as a physician and scientist was interrupted by World War II; he and his family were deported to Auschwitz and later to Buchenwald concentration camps. After the war he lived in Poland and then Israel, where he died in 1961.

science. Normal science is governed by the generally accepted concepts and theory that prevail in a given era. It is reinforced by standard texts with a standard perception of the history of facts as well as the social institutions, culture and beliefs engendered or acquired by a mature science. In normal science, regular advances occur in steps that conform to the ruling theory. Normal science, in the view Kuhn derived from the proto-perspective of Fleck, is overturned only by the advent of revolutionary theory that provides a new and convincing conceptual base for the pursuit of explanation of an evolving set of facts. Kuhn's theory provoked hot debate among scientists and philosophers. Not a few, secure in a belief of the impeccable objectivity of science, were averse or hostile to the premise that social elements extraneous to science could influence thought and discovery.

One may note here that in epidemiology major changes in epidemiological concepts over the past two centuries conform well to the revolutionary concept. Thus, in later pages of this text, in the nineteenth-century scientific revolution is well-illustrated by the ultimate displacement, as a theory of disease, of the centuries-old idea of miasma by the germ theory of infection. Other equally significant changes, however, have attended a slower process of attrition rather than revolution, as when during the mid-twentieth century the germ theory of specific infection was displaced by a broader theory of multiple causes, which in turn explained noninfectious chronic disease. We consider later the large extent to which the course of modern epidemiology conforms to these ideas, in particular the role of revolution or attrition in such conceptual shifts. One may speculate that revolution is a consequence of new knowledge, which is the outcome of definitive discovery, as is characteristic of the advances of physics and the so-called hard sciences. Attrition is perhaps more common in the relatively fluid human sciences, which in medicine must draw on human biology, epidemiology, the clinical sciences, and much else.

To illustrate, with Kuhn we turn briefly to the philosophy of science and the scientific history of cosmology. Familiar examples illustrate the historical succession of concepts of the universe and the relation of theory to practice. In the first millennium CE and the first half of the second, the cosmos was explained according to the theory of Ptolemy of Alexandria (Claudius Ptolemaius) who flourished roughly in the period 140–160 CE: Ptolemy himself followed the theory of Aristotle in the fourth century BCE. This theory, not an unrealistic one, held mistakenly that the earth was fixed at the center of the universe, that the sun and all heavenly bodies revolved around it and, being perfect, must move in perfect curves, namely circles. Ptolemy's explanation worked well for the purposes of that time and for many later generations. Thus the theory enabled Greek and Egyptian sailors to chart sea routes and navigate the open seas. Their needs were adequately accommodated by the central principles of the theory.

The revolutionary cosmic system of Nicolaus Copernicus in Poland published in 1543, however, placed the sun at the center of our universe. The new theory explained more mysteries than the Ptolemaic and eventually supplanted it. Half a century later, in the face of the new realities of the heavens discovered by Galileo Galilei, the new Copernican theory proved better able to withstand the test. Galileo constructed a telescope, which he was the first to apply to a wonderful new world of stellar observation. He reinforced his findings by his mathematical analysis of those first direct observations of the stars and their movements. As everyone knows, however, he failed to do so to the satisfaction of all the important arbiters within the papal ambit, and was humiliated accordingly by the papal court.

Isaac Newton followed with his theory of gravitation. In our own time, the case of Newton's long-standing and seemingly unassailable theory can be seen to resemble that of the Ptolemaic theory. The gravitational theory fitted near perfectly those facts of the cosmos known through the nineteenth century. Hence it produced accurate predictions of the movement in the stellar universe and still does. Yet early in the twentieth century we learned, and are still learning, that the then young Einstein's *theory of relativity* better explained the cosmos. Indeed, starting from the Einsteinian foundation, today's cosmologists have moved on still further to a theory of the origin of an ever-expanding universe originating in a Big Bang.

The key object lesson of this strictly amateur potted history of cosmic theory is that it relies on a general historical model of the way new facts have followed new concepts. Each such concept changed the perception of what facts were relevant or germane. With the qualifications just noted, epidemiology too is replete with shifts in theory and their consequences. Early in nineteenth-century Britain, for example, the seventeenth-century theory of miasma had gained purchase and risen to dominance both in medicine and in the nascent public health movement. One must note, however, that the theory did not go unchallenged by contagionists or by those who saw poverty as critical to health (Rosen 1958, 1993).

Thomas Sydenham (1624–1689) had forcefully advanced the miasma theory. Dubbed "the English Hippocrates," he earned the appellation in the light of the derivation of a humoral theory, which governed both individual predisposition and its extension to the atmospheric conditions that precipitated epidemic disease. In the main, perhaps, his repute rested on his excellence and fame as a clinician. Sydenham achieved all this despite the revolutionary turmoil of the times, during which he followed Oliver Cromwell in the civil war that deposed King Charles the First. Subsequently, with the restoration of the crown under Charles the Second, Sydenham lost political favor, but seemingly this did little harm to his repute and practice as a physician (Figure 2.1).

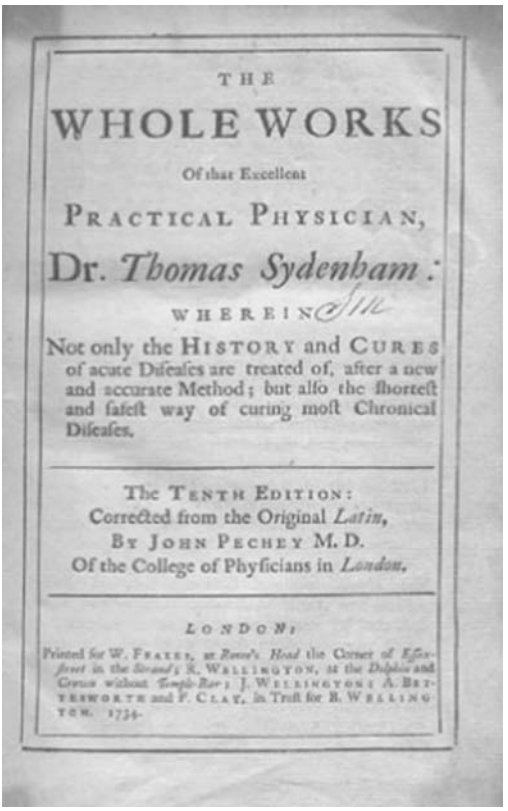

THE
WHOLE WORKS
Of that Excellent
PRACTICAL PHYSICIAN,
Dr. *Thomas Sydenham*:
WHEREIN
Not only the HISTORY and CURES of acute Diseases are treated of, after a new and accurate Method; but also the shortest and safest way of curing most Chronical Diseases.
The TENTH EDITION:
Corrected from the Original *Latin*,
By JOHN PECHEY M. D.
Of the College of Physicians in *London*.
LONDON:
Printed for W. FRASER, at *Rose's Head* the Corner of *Essex-street* in the *Strand*; R. WELLINGTON, at the *Dolphin* and *Crown* without *Temple-Bar*; J. WELLINGTON; A. BETTESWORTH and F. CLAY, in Trust for B. WELLINGTON. 1734.

Figure 2.1 Front cover of an edition of Thomas Sydenham's writings. This one was published in 1734.

Sydenham's clinical approach was to assemble detailed observations of the features held in common by each particular type of illness. To accomplish this aim, he gave great emphasis to direct observation of the natural course of disease. From his systematic notes on the cases of the many patients he attended, he defined for the first time the characteristics of several disease entities. These included the eponymous choreiform tremor of rheumatic fever—*Sydenham's chorea*—dysentery, measles, scarlatina, and the fever of malaria (literally, from the Latin, bad air). For the treatment of fever, he advocated in England the native Amazonian remedy of cinchona bark (the source of quinine). Among much else, he gave a classic description of gout, attributing it to humoral imbalance in the same sense as that used, one may reasonably assume, by Hippocrates and Galen (Greenwood 1935; Riley 1987).

Excepting much precise clinical description, however, readers should not be under the illusion that Sydenham's theoretical writings are readily

understood in a modern sense**.** His thinking was not static, but evolved as he practiced, observed, and learned. In the early twentieth century, his ardent admirer Sir William Hamer was, like William Farr in the nineteenth century, among early analysts of epidemic patterns. In Hamer's own book on epidemics, he too underlines the difficulties in following Sydenham's theories and grasping their full compass (Hamer 1928). Beyond this, and more relevant to the theme of this text, Sydenham searched also for identifying patterns of specific epidemic disorders. Acute febrile illness, he held, could be assigned to two main classes: *Intercurrent illnesses* such as scarlatina, pleurisy and rheumatism were supposedly owed to a combination of host susceptibility and atmospheric constitution. *Epidemic distempers*, diseases such as plague, smallpox, and dysentery, were owed to poisonous emanations from the soil (miasma) arising also with atmospheric change (which Sydenham labeled the *epidemic constitution*). To complete his classification he added still another class of epidemic disorder, namely, *stationary fevers* which manifested only in the ripe conditions of particular years.

Richard Mead (1673–1754) was an eminent physician of his time and renowned for his wide scholarship. In setting out his own views on the source of epidemic diseases, a half century after Sydenham, specifically plague, smallpox, and measles, Sydenham hardly gains a mention. Unlike Sydenham, Mead recognized the specificity of each of these conditions. Also, unlike many physicians both before and after his time, he was quite emphatic about the role of "contagion," responding to outbreaks of plague in the south of France by publishing his *Short Discourse*, which went through nine editions (Mead 1720; Zuckerman 2004, p. 285). Mead recognized person-to-person transmission and extended the concept beyond that: "contagion is propagated by three causes" he wrote: "the air, Diseased Persons and Goods transported from Infected Places" (Mead 1720, Part 1, p. 2).

However, this explicit anti-miasmatist, contagionist position apparently did not carry a theoretical message to his contemporaries, as did Sydenham's. Thus Sydenham's ideas of miasma connecting disease to dangerous emanations of climate, weather, and environment were adopted, developed, and expressed in varying form by several seventeenth-century figures in Britain. Among them were John Arbuthnot, a physician, author, and wit, John Locke (1632–1704), a major philosopher noted mostly for his *Essay on Human Understanding* (1690), and Robert Boyle, 14th child of the Earl of Cork and a medical practitioner, physicist, and chemist fertile in ideas. Coincidentally, mortality in Western and Central Europe began a steady decline that has not abated since (except for a brief rise in Britain in the early years of the Industrial Revolution). For centuries, smallpox had been one of the most fatal and terrifying epidemic disease. Beginning about the mid-eighteenth century, immunity induced by the variolation advocated by Lady Mary Wortley Montagu,

lately returned from the British embassy in Turkey, began to contribute to the decline, and by the mid-nineteenth century William Jenner's discovery of the effects of vaccination began to have major effects. In Europe, earlier Renaissance campaigns had promoted cleanliness in the urban environment, and in all likelihood reinforced the undoubted effects of vaccination on mortality. In the time of Sydenham, consciousness about the importance of cleanliness for health grew concurrently (Figure 2.2) (Riley 1987).

The concept of miasma prevailed through the whole of the nineteenth century. At the turn of that century, the theory was still alive if less well. Given the cast of these theories and the meager development of the science of epidemics of the time, a not illogical line of preventive investigation was to postulate and seek out both the sources and ill effects attributable to miasma. At least one such remedy seemed to be within reach. A logical step toward both proof and prevention was to eliminate the suspect sources of miasma in slums and poor sanitation. The modern public health movement arose in the early nineteenth century in France, with England following two or three decades later. Miasma theory guided the initial approach. Indeed, Edwin Chadwick and Florence Nightingale, founders of the Sanitary Movement earlier in the nineteenth century, were proponents of miasma theory to the end of their long lives (in 1890 and 1910 respectively). Even in the later nineteenth century, such influential and canny English proponents of public health as William

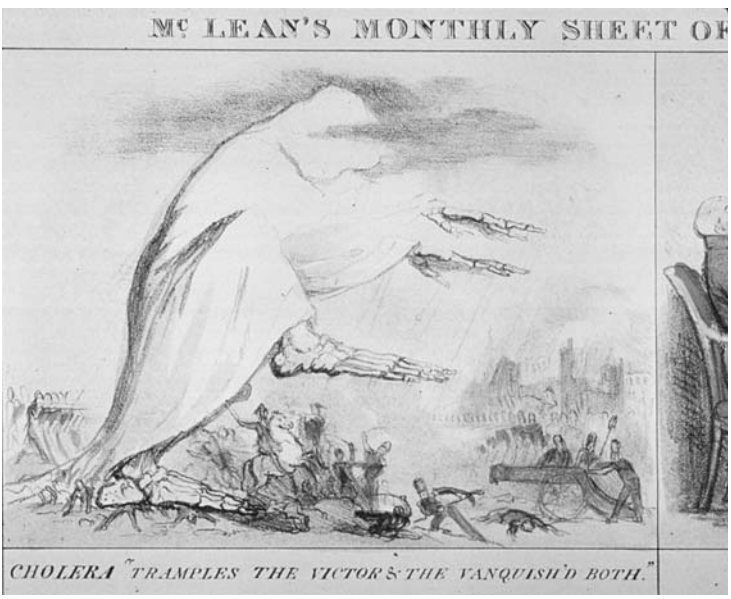


Figure 2.2 Cholera depicted as an indiscriminate miasma in McLean's (1831). Credit: Courtesy of the National Library of Medicine.

Farr, John Simon, and their several medical allies had adhered to miasma theory for the greater part of their early careers (see Chapter 6).

In England, 1831 was a time of a rising cholera epidemic imported into Europe from the East. The idea of *contagion* was once more gaining cogency and currency alongside that of miasma. At that time, "contagion" spoke specifically of transmission either by direct contact or by fomites. A third form of transmission was then characterized as *infection,* in the then narrow sense of the *inhalation* of so-called viral spread through the air. In 1530 in Italy the many-sided poet and physician Girolamo Fracastoro, in his long poem about the shepherd Syphilis, had described sexual transmission by contagion of the so-called French pox; the venereal disease was the terrible fate visited on him by the gods for the transgression of a forbidden love (Fracastoro 1530 and Eatough 1984).

Apart from occasional mention in various writings exemplified by Jacob Henle in Berlin, only in the 1840s did ideas of specific infection began to gain a foothold, if a slippery one, in then current medical thought. As we shall describe later, in that decade a few medical men—the English contemporaries and nascent epidemiologists John Snow and William Budd prominent among them—began to assemble compelling evidence that infectious agents harbored by affected persons could be excreted in the feces and transmitted to others by the route of ingestion. In the same decade, Ignác Semmelweis likewise arrived at the similar conclusion that pathologists and obstetricians, having dissected the corpses of women killed by puerperal fever, were liable to transmitting the fatal infection to patients whom they then attended. As described in later pages, in different ways each of the three set about demonstrating that fact.

For compelling proof in counterargument, anti-contagionists cited a Warsaw physician who in 1831 was said to have ingested material from cholera stools; supposedly, he had done so without apparent ill-effect. A similar experiment was repeated late in the century by the eminent German public health authority Max von Pettenkofer (as we shall describe later in explaining these tests); on the face of it dismissive of the theory of microbial infection by ingestion. Yet resistance to the germ theory was not simply blockheadedness. Through the nineteenth century, much remained inexplicable by the science of the time. For example, without awareness of carrier status, how can one explain occasional sporadic and even consecutive outbreaks of childbed fever? Loudon describes how one *accoucheur*, although scrubbed and aproned, spread the infection to case after case (Loudon 1998). The theory of specific cause, in the form of microbial infection, had to wait for the twentieth century to achieve full acceptance and to totally displace miasma theory. The germ theory in turn sustained dominance for a half-century before the specific risk factor theory of multiple causes supplemented and

to a degree supplanted it. In turn, risk factor theory withstood challenge until the twenty-first century loomed, when it began to be supplemented by a broader multidimensional perspective, which we refer to as eco-epidemiology. Below we discuss the circumstances of succeeding developments in more detail.

In epidemiology as in other sciences, the place of guiding concepts and theory is to enable predictions to be tested and choices to be made between paths of action. Yet in any given period it remains possible for different theories to guide responses that are seemingly equally successful under the same circumstances. One theory is better than another only when it is faced with more specific tests and passes them. Displacement of a theory requires specific tests based on narrow *a priori* assumptions, which the less effective theory cannot meet. Sanitary reform could be justified both by the miasma theory and the germ theory. Although sanitation alone is able to prevent the transmission of infectious disease, it does so only for the one broad class of ingested infections. To prevent infection transmitted by other means, effective hygiene, immunization, chemotherapy, antibiotics, and more may be needed.

Thus concepts and theory, sometimes disdained as remote from the world of facts and action, are critical to practice. The causes of disease sought by any scientists, medical or other, are bound to be limited by their concepts and by the frame of reference within which they work; these guide and generate operations in terms appropriate to theory. Different concepts lead scientists to seek different explanations of disease and to follow different courses of action to eliminate disease. For instance, Schwann's discovery of the cell as a basic structural element of living organisms provided the foundation for Virchow's realization that cells were also the locus or seat of disease, which led him to found the new discipline of cellular pathology (Ackerknecht 1948; Boyd 1991). Which course should be followed depends on whether we have in mind miasma or microorganisms and, beyond that, molecules, enzymes, cells, human behavior, or the structure of societies. In tracing the evolution of the chief concepts of causation in modern epidemiology, we shall aim to illuminate their nature and limitations.

The formulations above stress the role of concepts in guiding scientific and intellectual development, and the intimate connection of such development with the historical setting in which they arise and flourish. In the following descriptions and analyses of the history of ideas in epidemiology, however, the reader will find throughout that many accounts are tied to the ideas and work of individuals. The intended stance is neither to overemphasize, nor to neglect, either the social or the individualistic aspects of history. The evolution of our intellectual world is invariably driven by the interaction of gifted or effective or powerful individuals with the historical and environmental

forces, which, at given times and places, govern the possibilities of individual action.

References

Ackerknecht EH (1948). Anticontagionism between 1821 and 1867. *Bull Hist Med* 22:562–693.

Boyd BA (1991). *Rudolf Virchow: The Scientist as Citizen*. New York: Garland.

Fleck L (1935). *Genesis and Development of a Scientific Fact*. Chicago: University of Chicago Press. (See 1979 edition with comment by Kuhn).

Fracastoro G (1530). See Eatough G (ed. and transl.) 1984. *Fracastoro's Syphilis*. Liverpool, Great Britain: F. Cairns.

Greenwood M (1935). *Epidemics and Crowd Diseases: An Introduction to the Study of Epidemiology*. New York: MacMillian. (See 1977 reprint of the 1935 ed. published by New York: Arno Press.)

Hamer W (1928). *Epidemiology, Old and New*. London: Paul, Trench, Trubner and Company.

Institute of Medicine, Committee on the Use of Complementary and Alternative Medicine by the American Public, February I (2005). *Complementary and Alternative Medicine in the United States*. Washington, DC: National Academies Press.

Kronenberg F, Cushman LF, Wade CM, Kalmuss D, Chao MT (2006). Race/ethnicity and women's use of complementary and alternative medicine in the United States: results of a national survey. *Am J Public Health* 96:1236–1242.

Kuhn TS (1962). *The Structure of Scientific Revolutions*. New York: Oxford University Press. Republished 1970.

Langholf V (1990). *Medical Theories in Hippocrates: Early Texts and the 'Epidemics'*. Berlin: W. de Gruyter.

Locke J (1690). *An Essay Concerning Human Understanding*. See Locke J (1975) Nidditch PH, ed. Oxford: Clarendon Press.

Loudon I (1998). The tragedy of puerperal fever. *Health Libr Rev* 15:151–156.

Mead R (1720). *A Short Discourse Concerning Pestilential Contagion and the Methods to Be Used to Prevent It. By Richard Mead*. 4th ed. Dublin: reprinted by and for George Grierson.

Needham J (1956). *Science and Civilisation in China: History of Scientific Thought, Vol 2*. Cambridge, UK: Cambridge University Press.

Riley JC (1987). *The Eighteenth-Century Campaign to Avoid Disease*. New York: St. Martin's Press.

Rosen G (1958, 1993). *A History of Public Health*. Baltimore: Johns Hopkins University Press.

Singer CJ (1979). *Greek Biology & Greek Medicine*. 1st AMS ed. New York: AMS Press.

Susser M (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology*. New York: Oxford University Press.

Unschuld PU (1985). *Medicine in China: A History of Ideas*. Berkeley: University of California Press.

Zuckerman A (2004). Plague and contagionism in eighteenth-century England: the role of Richard Mead. *Bull Hist Med* 78:273–308.

3

The Concept of Environment

To construct an effective epidemiological study, a clear initial conception of the variables to be involved is critical. In the most general sense, one can say that this has been an essential foundation of the major medical and public health advances contributed by epidemiology: each advance attained new clarity in the specification of variables constituting causal models, clarity that in turn depended on new concepts. These advances have followed as investigators specified, either formally or only by implication, the nature of the variables they first attempted to relate to each other and then to elucidate their causal links and time sequences.

Hippocrates: Causal Theory

Historical examples from early times illustrate the process. The first epidemiological ideas available to Western medicine, including but not limited to epidemic disease, are found in the Hippocratic corpus and date from as early as 400 BCE. These works are certainly the product of many authors. In the book, *Airs, Waters, Places*, the authors speculate on the relations between disease and the physical environment of climate, water, soil, and prevailing winds (see Hippocrates, Chadwick and Mann 1950). This book gave

a description of diseases related to stagnant water in marshes and lakes, among which we recognize malaria. The authors go on to infer the nature of the connection of the disease with water.

Here is a passage:

> Stagnant water from marshes . . . will necessarily be warm, thick, and of unpleasant smell in summer, *because* such water is still and fed by rains; it is evaporated by the hot sun. *Thus* it is colored, harmful and bilious looking. In winter it will be cold, icy and muddied by melting snow and ice. This makes it productive of phlegm and hoarseness. Those who drink it have large firm spleens. . . . Their faces are thin *because* their spleens dissolve their flesh. . . . Their spleens remain enlarged summer and winter, and in addition cases of dropsy are frequent and fatal to a high degree. The *reason for this* is the occurrence during the summer of dysentery and diarrhea with prolonged fever. Such diseases . . . *cause* dropsy [italics added].

It was probably true that the water was stagnant and evaporated, that the people who drank it had large spleens, and that they suffered from diarrhea and fevers in the summer. The reasoning from these observations was good in its context, but at least half of the six causal statements (indicated by the operative words italicized above) are mistaken. The Hippocratic writer could not know that the malaria was caused by protozoa carried by Anopheles and other mosquitoes bred in the summer pools and not, as he thought, by the water itself. Nor could he know that enlargement of the spleen was a reaction to parasitic infection and not to drinking water. He could not even have guessed these things because no notion of biological agents of disease existed. The concept of microscopic living agents of disease is an obvious prerequisite for the supposition that particular microorganisms and parasites are capable of causing disease. The author was right to think that drinking the stagnant water could be harmful and often gave rise to diarrhea. Again, he could not have known that the harm would be caused by bacteria and protozoa carried by human excreta contaminating the water and not by the nature of the water itself.

This Hippocratic reasoning was mistaken on other grounds as well. It is clear to us now that one source of error, uncorrectable then, lay in a failure to differentiate between variables. Physicians had no basis adequate to conceive and characterize variables and to isolate them one from another. First, the author failed to differentiate certain unrelated manifestations of disease. He threw together malaria, dysentery, and dropsy (this last edematous condition has no connection explicable in modern terms). Second, he had few if any means for differentiating which phenomena might be causes, which might be effects, and which were associated without being linked to the causal sequence. He correctly observed an association between the icy, muddy water of winter and the hoarseness of respiratory infection, and he

wrongly concluded that the icy water itself caused the phlegm and hoarseness (Figure 3.1, Example 1). He correctly observed an association of the colored and bilious-looking water of summer with large spleens and wasting of the flesh, and he wrongly concluded that the large spleens caused the wasting (Figure 3.1, Example 2).

The complex relations between variables made understanding difficult. All the same, these Hippocratic works provided what are the first identified

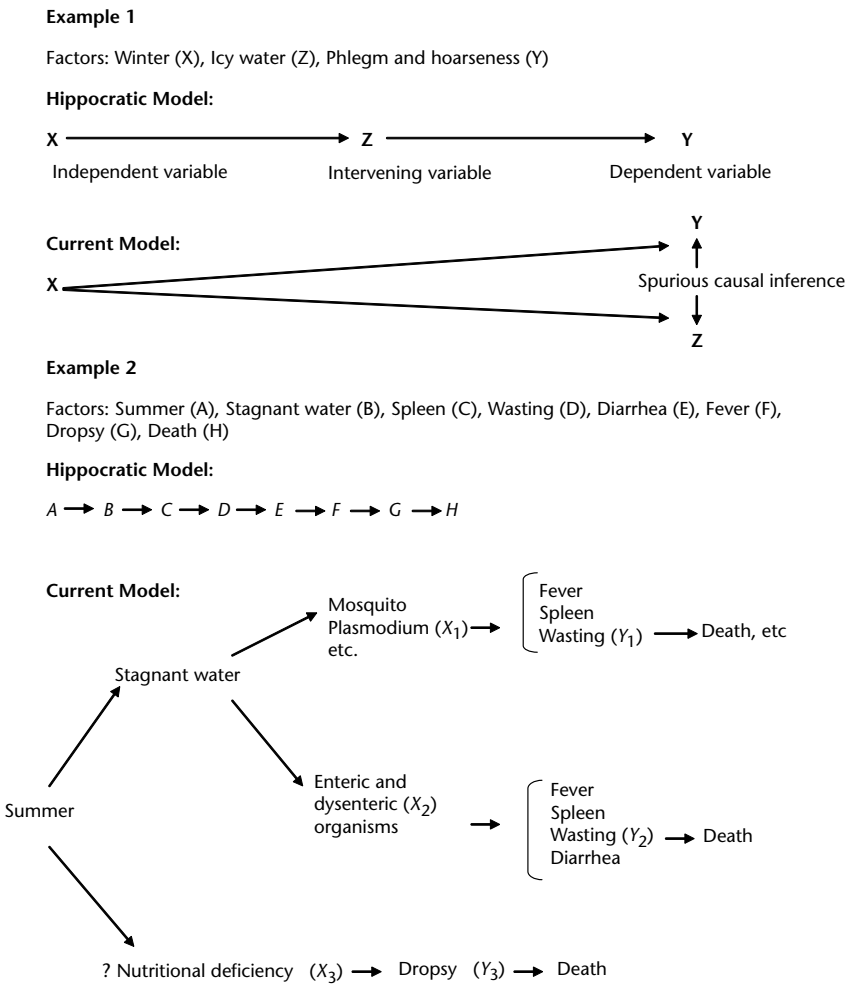


Figure 3.1 Hippocratic causal models.
Source: Figure first published in *Causal Thinking* (Susser 1973). Reprinted with permission from Oxford University Press, Inc.

concepts of disease causation in the Western corpus and in epidemiology, and indeed in the health sciences overall. These works distinguished the environment, as represented by air, water, and place, from the person, as represented by the individual constitution. Thus, early and perspicaciously, they separated the factors of person (or, as we might anticipate in later usage, the host) and environment. They saw both as contributing to the specific manifestations of disease.

After the Hippocratic era, fundamental advances in the concepts of disease and causation were few and long in gestation. In second-century Rome, Galen (129–200 CE) was the unquestioned leader in medicine. His codification of disease was enormously influential. In the main, however, he codified and consolidated Hippocratic theory. The authority and conviction carried by this work persisted well into the second millennium. Andreas Vesalius (1514–1564) published the chief work on the anatomical dissection of the human body only in 1543; he thereby corrected several of Galen's mistaken notions of anatomy (Moes 1976; Kemp 1998; Touwaide 2006).

Likewise, scientific physiology may be said to begin only in 1628. In that year William Harvey (1578–1657) published his treatise on the circulation of the blood. He had first dissected and studied intensely the anatomy of the heart. He then devised remarkably simple and creative experiments. These demonstrated unequivocally that blood circulated: blocking the veins in a limb caused them to fill with blood and swell; blocking the arteries caused the limb to blanch. Thereby, he also dispelled several now fantastic myths about the formation and function of the blood and the heart. Harvey was the exemplar who, by demolishing unfounded assumptions about the physiology of the blood, first opened the way to the experimental approach in medicine. More than a century after Harvey, in 1761, G.B. Morgagni (1682–1771) published the first work on morbid (or pathological) anatomy. These works are landmarks that provide a basis for the development—excruciatingly slow from a modern perspective—of the recognition and definition of specific disease entities.

Significant causal discoveries about environmental factors came equally slowly. The wide-ranging, original, and controversial thought of the Swiss physician known as Paracelsus (his given name was Theophrastus Bombastus von Hohenheim, 1493–1541) leaned toward alchemy, in which one sees the early stirrings of chemistry. He also arrived at a notion of specific disease entities. In his scheme, disease was attributed to multilevel environmental causes. These causes included the cosmos itself and hence were interpretable by astrology. Paracelsus did not ignore the mundane, however. Thus he localized silver mining as an important source of disease: in a text on the diseases of miners, he built the argument for environmental poisons as a cause of respiratory disease (Temkin and Sigerist 1996; Ball 2006). It was 1700

before Bernardo Ramazzini in his *Discourse on the Diseases of Workers* addressed the whole range of potential and known occupational hazards in a compilation about diseases connected with occupation. This work remained a major source into the nineteenth century (Ramazzini 1700/1964; Franco 1999).

As noted more fully in Chapter 2, the theory of disease proposed by Thomas Sydenham still had an Hippocratic orientation. He advanced and crystallized two central ideas: First, diseases were specific entities. In this respect, Sydenham's case reports formed an influential descriptive compendium that provided the core of clinical understanding in Europe for more than a century (Greenhill 1979). Second, environmental conditions governed the emergence and unfolding of disease—that is, in a phrase of our times, their natural history. In an individual, susceptibility to a disease was governed by the balance in a person's makeup of some sort of predisposing humors. The overarching environmental factor which precipitated disease he described as the *epidemic constitution*, a state Sydenham attributed to miasmata arising from a combination of season and atmosphere.

In the industrializing England of the eighteenth century, however, a few alert and perceptive physicians began to report convincing and epidemiologically slanted perceptions of better defined and more specific sources of environmental hazards. In 1753, in what is generally recognized as the first real prototype of a clinical trial, among sailors on prolonged service in the English channel, the ship's doctor James Lind reported that lemons cured the scurvy that afflicted sailors on long voyages (Carpenter 1986). Thereafter, lemons came to be used preventively if inconsistently, and Lind eventually was made chief physician at the naval hospital of the Royal Navy in Portsmouth. In the early nineteenth century, the Navy substituted cheaper West Indies limes for lemons (hence *limey* as a nickname for British sailors). Compared with lemons, however, limes are relatively deficient in vitamin C, thus epidemics of scurvy recurred aboard naval vessels kept long at sea.

In 1767 in Devon in southwest England, George Baker provided consummate proof that lead poisoning was the source of the Devonshire colic—expressed in a complex of colic, palsy mostly of the arms, and epilepsy—a disease suffered uniquely by cider drinkers in Devonshire county (Hernberg 2000). Like Lind before him, Baker supported his observations by research-based tests of his perceptions of the cause. Lind had done so by a prototype of the controlled trial; Baker confirmed the suspected source by chemical tests. In that county only, lead was used in several steps of cider production, and Baker's tests demonstrated the presence of lead.

In 1775, the leading London surgeon Percivall Pott (of the eponymous Pott's fracture) provided the first known evidence of an environmental source for cancer (Dobson 1972; Doll 1975). A sharp observer, he noted several

Figure 3.2 Young chimney sweeps from 1890. Credit: © Bettmann/Corbis. Reproduced by permission.

cases of scrotal cancer uniquely affecting postpubertal and adult chimney sweeps. Sagely, he connected this observation to their occupational history. "Climbing boys," Pott knew, were recruited at ages as young as five to eight (see Figure 3.2). From this fact, he further deduced that the continuous exposure to soot implied a long latency period. (*3–4 benzpyrene*, the responsible chemical agent in soot, was identified a century and a half later, in 1934). Legislation against such employment of young children was eventually put in place, though improvements were slow and incremental.[1] Aside from isolation and quarantine in the face of epidemics, these laws must surely rate as the first application on so large a scale of a sound epidemiological discovery to public policy in law. Indeed, it probably rates as the first national application ever.

[1] It took generations to eradicate the practice of using boys as human brooms to clean the numerous, narrow chimneys in England, despite the existence and growing availability of mechanical devices for these tasks after 1803. Initial attempts to legislate were weak and ineffective, and laws were not actively enforced. Regulatory measures dating from 1834 began to alleviate the pressure on children and to pave the way for more widespread use of other methods, including machinery. Not until 1842 were chimney-sweeper apprentices under the age of 16 forbidden; finally, in 1875 requirements for licensing chimney-sweepers further discouraged the residual exploitation of children in this manner across the whole of Great Britain (Phillips 1950). Other attempts to create legal protections for child workers in England and other industrializing countries arose from the backlash against child labor practices in the textile mills and the mines.

References

Ball P (2006). *The Devil's Doctor: Paracelsus and the World of Renaissance Magic and Science*. New York: Farrar, Straus and Giroux.

Carpenter KJ (1986). *The History of Scurvy and Vitamin C*. Cambridge: Cambridge University Press.

Dobson J (1972). Percivall Pott. *Ann R Coll Surg Engl* 50:54–65.

Doll R (1975). Pott and the path to prevention. *Arch Geschwulstforsch* 45(6):521–531.

Franco G (1999). Ramazzini and workers' health. *Lancet* 354:858–861.

Greenhill WA (1979). *The Works of Thomas Sydenham, M.D.* Special ed. Birmingham, AL: Classics of Medicine Library.

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

Hippocrates (1950). *The Medical Works of Hippocrates*. Chadwick J and Mann WN, transl. Oxford, England: Blackwell Scientific Publications.

Kemp M (1998). Vesalius's veracity. *Nature* 393(6684):421.

Moes RJ (1976). Andreas Vesalius and the anatomy of the upper extremity. *J Hand Surg [Am]* 1(1):23–28.

Phillips G (1950). The abolition of the climbing boys. *Am J Econ Soci* 9:445–462.

Ramazzini B (1700; reprinted 1964). *Diseases of Workers.* The. History of Medicine Series No. 23 New York: Hafner Publishing Co.

Temkin CL, Sigerist HE (1996). *Four Treatises of Theophrastus Von Hohenheim, Called Paracelsus*. Johns Hopkins paperbacks ed. Baltimore, MD: Johns Hopkins University Press.

Touwaide A (2006). The kidney from Galen to Vesalius—a first approach. *J Nephrol* 19(Suppl 10):S4–S8.

4

Numeracy in Epidemiology

The Beginnings: John Graunt and William Petty

No later than biblical times, as the story of King David attests, censuses had provided rulers of kingdoms or states with one-dimensional, cross-sectional measures of the numbers of their subjects. In essence, they were instruments of government, which measured potential resources. The focus was at once on revenue garnered from taxes, and on the drafting of men for soldiering. Point estimates from such cross-sectional data did not foster any kind of longitudinal view. They did not yield estimates of population dynamics and trends in populations over time.

Only in the seventeenth century were population counts put to fuller use. In 1662, the London haberdasher John Graunt (1620–1674) published a slender book that was to be a foundation of much that was to follow in the population sciences (Graunt 1662). The book, *Natural and Political Observations made upon the Bills of Mortality*, reports deaths by apparent cause that were used to provide a monitoring system for plague and pestilence. Women had been employed to visit homes in the narrowly defined area of the City of London (not a large area at that time) and to report any deaths. Possibly, though not certainly, these reports were so intended from the outset. As Daniel Defoe observed in 1722 in his fictional account, *A Journal of the Plague Year*, the

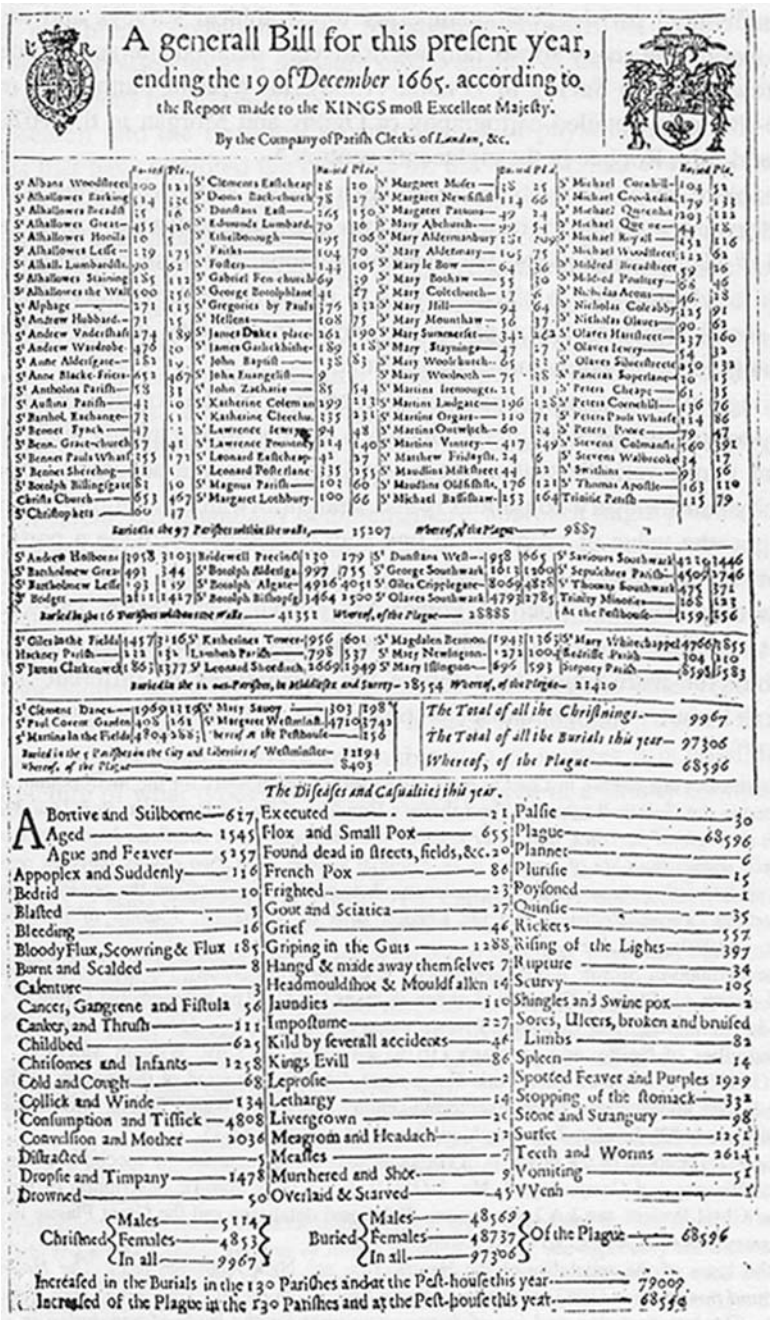
Bills afforded forewarning of epidemic outbreaks and allowed the wealthier classes to make timely departures from the city (Figure 4.1).

By his own account Graunt began his study as a "matter of curiosity," [cap II, p. 12] a stimulus many great scientists have considered critical to discovery. He writes, "Now having (I know not by what accident) engaged my thoughts upon the *Bills of Mortality.* …" and goes on to say, "observations which I happen'd to make (for I designed them not). …" Yet his perspective could well fit modern epidemiology. Later he remarks of these observations, "finding some Truths, and not commonly believed Opinions.… I proceeded farther to consider what benefit the knowledge of the same would bring to the World" (published in London, 1662). This work was recognized as sufficiently original to earn him a place in the Royal Society, then newly founded for the study of science. It turned out to be a place well merited.

Graunt's major and unprecedented contribution was to use the weekly returns of the Bills of Mortality for the parishes of London to quantify the occurrence of deaths by compiling them longitudinally over time. No denominators were available, and he was thus limited to enumeration of the deaths solely by age group, sex, location, and time of death. No evidence emerges to indicate whether he was aware of that lack, which in any case did not impede him. Thus his first fundamental contribution was to classify the data on deaths into groups. In current terms, he could thus more nearly compare like with like to detect differences. In groups differentiated at death by individual age, location, and time of death, Graunt found differences in mortality across the parishes of the City of London, and noted changes in causes of death over time. Seeking for an index of the health of the community, a novel idea for the time, he proposed that the proportion of deaths among persons 70 years and over would serve the purpose.[1] Finally, he offered statistical interpretations of his findings. Among his observations, he speculated on the excess numbers of deaths among men as compared with women in London (14 to 13), noted that deaths occurred in the same proportion, but that more men than women died violent deaths.

In interpreting these data, he displayed due caution, observing that the classification stemmed from the causes of death assigned in the reports of the ancient matrons who had collected the information recorded in the Bills of Mortality. Thus, Graunt asserted, during plague years denial of affliction by the plague was common. Resistance to reporting the presence of plague

[1] "Onely the question is, what number of Years the *Searchers* call *Aged*, which I conceive must be the same, that *David* calls so, *viz.* 70. For no man can be said to die properly of *Age*, who is much less: it follows from hence, that if in any other Country more than seven of the 100 live beyond 70, such Country is to be esteemed more healthfull then this of our City." [Graunt's No. 18 in Cap.II, *General Observations upon the Casualties*]



A generall Bill for this present year, ending the 19 of *December* 1665. according to the Report made to the KINGS most Excellent Majesty.

By the Company of Parish Clerks of *London*, &c.

	Buried	Plo.		Buried	Plo.		Buried	Plo.		Buried	Plo.
S. Albans Woodstreet	200	121	S. Clements Eastcheap	28	20	S. Margaret Moses	28	25	S. Michael Cornhill	104	52
S. Alhallows Barking	124	126	S. Dionis Backchurch	78	27	S. Margaret Newfishst.	114	66	S. Michael Crooked-la	179	131
S. Alhallows Bread-st	15	16	S. Dunstans East	165	150	S. Margaret Pattens	47	24	S. Michael Quernhith	203	122
S. Alhallows Great	455	430	S. Edmunds Lombard.	70	36	S. Mary Abchurch	99	54	S. Michael Queenh.	44	18
S. Alhallows Honila	10	5	S. Ethelborough	195	106	S. Mary Aldermanbury	151	109	S. Michael Royall	451	116
S. Alhallows Lesse	139	173	S. Faiths	104	70	S. Mary Aldermary	105	75	S. Michael Woodstreet	122	62
S. Alhall. Lumbardst.	30	62	S. Fosters	144	105	S. Mary le Bow	64	36	S. Mildred Breadstreet	59	30
S. Alhallows Staining	185	112	S. Gabriel Fen church	69	39	S. Mary Bothaw	55	30	S. Mildred Poultry	86	46
S. Alhallows the Wall	300	256	S. George Botolphlane	41	27	S. Mary Colechurch	17	6	S. Nicholas Acons	46	28
S. Alphage	171	115	S. Gregories by Pauls	176	112	S. Mary Hill	94	64	S. Nicholas Colesbby	125	91
S. Andrew Hubbard	71	55	S. Hellens	108	75	S. Mary Mounthaw	36	37	S. Nicholas Olaves	90	62
S. Andrew Undershaft	274	189	S. James Dukes place	161	140	S. Mary Somersetst.	342	262	S. Olaves Hartstreet	237	160
S. Andrew Wardrobe	476	210	S. James Garlickhithe	189	118	S. Mary Staynings	47	27	S. Olaves Ivery	54	32
S. Anne Aldersgate	181	119	S. John Baptist	138	83	S. Mary Woolchurch	65	21	S. Olaves Silverstreet	250	132
S. Anne Blacke-Friers	452	467	S. John Evangelist	9		S. Mary Woolnoth	75	28	S. Pancras Soperlane	20	15
S. Anthons Parish	58	31	S. John Zacharies	85	54	S. Martins Ironmonger.	21	11	S. Peters Cheape	61	35
S. Aultons Parish	42	10	S. Katherine Colemans	199	113	S. Martins Ludgate	196	128	S. Peters Cornehill	136	76
S. Bartho. Exchange	72	61	S. Katherine Cheechs.	135	131	S. Martins Organ	110	71	S. Peters Poole Wharfe	114	86
S. Bennet Tynch	47	2	S. Lawrence Iewry	94	48	S. Martins Outwhich	60	24	S. Peters Poore	79	47
S. Benn. Grace-church	57	41	S. Lawrence Pountney	114	140	S. Martins Vintry	417	149	S. Stevens Colmanst.	160	391
S. Bennet Pauls Wharf	155	171	S. Leonard Eastcheap	42	27	S. Matthew Fridaystr.	24	6	S. Stevens Walbrooke	34	17
S. Bennet Sherehog	11	1	S. Leonard Fosterlane	135	155	S. Maudlins Milkstreet	44	22	S. Swithins	93	56
S. Botolph Billingsgate	81	10	S. Magnus Parish	102	60	S. Maudlins Oldfishst.	176	121	S. Thomas Apostle	163	110
Christs Church	553	467	S. Margaret Lothbury	100	66	S. Michael Bassishaw	153	164	Trinitie Parish	125	79
S. Christophers	60	17									

Buried in the 97 Parishes within the walls — 15207 Whereof, of the Plague — 9887

S. Andrew Holborne	1938	3103	Bridewell Precinct	130	179	S. Dunstans West	958	1665	S. Saviours Southwark	1229	1446
S. Bartholomew Great	491	344	S. Botolph Aldersgate	997	1755	S. George Southwark	1011	1360	S. Sepulchres Parish	4509	3746
S. Bartholomew Leste	193	119	S. Botolph Algate	4916	4051	S. Giles Cripplegate	8069	4828	S. Thomas Southwark	475	371
S. Bridget	2111	1417	S. Botolph Bishopsg.	1464	1500	S. Olaves Southwark	1479	1785	Trinity Minories	168	123
Buried in the 16 Parishes without the walls — 41351 Whereof, of the Plague — 28888									At the Pesthouse	159	156

S. Giles in the Fields	1457	1165	S. Katherines Tower	956	601	S. Magdalen Bermon.	1941	1136	S. Mary Whitechappel	4766	1855
Hackney Parish	522	134	Lambeth Parish	798	537	S. Mary Newingtons	1272	1004	Redriffe Parish	304	210
S. James Clarkenwell	865	1377	S. Leonard Shoreditch	2669	1949	S. Mary Islington	496	593	Stepney Parish	6598	1583
Buried in the 11 out-Parishes, in Middlesex and Surrey — 28554 Whereof, of the Plague — 21410											

S. Clements Danes	1969	1126	S. Mary Savoy	303	198						
S. Paul Covent Garden	405	161	S. Margarets Westmins.	4710	1742						
S. Martins in the Fields	1804	288	Whereof in the Pesthouse		156						
Buried in the 5 Parishes in the City and Liberties of Westminster — 12194 Whereof, of the Plague — 8403											

The Total of all the Christnings	9967
The Total of all the Burials this year	97306
Whereof, of the Plague	68596

The Diseases and Casualties this year.

Abortive and Stillborne	617	Executed	21	Palsie	30
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in streets, fields, &c.	20	Planet	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	15
Bedrid	10	Frighted	23	Poysoned	1
Blasted	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Rising of the Lights	397
Burnt and Scalded	8	Hangd & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	14	Scurvy	105
Cancer, Gangrene and Fistula	56	Jaundies	110	Shingles and Swine pox	2
Canker, and Thrush	111	Impostume	227	Sores, Ulcers, broken and bruised	
Childbed	625	Kild by severall accidents	46	Limbs	82
Chrisomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	68	Leprosie	2	Spotted Feaver and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the Stomack	332
Consumption and Tisick	4808	Livergrown	26	Stone and Strangury	98
Convulsion and Mother	2036	Meagrom and Headach	12	Surfet	1251
Distracted	5	Measles	7	Teeth and Worms	2614
Dropsie and Timpany	1478	Murthered and Shot	9	Vomiting	51
Drowned	50	Overlaid & Starved	45	Vvenh	1

Christned	Males 5114	Buried	Males 48569	Of the Plague	68596
	Females 4853		Females 48737		
	In all 9967		In all 97306		

Increased in the Burials in the 130 Parishes and at the Pest-house this year — 79009

Increased of the Plague in the 130 Parishes and at the Pest-house this year — 68554

Figure 4.1 A page from *Graunt's Bills of Mortality.*

was unsurprising: the isolation of a household then imposed had severe consequences for the normal routines that sustained daily life. Graunt supposed that denials and resistance to reporting by households was to be expected in face of the complete isolation of households contingent on the notification of plague. On that ground, he estimated that during the epidemic, causes of death other than plague were inflated by no less than one fifth.[2]

> That the knowledg even of the numbers, which die of the *Plague*, is not sufficiently deduced from the meer Report of the *Searchers*, which onely the Bills afford; but from other Rationcinations, and comparings of the *Plague* with some other *Casualties*.
>
> For we shall make it probable, that in Years of *Plague* a quarter part more dies of that *Disease* than are set down; the same we shall also prove by the other *Casualties*. Wherefore, if it be necessary to impart to the World a good Accompt of some few *Casualties*, which since it cannot well be done without giving an Accompt of them all, then is our common practice of doing so very apt, and rational. (Graunt 1662)

From existing data accumulated in registers, Graunt carried out what would now be described as a secondary analysis. The form and perspicacity of Graunt's longitudinal analyses have moved authoritative scholars to attribute the invention of the life table to him, if only in a somewhat rudimentary antecedent form. Starting from baptismal records of births in the given population of the City of London parishes, he worked from three variables: the numbers, respectively, of those at risk, of deaths, and of survivors. When analyzed longitudinally at each successive age, these are indeed the essential basis of the life table. From these data, Graunt calculated comparative proportions and ratios for mortality at different ages over time.[3]

> Where as we have found, that of 100 quick Conceptions about 36 of them die before they be six years old, and that perhaps but one surviveth 76, we, having seven *Decads* between six and 76, we sought six mean proportional numbers between 64, the remainer, living at six years, and the one, which survives 76, and finde, that the numbers following are practically near enough to the truth; for men do not die in exact Proportions, nor in Fractions: from when arises this Table following
>
> ...

[2] In 2004, a similar approach, relevant to early estimates of HIV mortality in South Africa, was used by Bradshaw et al. (Bradshaw et al. 2004). Here, because rates of death rose by age, sex, and presumably AIDS-related causes (tuberculosis, pneumonia, diarrhea, for instance), the authors could confidently argue that AIDS mortality was underdescribed by counting only AIDS as cause.

[3] Greenwood pointed out that an assumption, implicit in any fully valid analysis of this type, requires a stationary population, not only in the sense that births and deaths were equal, but that death rates by age, too, do not vary over time (Greenwood 1977a).

Figure 4.2 William Petty, detail of a portrait attributed to Isaac Fuller, c. 1649–1650. Credit: © National Portrait Gallery, London.

From when it follows, that of the said 100 conceived there remains alive at six years end 64.

At Sixteen years end	40	At Fifty six	6
At Twenty six	25	At Sixty six	3
At Thirty six	16	At Seventy six	1
At Forty six	10	At Eighty	0

(Graunt 1662)

Graunt had not yet envisaged the notion of life expectancy, however. That was left to Edmond Halley (1656–1742), distinguished eighteenth-century mathematician and astronomer. In 1693, Halley obtained and analyzed a time series of funerals supplemented by a series of births from the city of Breslau. These data came into his hands when some rationalists of that city sought to disprove a contention of astrologers that particular years of one's life were especially dangerous (Lazarsfeld 1961). His life tables showed incidentally that individual age could be used as a predictor of mortality for actuarial purposes (Figure 4.2).[4]

[4] This great gift to the life insurance business made it a profitable industry. Through the succeeding centuries living conditions tended to improve and mortality rates to decline (if not altogether steadily), thus predictions routinely underestimated life expectancy and overestimated both risk to the insured and greater anticipated cost to the insurer, engendering higher insurance rates than appropriate to the insured.

It was probably the colorful physician William Petty, Graunt's friend and confrere, who nominated him for membership in the newly founded Royal Society. Petty's contributions were equally original. He collected and analyzed statistics for mortality by cause and location, as in a study of the city of Dublin (first published in 1690) with a view to relating variation to local environmental conditions (Petty 1690; see Petty and Graunt, republished 1986). His ideas have been significant and influential, more especially for economists in the health and insurance fields. Among other things, he devised a method for the costing of lost lives, and can thus be seen as the forerunner of cost-benefit analysis. The ideas of both Graunt and Petty long lay dormant. But each pointed the way to the triumphs of enumeration a century or more later. Variously accumulated data began to drive a growing awareness of health and social problems, first among the urbanizing and industrializing societies of France, and then of England. Today, Graunt is recognized as forefather of demography (Glass 1963; Greenwood 1977b). He has equal importance for epidemiology, as J.N. Morris observed in his stimulating text (Morris 1957; see also Rothman 1996). Graunt had laid the foundation both for the forerunner of life tables that Halley first constructed, and for the estimates of life expectancy and the tabulations of causes of death so familiar now in both epidemiology and demography.

In epidemiology, as in other sciences, the interplay between theory and method, and constant transpositions from other fields of study, are profoundly influenced by the social forces of the time. As with Graunt's work, these influences often find expression in apparent accidents of circumstance. Thus the existence of particular vital records seemingly kept to confer on privileged classes the particular social advantage of forewarning of epidemic plague led to new studies and understanding.

References

Bradshaw D, Laubscher R, Dorrington R, Bourne DE, Timaeus IM (2004). Unabated rise in number of adult deaths in South Africa. *S Afr Med J* 94:278–279.

Defoe D (1722). *A Journal of the Plague Year.* See Defoe D (1992), Backscheider P, ed. New York: W.W. Norton and Co.

Glass DV (1963). John Graunt and his natural and political observations. *Proc R Soc Lond B Biol Sci* 159:1–37.

Graunt J (1662). *Natural and Political Observations Made Upon the Bills of Mortality.* Willcox, W.F., ed. Baltimore: The Johns Hopkins Press, 1939. Reprinted Ann Arbor, MI: UMI Books on Demand, 2001.

Greenwood M (1977a). *Epidemics and Crowd Diseases an Introduction to the Study of Epidemiology.* Reprint of the 1935 ed. published by Macmillan, New York. New York: Arno Press.

Greenwood M (1977b). *Medical Statistics From Graunt to Farr.* 1941, 1943. New York: Arno Press.

Lazarsfeld PE (1961). Notes on the history of quantification in sociology-trends, sources and problems. In Woolf, eds. *Quantification*. New York: Bobbs-Merril.

Morris JN (1957). *Uses of Epidemiology.* Edinburgh: Livingstone.

Petty W, Graunt J (republished 1986). *The Economic Writings of Sir William Petty: Together With the Observations Upon the Bills of Mortality More Probably by John Graunt*. (Hull CH, ed.). Fairfield, NJ: A.M. Kelley.

Rothman KJ (1996). Lessons from John Graunt. *Lancet* 347:37–39.

5

The French Enlightenment, Epidemiology, and Public Health

After Graunt and Petty and through the eighteenth century, advances in method applied to enumerating disease distributions were somewhat sporadic. In France late in that century the Enlightenment, followed by the climactic French Revolution with its slogan of liberty, equality, and fraternity, opened new intellectual avenues (so most if not all historians believe). The period of the Enlightenment of the eighteenth century began and had its maximal impact in France.[1] Among the philosophers whose writings inspired philanthropists and physicians were Voltaire, who sought to undermine the Church's hold over many institutions, including hospitals; Rousseau, who laid great weight on the election of individuals in a democratic state; Diderot, who provided many ideas relevant to medicine and health in his Encyclopedia (hygiene was a chapter in 1760; Hallé, 25 years later defined it with more focus); Montesquieu, who explicitly wrote (1748, *The Spirit of Laws*) "the state owes every citizen assured subsistence, food, clothing, and a way of life not injurious to health."

[1] The impact of the new libertarian ideas was not confined to France. Thus, Mary Wollstonecraft in England closely followed these ideas, and wrote her famous pamphlet on "A Vindication of the Rights of Woman" in 1792, demanding "JUSTICE for one half of the human race."

A major change in the focus of the reigning intellectuals of the time thus created a new social and political consciousness and gave substance to new and literally revolutionary ideas. Central to this thinking was the idea of creating a freer and fairer society, goals clearly entailing improved conditions of life for the least fortunate. (As noted in the previous chapter, such an idea was not foreign to Graunt, given his proposal of an index of a healthful society, a century and more before.) These ideas included concepts relevant to health, hence in the late eighteenth and early nineteenth century, a new social and epidemiological consciousness arose. This movement has been elucidated in the English literature much more fully than before by recent historians of public health and medicine (Ackerknecht 1948; La Berge 1977, 1992; Coleman 1982; Rosen 1993; Weiner 1993; Barnes 2006; Quinlan 2007).

Shortly before the Revolution, in response to rising reformist fervor Louis XVI instructed his finance minister Jacques Necker (1732–1804) to attend to the scandalous conditions of hospitals and hospices that housed the indigent throughout France. Necker appointed Dr. Jean Colombier, previously a physician in the military, to inspect all hospitals, laboratories, and prisons in the country. In carrying out this charge, Colombier documented an alarming number of degrading conditions, and he proposed reforms and rules modeled on the military for nurses, physicians, and pharmacists. Colombier had two famous sons-in-law, both physicians. The career of one of these, Michel Augustin Thouret, member of the Royal Society of Medicine and Dean of the Paris Health School, takes us through the first actions of the Revolutionary government which took over the role of governing France in 1789.[2]

The National Constituent Assembly created two Committees. In setting up these committees, the Assembly intended that equity should apply to the health and health care of all citizens. The Committee on Poverty, chaired by the nobleman Duc de La Rochefoucauld-Liancourt was the first to be appointed; Michel-Augustin Thouret and Pierre Jean Georges Cabanis (a physician-philosopher) both served on it.[3] The committee not only made a detailed on-the-spot survey of hospitals and physicians across France, but

[2] The States-General, representing nobles, clergy, and the Third estate, had been summoned by the King in May 1789; by June, the Third estate had assumed leadership, creating the National Assembly. On July 14 the Bastille was stormed, and on August 4, nobles and clergy relinquished their privileges, and the assembly proclaimed the Rights of Man. Within months, illness was identified as a major cause of poverty, and the Poverty Committee began to consider the citizens' claim to health in 1790 and 1791 (Weiner 1993).

[3] The Committee was guided by Jacques Tenon who in 1788 had written: "Hospitals are a measure of civilization: as people agglomerate and grow more humane and better educated, the hospitals become more appropriate to their needs and better kept" (Weiner 1996).

they also recommended training and roles for health officers each of whom was to be elected in his *arrondissement* (district) and assigned responsibility both for preventive health and care of the indigent sick. Almost simultaneously, the National Assembly set up a Health Committee comprising the 17 doctors among the elected legislators, and chaired by Joseph Ignace Guillotin.[4] Although Dr. Guillotin had participated in some of the work of the Poverty Committee, he and his fellows on the Health Committee were flatly opposed to the recommendation that doctors should be salaried public servants and the proposal failed.

A major task facing the Committee on Poverty was to discriminate between services for the sick, and services for the indigent. Thouret estimated that the numbers of the needy poor would be 1/25 of the population, among whom half would be either aged, or homeless children, or chronic invalids. These numbers provided the budgetary estimates that the committee put to the National Assembly and which were duly accepted. However, the Health Committee again rejected the estimates and no action followed.

Among the many new ideas stemming from the Enlightenment that contributed to the development of epidemiological thought was the new mathematical principle of uncertainty, illuminated by the renowned mathematician Pierre Laplace (1749–1827). His works of 1812 and 1814 elaborated the thought-provoking theorem of Thomas Bayes (1702–1761) on two types of probability. Soon thereafter P.J.G. Cabanis (1757–1808), a physician and philosopher and an active member of the Committee on Poverty, applied the idea of probabilities to the study of medical and social phenomena. He argued that the evaluation of treatment required carefully observed and enumerated data analyzed in terms of probabilities. Moreover, he drew a sharp distinction between observations of individuals and those of the masses and was probably the first to do so. The distinction, fundamental for epidemiology and any other science engaged in analyzing populations, is not always respected even in our times.

Another member of the committee on poverty was Philippe Pinel (1745–1826). In early nineteenth-century France, Pinel, in his work in mental hospitals, soon applied the ideas of enumeration to the analysis of data on disease in populations, more particularly in psychiatry. Like William Tuke before

[4] Guillotin, Joseph Ignace (1738–1814). Member of the Legislative Assembly, Chairman of the Health Committee, member of the Poverty Committee, member of the Paris Vaccination Committee. He thought that doctors, rather than administrators and philosophers should play the determinant role in reforming health services; he was especially interested in the training of doctors and developing the curricula for their training. He argued for an egalitarian approach for beheading aristocrats and commoners, and invented the "simple mechanical device" for doing this, which still bears his name.

him in late eighteenth-century York, and John Connolly in early and mid-nineteenth century London, Pinel was a pioneer of the moral treatment of mental patients, then objects of public ridicule confined in prison-like conditions. This was a matter first of freeing the patients from the cruel constraints of shackles and close confinement, and then ensuring as much humane social intercourse as possible[5] (Deutsch 1949; Susser 1968; Armitage 1983).

In France, medicine flourished as nowhere else in the first half of the nineteenth century, most especially in the School of Paris. Pierre Louis (1787–1872) had returned from the Napoleonic wars as a young physician with new ideas. Unlike many of the Parisian eminences of the time in clinical medicine, Louis was thoroughly committed to statistical analysis (Armitage 1983). Among his hospital patients, he was dedicated to studies of disease manifestations and frequencies as well as to their treatment through testing, careful quantification, and statistical analysis of his hospital patients (he was also probably the first to distinguish typhus and typhoid as separate entities).

For instance, with the advocacy of leading physicians, bleeding by leeches had reached vast proportions; perhaps 30,000 leeches were imported each year. Indeed, so widespread was the practice that, colloquially, physicians themselves came to be called leeches. Louis suspected that bleeding brought more danger than help. Eventually the study of the practice of this ineffectual treatment for fatal pneumonias damped down the vogue for bleeding as a panacea. Among hospital patients admitted for pneumonia at successively later intervals after onset, fatalities were highest among those bled soonest, and thus bled longest and presumably most. He carefully noted that a presentation to the doctor of the illnesses of those bled early or late were equally grave. The data from this now famous observational study shown in the Table 5.1 cannot, however, provide assurance that selective survival was controlled thereby. For instance, those bled early in the course of the illness may well have been admitted early because of the severity of the attack. In fact, Major Greenwood, the founding professor of epidemiology and statistics in the London School of Hygiene, noted the insufficiencies of Louis's data although, it must be allowed, by the standards of a century later (Louis 1835, 1836; Greenwood 1936; also see Armitage 1983; Morabia 1996).

[5] Pinel promoted human treatment and reduced confinement for the mentally ill, both at Bicêtre and then at Salpêtrière, where he was physician in chief from 1795 to 1826. Though asylum reform took place years after the French Revolution, it was influenced by the doctrine of the rights of man. A painting of Pinel by Robert-Fleury (1876) portrays an idealized moment of removing chains from female patients in the Salpêtrière Hospital in Paris; it hangs in Charcot Library of the Salpêtrière Hospital Medical School and was viewed by Sigmund Freud (Harris 2003a; Fee and Brown 2006). Another painting, by Charles Louis-Muller (1849), depicts a similar scene at Bicêtre (Harris 2003b).

Table 5.1 Pierre Louis's Analysis of Bleeding as a Remedy for Pneumonia

Day Bled After Onset	Deaths (*n*)	Survived (*n*)
1–3	12	12
4–6	12	22
7–9	3	16

Louis established an international reputation as a teacher of rigorous statistics in medicine. Many of the students he attracted from Western Europe and the United States later became distinguished in their own right. Abraham and David Lilienfeld constructed a genealogical tree to trace Louis's influence (Lilienfeld and Lilienfeld 1980). Several of his students from Britain were major contributors to nineteenth-century epidemiology. One of these was William Farr (1807–1883), a founder of modern vital statistics and epidemiology, who held the post of Compiler of Medical Statistics from 1839 until 1880 (Eyler 1979; Susser 1987). Another was William Augustus Guy (1810–1885), important in the Sanitary Movement and also for his epidemiological studies of occupation and nutrition.

William Budd (1811–1880), another of Louis's students, like his mentor also differentiated typhoid as a disease distinct from typhus, and established typhoid as a transmissible infectious disease with an incubation period. He observed that survival of an attack conferred immunity. Although one gives due credit to Lind, Potts, and Baker for their singular early observations, as a founder of clinical epidemiology, Budd ranks with Louis and Parent-Duchatelet. John Simon (1816–1904), was another student of Louis's. He held the first appointment as Medical Officer of Health for London from 1848, and thereafter was appointed Chief Medical Officer of the newly founded General Board of Health. Louis's students from the United States also included some who became well known; for instance, Elisha Bartlett (1804–1855), and George C. Shattuck Jr.

Nonetheless, Louis was not himself a direct contributor to epidemiology and public health in their broadest sense. Those interests were pursued by several other contemporary French physicians who shared a belief in the utility of statistical analysis. Many, like Louis, were among the large body of medical graduates discharged from Napoleon's armies. Their main early experience was thus in the military. Then, as always, soldiers were selected for youth and fitness. Surgeons and medical hygienists best served the predominant medical needs of the military. The pressing work of surgeons was to contend with the sporadic mass trauma of battle and its aftermath; that of hygiene was to ensure that infection did not spread and disable the bodies of troops assembled *en masse*.

A second son-in-law of Jean Colombier was Rene-Nicholas Desgenettes, also a military medical man. When Napoleon's armies reached new territories and climates, they of course encountered new health hazards. Although Napoleon openly derided the therapeutic ability of doctors, he did recognize that some study of these new hazards was needed, and in 1798 he set up an Egyptian Institute to do this. Bonaparte had 164 health officers and 167 scientists, mathematicians, and artists among his forces and Dr. Desgenettes was prominent among them. Many of these diarized and later published their experiences. Although they diagnosed bubonic plague, typhoid, and yellow fever, they could not stem the enormous mortality they and their men suffered from these and other exposures.

Accordingly, in the aftermath of the wars, medicine in Paris gained renown beyond its dominance in clinical medicine. A main preoccupation of physicians in the army had inevitably been with the hygiene of food, sanitation, and epidemic infection among men massed in camps. Thus, with the end to Napoleon's wars after his final defeat and capture at Waterloo in 1815, there followed a post-Napoleonic era rich in young returning medical men trained in matters of hygiene and a degree of competence in their maintenance. Some of these medical men became hygienists and initiators of a general public health movement. Intent on extending their knowledge, they swelled the ranks of the students of Jean Noel Hallé,[6] who was by then formulating a public health curriculum that he was teaching to overflow audiences at the National Academy of Medicine in Paris.

The accumulation of recent histories of Parisian scholars of that time makes clear their central contribution. The aims of the new discipline have a modern ring. They went well beyond the historic emergency functions either of warfare or of trying to stem epidemics (predominantly cholera in the nineteenth century). In short, the aims were to establish a discipline and a profession of public health; to gain an understanding of public health problems; to devise the means to control or eliminate such problems; and finally to reform law and administration and thereby ensure that advances would be institutionalized and sustained. The development of the discipline relied heavily on the ideas and investigations of physicians with broad social and political concerns. Notable among them are Villermé and Parent-Duchatelet. These

[6] Hallé, Jean Noel (1754–1822). Hallé was a product of the prerevolutionary Enlightenment and its health reformists. He kept out of the way during the height of the revolution, returning to prominence with the Napoleonic consulate. As a Professor of Hygiene at the Paris Health School, he had written a chapter on hygiene in the Encyclopedia of 1798, and subsequently, in 1815, revised his thinking on this. His lectures, which drew large crowds, outlined the structure and content of public health.

two contemporaries pursued population studies of epidemiological type, and devoted themselves to creating a new public health movement.

Thus the surveys, estimates, and reforms developed by the Committees of the National Assembly did not follow a straightforward course. Nevertheless, in the post-Napoleonic era and through the first half of the nineteenth century, many of the new plans were implemented.

Luis-René Villermé (1782–1863), contemporary with Louis, was preeminent in his time as a student of social and occupational environmental effects on mortality and sickness (Coleman 1982). He is said to have been of liberal but not revolutionary persuasion. Although concerned for the poor in the growing industrializing cities of France, he was averse to government control and intervention. Beginning in the 1820s, he devoted himself to studying the connection of human biological states of health with social circumstances. To accomplish his many investigations of the ill-consequences of particular environments, Villermé created new survey methods for epidemiological studies.

Villermé undertook his intense studies of mortality in Paris as head of a commission of the Royal Academy of Medicine (of which he was a member, as he was also of the influential Paris Health Council). The spirit of the times remained anti-contagionist and wedded to miasma theory. Thus, in the first instance these studies addressed factors primarily in the physical environment. In successive, systematic investigations, the Commission tested the relationship of mortality in the city across its 12 *arrondissements*. One by one, he eliminated each potential environmental cause as bearing no relation to the distribution of mortality. Step-by-step, he tested the potential sources of miasma that could be measured. These tests included patterns of humidity, altitude, soil, the prevailing winds, and proximity to the fouled waters of the Seine. Population density and congestion, too—examined first in terms of the ratio of open space to housing, and then of household density—also failed to relate to mortality patterns.

At the last, Villermé hit upon a means of testing the relation of poverty to mortality. Below a certain level of resources, households were exempt from the tax levy. It came to him that, in each *arrondissement*, data according to their relative wealth could be derived from the taxes paid by households. In 1826 he published an ecological study (Villermé 1826). Districts were compared in terms of the proportion in each exempted from taxation by virtue of average income below the taxable threshold. Across the 12 arrondissements, a tabulation of these nontaxable proportions against mortality correlates the two indices perfectly with but one exception (Table 5.2).

A potential flaw remained. The deaths recorded were those that occurred within the home or elsewhere and outside the hospitals and institutions. Such deaths were predominantly those of the poor. Villermé obtained a count of the missing hospital deaths and added them to his tabulations to correct the

Table 5.2 Villermé's Data: Linking Poverty to Mortality in Paris

Arrondissement	Proportion Exempt	Deaths at Home: 1 in
Montmartre	0.07	62
Chaussée d'Antin	0.11	60
Roule, Tuileries	0.11	58
St. Honoré, Louvre	0.15	58
Luxembourg	0.19	51
Porte St. Denis, Temple	0.21	54
Faubourg St. Denis	0.22	53
St. Avoie	0.22	52
Monnaie, Invalides	0.23	50
Ile St. Louis	0.31	44
St. Antoine	0.32	43
Jardin du Roi	0.38	43

Mortality rates from 1817 to 1821. Arrondissements in Paris ranked by proportion of properties exempt from taxation.

potential bias from these institutional omissions. Although the revised and expanded analysis that included an additional measure of income and also mortality over time among *arrondissements* produced a slightly different order of mortality rates, still large mortality differences supported the initially observed relationship of income to death rates (Table 5.3).

In all this work Villermé gained much from his friendship with the Belgian Adolphé Lambert Quetelet (1796–1874), a founder of statistical method (Ackerknecht 1952; Lazarsfeld 1977). Quetelet's ingenuities greatly advanced quantitative methods for the study of social and biological phenomena. He used rates specific to age and sex groups, and introduced the notions of averages, of the normal and binomial distributions and sampling. He also devised an anthropometric index named for him that adjusted body weight for height and is still in common use. This new package of statistical methods was first taken up in France and then in England (Shryock 1961).

In his several studies across France over subsequent years, Villermé reinforced his finding of the strong relationship of mortality with poor material circumstances and poverty. For instance, in 1834 in the textile manufacturing town of Mulhouse he began an original and detailed six-year study of mortality by occupation. In these latter studies, Villermé relied mainly on direct observation and inspection rather than on statistical and epidemiological approaches but still produced convincing accounts. His two-volume publication of 1840 demonstrated the high mortality risk of the mill workers, many of them women. Perhaps in the light of his political bent (and probably correctly at least in part), he preferred to attribute their dismal mortality rates to their miserable living standards rather than to their occupation (Villermé 1840). Industrialization in France had occurred largely outside

Table 5.3 Villermé's Data: Including Institutional Mortality in Paris

Arrondissements	Proportion of Tenements Exempted from Taxation	Annual Average Value of Tenement	Deaths in Private Houses		Total of Deaths in the House and at the Hospitals		Cholera
			Period from 1817 to 1821	Period from 1822 to 1826	Period from 1817 to 1821	Period from 1822 to 1826	
		Fr.	1 in	1 in	1 in	1 in	1 in
Montmartre	0.07	425	62	71	38	43	90
Chausssée d'Antin	0.11	604	60	67	43	48	107
Roule, Tuileries	0.11	497	58	66	45	52	82
St. Honoré, Louvre	0.15	328	58	62	33	34	54
Luxembourg, &o	0.19	257	51	61	33	39	17
Porte St. Denis, Temple	0.21	242	54	58	35	38	62
Faubourg St. Denis	0.22	225	53	64	34	42	67
St. Avoie	0.22	217	52	59	35	41	34
Monnaie, Invalides	0.23	285	50	49	36	36	34
Ile St. Louis	0.31	172	44	50	25	30	22
St. Antoine	0.32	172	43	46	25	28	36
Jardin du Roi	0.38	147	43	11	24	26	35
				In all Paris	32	36	

Paris, especially in textile industries. The production of linen from flax had been a delicate process not easily converted into a mechanical procedure. The textiles that did lend themselves to manufacture were cotton, wool, and silk. As industrial production of these cloths rose and flourished, they provided employment for the flax farmers and their workers in the new textile mills, but at very low rates of pay and under factory conditions.

A case has been made for Parent-Duchatelet (1790–1836) as the prime contributor to the new public health movement in France from 1820 until his early death in 1836 (La Berge 1977, 1992). For some aspects of public health, and especially for the effects of physical and occupational environment on workers, abundant evidence from his work backs the claim. He enumerated carefully and freely but in the main descriptively. Thus with respect to epidemiological innovation, in so far as his biographers discuss this aspect of his works, the claim for his primacy seems rather less cogent. Villermé was Parent-Duchatelet's friend and colleague. They served together on the Paris Health Council, in the Royal Academy of Medicine, and as co-founders and editors of the *Annales des Hygiene Publique et Medicine Legale*. With regard to an epidemiological approach, however, on our reading Villermé would seem to have been the greater innovator.

Parent (as his biographer La Berge familiarly speaks of him) relied primarily on direct observation and analysis. He used personal interviews and detailed physical measures to support his observations. In contrast to the liberal Villermé who resisted state action, Parent promoted state and legal action and indeed has been described as a statist. Indefatigable in his investigations and reporting, Parent is famed most for his two-volume study of prostitution in Paris. For this purpose, he visited the hospitals and prisons in which the women spent much time, interviewed physicians who dealt with them, and repeatedly visited the women in brothels and elsewhere at all hours day or night to obtain the details of their biographies and conditions of life.

Subsequently, he investigated with similar intensity the sewers of Paris, which at that time served only as storm drains. Human wastes were deposited in cesspits, and these too came under his scrutiny. His observations and measurements of many aspects of the physical environment once compiled, he conducted tests of possible solutions. As a final step, he offered his recommendations for the reform of the problematic aspects of the city's systems. At least one epidemiological conclusion emerged from his studies of dumps and horse-butcheries. His enumeration of sickness and death in these foul-smelling and seemingly noxious vicinities did not show the excess expected by anti-contagionist miasmatists. No mention is made of miasma theorists either absorbing or seeing a need to refute this result. The dogma was thoroughly entrenched.

References

Ackerknecht EH (1948). Hygiene in France, 1815–1848. *Bull Hist Med* 22:117–155.

Ackerknecht EH (1952). Villermé and Quetelet. *Bull Hist Med* 26:317–329.

Armitage P (1983). Trials and errors: the emergence of clinical statistics. *J R Stat Soc Ser A Gen* 146:321–334.

Barnes DS (2006). *The Great Stink of Paris and the Nineteenth-Century Struggle against Filth and Germs.* Baltimore: Johns Hopkins University Press.

Coleman W (1982). *Death Is a Social Disease: Public Health and Political Economy in Early Industrial France.* Madison, WI: University of Wisconsin Press.

Deutsch A (1949). *The Mentally Ill in America: A History of Their Care and Treatment from Colonial Times.* 2nd ed. New York: Columbia University Press.

Eyler JM (1979). *Victorian Social Medicine: The Ideas and Methods of William Farr.* Baltimore: Johns Hopkins University Press.

Fee E, Brown TM (2006). Freeing the Insane. *Am J Public Health* 96:1743.

Greenwood M (1936). *The Medical Dictator, and Other Biographical Studies.* London: Williams and Norgate.

Harris JC (2003a). Pinel delivering the insane. *Arch Gen Psychiatry* 60:552.

Harris JC (2003b). Pinel orders the chains removed from the insane at bicetre. *Arch Gen Psychiatry* 60:442.

La Berge AF (1977). A. J. B. Parent-Duchatelet: hygienist of Paris, 1821–1836. *Clio Med* 12:279–301.

La Berge AF (1992). *Mission and Method: The Early Nineteenth-Century French Public Health Movement.* Cambridge: Cambridge University Press.

Lazarsfeld, PF (1977). Notes on the history of quantification in sociology—trends, sources and problems. In Kendell MG, Plackett RL, eds. *Studies in the History of Statistics and Probability II.* London: C. Griffin, pp. 213–270.

Lilienfeld AM, Lilienfeld DE (1980). *Foundations of Epidemiology.* 2nd ed. New York: Oxford University Press.

Louis P (1835). *Recherches Sur Les Effets de la Saignee.* Paris: de Mignaret; Transl. provided by MM Curnen. [Table adapted by OL Wade. Adverse reactions to drugs. London. Heinemann, 1970. p. 2].

Louis PCA (1836). *Researches on the Effects of Bloodletting in Some Inflammatory Diseases and on the Influence of Tartarized Antimony and Vesication in Pneumonitis.* Boston: Hilliard, Gray & Company.

Montesquieu (1748). *The Spirit of Laws.* See Montesquieu (2002). Amherst, NY: Prometheus Books.

Morabia A (1996). P. C. A. Louis and the birth of clinical epidemiology. *J Clin Epidemiol* 49:1327–1333.

Quinlan SM (2007). *The Great Nation in Decline: Sex, Modernity, and Health Crises in Revolutionary France C.1750–1850.* Aldershot, England: Ashgate.

Rosen G (1993). *A History of Public Health.* Expanded ed. Baltimore: Johns Hopkins University Press.

Shryock RH (1961). The history of quantification in medical science. In Woolf H, ed. *Quantification*. New York: Bobbs-Merrill.

Susser M (1968). *Community Psychiatry: Epidemiology and Social Themes.* New York: Random House.

Susser M (1987). *Epidemiology, Health & Society: Selected Papers.* New York: Oxford University Press.

Villermé L (1826). Rapport Fait Par M. Villermé, Et Lu a L'Academie Royale De Medicine, Au Nom de la Commission de Statistique, Sur Une Serie de Tableaux Relatifs au Mouvement de la Population dans les Douze Arrondissements Municipaux de la Ville de Paris Pendant les Cinq Annes 1817, 1818, 1819, 1820, et 1821. *Arch Gen Med* 10:216–245.

Villermé L (1840). *Tableau de l'État Physique et Moral des Ouvriers Employés dans les Manufactures de Coton, de Laine et de Soie*. Paris: Edition Jules Renouard et Cie (new ed. 1989 by Etudes et documentations internationales).

Weiner DB (1993). *The Citizen-Patient in Revolutionary and Imperial Paris*. Baltimore: Johns Hopkins University Press.

Weiner DB (1996). *Memoirs on Paris Hospitals*. Canton, MA: Science History Publications.

Wollstonecraft M (1792). *A Vindication of the Rights of Woman*. See Wollstonecraft M (2004). Harmondsworth: Penguin.

6

The British Sanitary Movement: Edwin Chadwick

In Britain during the first half of the nineteenth century and even later, smallpox remained a familiar and still recurring scourge. At the same time, mortal cholera epidemics spreading from the East posed new recurring threats. Their frightening onslaught could not be ignored. As one response, the birth of modern epidemiology may be attributed in part to this stimulus. Quantitative studies of disease and mortality in the tradition of Graunt and Petty provided the scientific foundation for this new field of public health.

In the later eighteenth century, the Industrial Revolution in England had developed momentum unexcelled anywhere. With that revolution, following on the height of the era of mercantile capitalism and its warring powers, the nautical English began to consolidate their gains on home ground. Redistribution of the population accelerated change in the social structure. Workers were needed as factories multiplied. At the same time, the gentry—squires and rural landowners—sought to raise agricultural productivity. They devised new farming methods, enlarged their acreage by enclosing land held and used in common for centuries by the peasant tenants beholden to them, and then simply usurping ownership. Thousands of displaced peasantry flocked to the cities in search of the work then becoming available

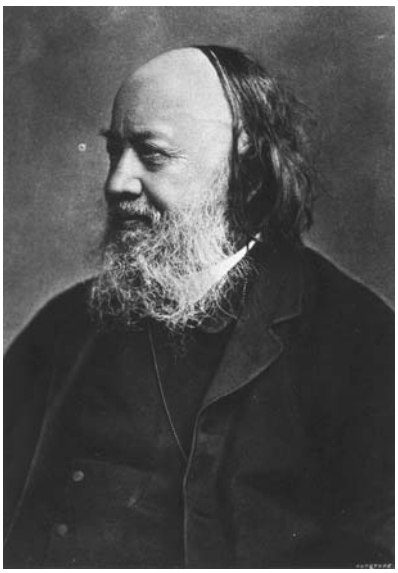


Figure 6.1 Edwin Chadwick (1800–1890). Credit: Courtesy of the National Library of Medicine.

in the new world of factories and industry. There the rural migrants lived in miserable, crowded housing and under abysmal insanitary conditions. To absorb human wastes no outlets but cesspits on the street or under tenement houses were available; every kind of refuse littered the streets. Those who were able to find employment worked long hours for small reward. To secure a living wage often entire families, not excluding young children, would enter the multiplying textile mills. The more hands, the less marginal were their lives. David Livingstone, the intrepid Scottish missionary and African explorer, provides an example. As a young child in the mid-nineteenth century, together with his whole family he worked 14-hour days in the textile mills of Glasgow (Jeal 1973). In these conditions, further exacerbated by the Napoleonic Wars, unrest simmered.

In this changing society Edwin Chadwick (1800–1890), trained as a barrister-at-law, and became an effective and soon a leading social planner (Figure 6.1). As a young man in the 1820s, he concerned himself at first with the problems of poverty and productivity of the poor and the unemployed in the growing cities. The England in which he was growing up was still recovering from both the threat and the actuality of war with Napoleonic France. The Duke of Wellington, fresh from his Waterloo victory in 1815 over Napoleon after the escape from Elba, was the national

hero. Born into the Anglo-Irish aristocracy, Wellington was a high Tory and an influential leader of the long-lasting and austere parliamentary regime, which kept tight rein on the budget and on the social demands of the poor and the needy.[1]

Chadwick, an ideologue in accord with Wellington, shared his essentially conservative concerns. In the aftermath of the Napoleonic Wars and into the 1830s, in the United Kingdom and especially in the manufacturing cities of the North, social and political disturbances were evident everywhere. Chadwick had meanwhile grown firm in the belief that adverse social conditions were at the root of this dangerous political unrest. To eliminate this prime cause of instability was therefore crucial for a rational procedure to restore political peace.

As a relatively young man Chadwick was strongly informed by the technical innovations in epidemiology and statistics of the early French hygienists, and not least by their emphasis on the quality of the social and physical environment. He learned much from Villermé in particular and from Parent-Duchatelet, and corresponded with both.[2] Chadwick's analytic approach, illustrated below, was formulated in the light of close knowledge of the ongoing work in France (Lewis 1952; Chadwick republished, 1965; Finer 1970; La Berge 1988; La Berge 1992; Hamlin 1998).

Chadwick adopted the Utilitarian social engineering approach of his mentor, the influential philosopher Jeremy Bentham.[3] Chadwick had already been the mainspring for the Poor Law Inquiry of 1832. From that base, he proved himself persuasive and forceful enough to provoke the appointment of a Parliamentary Commission into the health and conditions of the poor. In accord with the new law, Parliament established a permanent Poor Law Commission and appointed Chadwick as Secretary.

The rigorous Poor Law of 1834 followed the recommendations of that commission and created workhouses to shelter the homeless poor. In return for such shelter the law imposed harsh conditions, encompassed by a principle described as "less eligibility." The harshness was by calculation sufficient to give inmates an incentive to choose work outside the workhouse rather than poor relief within it. In practice, this application

[1] The Duke of Wellington was Prime Minister 1828–1830, 1834; preceded by the second Earl of Liverpool who served as Prime Minister from 1812–1827. Liverpool was concerned with maintaining social order and opposed to most reforms. When good harvests lowered the price of corn, his government passed the Corn Laws which prohibited imports until prices rebounded, showing a distinct lack of interest in the social demands of the needy.

[2] Although neither of these French scholars believed the miasma theory; in fact, Villermé tested its existence as a cause of disease, and on the basis of his findings, rejected it.

[3] Nevertheless, there is an important difference between Bentham and Chadwick. Bentham aimed to engineer happiness, Chadwick stability.

of utilitarian principles in providing for the poor replaced the prior less systematic but more humanitarian principles. These new emphases on alms and charity undermined the dignity of those subjected to this new approach, and the poor seemingly lost more than they gained. Charles Dickens, in his moving accounts of poverty, used his own experiences as a child in a family unable to make ends meet and always in debt. He graphically described the circumstances of life in the workhouse, orphanage, and prison in *The Pickwick Papers* (1836–1837), *Oliver Twist* (1837–1839), and elsewhere.

As noted above, Chadwick had begun his inquiries into the ways of life of the poor with the same obvious causal assumptions as the French hygienists: poverty caused disease, disease was costly to society, and to prevent disease one must alleviate poverty. Indeed, the view that poverty was a major cause of ill-health had been well argued in the eighteenth century by British as well as by French physicians who worked among the poor. The young Friedrich Engels (1820–1895) in his book of 1844 drew the same conclusion as the French writers. This work described working-class life in the adjoining northern industrial cities of Manchester, Salford, and Stockport. Engels was the Manchester agent for his family's textile business in Germany. He also continued as regular correspondent and ally of Karl Marx, who after the liberal uprising that swept across Europe in 1848 was perforce living and writing among many other such European exiles in London.

Engels had collected and studied all the statistics available for his illuminating account of the conditions of life among the factory workers (described in Chapter 14). He was well-acquainted with their conditions; his mistress in Manchester was herself a factory worker (Engels 1844/1993). Unsurprisingly, his social and political perspective was diametrically opposed to Chadwick's position then and later. Yet one of Engels' main concerns was in fact to rebut Malthus's notion, proposed in 1798, that the rapid population increase would eventually outrun food supply.[4] Instead, Engels argued that advancing technology and productivity would meet the need, and more assuredly so in the event of a successful workers revolution.

Chadwick too had been a protagonist of the need to alleviate the ills of poverty though also opposing the views of Malthus on poverty as the consequence of population growth. Chadwick's work on reforming the existing poor laws was based on the assumption that political instability

[4] "But though the principle of population cannot absolutely produce a famine, it prepares the way for one; and by frequently obliging the lower classes of people to subsist nearly on the smallest quantity of food that will support life, turns even a slight deficiency from the failure of the seasons into a severe dearth; and may be fairly said, therefore, to be one of the principal causes of famine" Bk.II,Ch.XIII in paragraph II.XIII.18 (Malthus 1826).

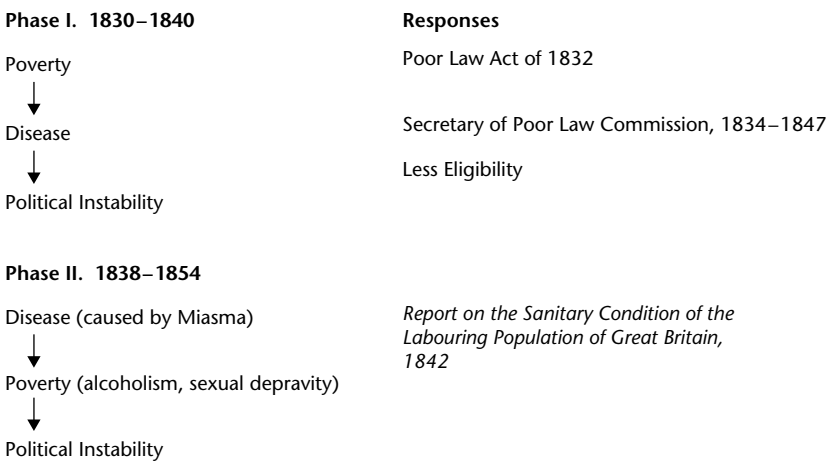


Figure 6.2 Chadwick's models of disease causation by phase.

could be reduced by alleviating poverty. Poor Law reform would in turn reduce disease, which initially he saw as a root cause of instability. Yet, at some unknown point in time, he had surprisingly and radically changed his assumptions about the causal chain leading from poverty to disease and sickness, and thus to ensuing political unrest (see Figure 6.2). The motivation and the reasoning behind this sea change remain somewhat obscure. He simply reversed the ordering of the elements of the causal chain that the French researchers had worked mightily to demonstrate, and that so markedly influenced his early thought.[5] He now attributed poverty to ill-health rather than the reverse (he never recognized the reciprocal interaction between the two conditions). His new stance became evident in his *Report on an Inquiry into the Sanitary Condition of the Labouring Population of Great Britain of 1842*. In the light of the revised causal construct, Chadwick was brought to contest with vigor his former more liberal view that poverty was the source of many ills.

What followed from the changed assumptions was clear enough, however. Instability underlying political unrest remained Chadwick's main political concern. He still held that poverty was a proximate cause of instability and

[5] One suggested reason for this change in Chadwick's thinking is that the 1832 Poor Law system he had created did not produce the hoped-for economic advantages from inducing the poor to find work rather than merely sheltering in the harsh conditions of the new workhouses. Any advantage in costs turned out to be outweighed by the costs of providing medical care.

ensuing unrest. His argument was about the time order of the important antecedents of instability in the causal chain. This turnaround ran contrary to the common medical understanding of the time. In his conception of the chain of events, however, the immediate cause of ill health still conformed to the common understanding in one important respect. This was its attribution to the miasma which, in Sydenham's terms, was the product of foul odoriferous emanations from soil, air, and water.

At this point, Chadwick became the chief agent of sanitary reform in Britain. He shared the unequivocal belief in the miasma theory of many at the time and addressed the question of disease and sanitation numerically. His Sanitary Report of 1842 is profusely illustrated with illuminating tables as in Figure 6.3.

The new formulation also modified the historic standard theory in another way. Miasma, it was proposed by Chadwick's General Board of Health of 1848–1853, need not be only atmospheric and climatic; it could also be local. This idea of local miasma aimed to meet and accommodate the etiological challenge posed by localized point epidemics with a common source (Vinten-Johansen et al. 2003). As noted earlier, such outbreaks of cholera and typhoid had already come under investigation by John Snow and William Budd respectively, both of whom reached a contagionist conclusion. In Chadwick's revised idea of localized miasma, the predisposing factors were still personal, although behavioral and not humoral as in Sydenham's concept.

The particular causal attributions were the allegedly widespread opportunities afforded the poor for alcoholism and avid sexuality, and their associated penchant for debauchery and dissipation. The relevant atmospheric changes were local, generated wherever masses of refuse fouled the streets and where human wastes were gathered in cesspits. For the political unrest among the poor all this supposedly generated, therefore, the ultimate cure was to break the cycle leading from the antecedent miasma, first to disease, and then to the ensuing poverty from which the discontent and unrest followed.

In summary, Chadwick's overriding purpose was to sustain political order. To this end, he arrived at a revolutionary concept that entailed a grand scheme of social engineering. His remedy for disease was to eliminate the sources of miasma arising from the execrable habits and filthy living conditions habitual to the poor. With the help primarily of engineers, and of one or two enlightened doctors such as Southwood Smith and Shuttleworth Kay, he devised a novel and ingenious system that would be installed in all cities. Street refuse was to be routinely and regularly collected to dispel the miasma arising from it. Human wastes and storm water in the built-up cities were to be cleared not by cesspits but by drainage.

Counties	Number of Deaths During the Year Ended 31st December, 1838, from…										Proportion of Deaths from the preceding Causes in every 1,000 of the Population, 1841	Proportion of Death from all Causes if Mortality in every 1,000 of the Population, 1841
	1 Epidemic, Endemic, and Contagious Diseases				2 Diseases of Respiratory Organs			3 Diseases of Brain Nerves, and Senses	4 Diseases of Digestive Organs	Total Deaths from four preceding Classes of Diseases		
	Fever: Typhus, Scarlatina	Smallpox	Measles	Hooping Cough	Consumption	Pneumonia	All other Classes					
ENGLAND												
Bedford - -	155	75	40	66	457	97	57	304	131	1,382	13	22
Berks - -	204	288	21	86	739	231	162	467	201	2,399	15	25
Bucks - -	256	85	61	27	575	131	61	348	152	1,696	11	19
Cambridge - -	231	136	57	90	686	156	70	318	189	1,933	12	21
Chester - -	592	279	178	87	1,742	366	345	1,442	421	5,452	14	21
Cornwall - -	443	135	168	491	1,270	342	124	631	228	3,832	11	18
Cumberland - -	165	188	11	83	562	75	142	278	169	1,673	9	21
Derby - -	394	77	79	71	905	200	205	777	268	2,976	11	18
Devon - -	615	460	287	312	1,649	564	298	1,237	471	5,893	11	18
Dorset - -	137	255	80	58	571	146	106	380	1 59	1,892	11	19
Durham - -	347	316	139	304	1,007	362	207	1,138	274	4,094	13	21
Essex - -	417	460	83	163	1,250	276	234	782	268	3,933	11	19
Gloucester - -	352	457	440	244	1,395	578	476	1,142	510	5,594	13	20
Hereford - -	84	83	17	36	333	56	57	238	62	966	8	18
Hertford - -	160	116	45	48	620	107	90	453	155	1,794	11	20
Huntingdon - -	61	18	1	17	216	45	42	140	72	612	10	18

Figure 6.3 Example of a table from Chadwick's Sanitary Report of 1842.

The Manner and Matter of Chadwick's Sanitary Report

In the Sanitary Report of 1842, Chadwick augmented his forceful advocacy with data, which he used well by the standards of policy debate of his times. Below are set out some of the illustrative tabulations in the Sanitary Report. These gain in cogency as the report proceeds (Table 6.1).

The initial condition of the population of the town of Wisbech, in Cambridgeshire, is first described by mortality rates before exposure to the experience of living in an area with drainage, and again by rates subsequent to drainage. Drainage is thus the independent variable postulated as cause, deaths the dependent variable for outcome. Table 6.2 shows the sophistication of the research design.

Here the simple before-and-after control is supplemented by contemporaneous comparison, at three parallel points in time, with a similar area that remained undrained. These comparisons controlled for general historical factors that might have brought about an overall decline in mortality, and protected better against wrongly crediting drainage with the improvement. Another comparison from Chadwick's report (Table 6.3) replaces

Table 6.1 Baptisms, Marriages and Burials in Wesbech

	Baptisms	Burials	Population in 1801
1796 to 1805	1627	1535	4710
1806 to 1815	1654	1313	5209
1816 to 1825	2165	1390	6515

"Drainage, embankments, engines, and enclosures have given stability to the soil (which in its nature is as rich as the Delta of Egypt) as well as salubrity to the air. These were very considerable improvements, though carried on at a great expense, have at last turned to a double account, but in reclaiming much ground and improving the rest, and in contributing to the healthiness of the inhabitants.

…

"**In the first of the three periods the mortality was 1 in 31; in the second, 1 in 40; in the third, 1 in 47**.…These figures clearly show that the mortality has wonderfully diminished in the last half century, and who can doubt but that the increased salubrity of the fens produced by drainage is the chief cause of the improvement." (Chadwick 1842)

Table 6.2 Comparing Bungay and Beccles

	Beccles	Bungay
Between the years 1811 and 1821	1 in 67	1 in 69
Between the years 1821 and 1831	1 in 72	1 in 67
Between the years 1831 and 1841	1 in 71	1 in 59

"You will therefore see that the rate of mortality has gradually diminished in Beccles since it has been drained, whilst in Bungay, notwithstanding its larger proportion of rural population, it has considerably increased." (Chadwick 1842)

Table 6.3 Chadwick's Table on Sickness in Male Prisoners

	Average Annual Sickness of Male Prisoners in the			Laborers and Operatives					
	No. 1 Glasgow Prison	No. 2 Edinburgh Prison	No. 3 Salford Prison	No. 4 Employed in East-India Company's Warehouses	No. 5 Average duration of sickness per annum of every person employed in cotton factories of Lancashire	No. 6 Males of families in wynds of Edinburgh		No. 7 Average annual sickness of members of benefit societies in Scotland	No. 8 Average annual sickness of provident portion of working classes throughout Great Britain, according to the experience of Mr. Finlaison
	Days and Decimals	Days and Decimals	Days and Decimals	Days and Decimals	Days and Decimals	Days and Decimals	Years of Age	Days and Decimals	Days and Decimals
Age									
Under 16 years	—	—	—	—	—	3.5	—	—	—
16–21	3.05	4.01	3.10	4.02	4.42	2.3	18	2.5	5.18
21–26	1.83	2.04	1.64	5.40	4.91	5.1	23	3.8	6.75
31–36	2.65	2.33	2.72	4.49	6.88	11.0	28	4.6	6.78
36–41	2.83	3.10	2.63	4.55	3.85	8.3	33	5.6	6.33
41–46	9.00	5.10	.85	5.57	4.13	4.1	38	6.2	7.86
46–51	.49	2.75	.51	5.18	5.09	15.1	43	8.8	9.02
51–56	—	—	—	5.43	7.18	30.0	48	9.1	11.76
56–61	—	—	—	6.80	3.47	16.2	53	14.8	16.77
61–66	—	—	—	7.21	12.68	30.4	58	17.8	23.57
66–71	—	—	—	10.24	—	42.7	63	20.0	33.22
71–76	—	—	—	9.93	—	64.2	68	36.0	61.22
76–81	—	—	—	10.60	—	41.0	73	38.6	101.44
				12.67	—	83.6	78	70.9	164.72

Source: Chadwick (1842). 1965 Edinburgh University Press edition, p. 280.

mortality with morbidity, a refinement of the dependent variable and outcome measure.

Prison reforms in the late nineteenth century included improvements in ventilation, drainage, and sanitation. Here, Chadwick was advancing the remarkable claim that the reforms had made the prison environment superior to that of the working population. And indeed, at first sight the tabulated prisoner sickness rates compare not unfavorably with others. Notably, the membership of Benefit Societies, which provided access to attendant sickness insurance, tended toward better-off workers. Chadwick's commentary does add a caveat (in response to criticism, which he invariably resisted to the utmost) about small numbers and the inducements and constraints in sickness reporting. William Farr, as we shall see later, properly poured scorn on a result that would recommend prisons for salubrity. Any epidemiologist might ask how large the effect might be if corrections were made for briefer durations of stay and greater turnover in prisons. Both went unreported in Chadwick's comparisons with Benefit Society members. Since the latter were known to be workers usually in longer-term employment, and since the prisons were filled with petty short-term offenders, the effect of the distorted comparison could be substantial.

Chadwick had purposefully deployed the data in the Sanitary Report with the endorsement of the Sanitary Commission he had guided or, one might better say, steered to its conclusions. He argued from his data to support his massive revolutionary scheme for a national sanitary system. The formidable did not daunt Chadwick and one may fairly attribute to him construction of a system for disposing of human wastes that is now familiar in the everyday experience of any developed society. Circuits for fresh water piped in for domestic use and for used water and sewage piped out for disposal, were to be strictly separated in closed systems and independent of each other.

In Paris, Parent-Duchatelet's sewers had removed only liquid waste; solid wastes were disposed of in cesspits and collected by manual labor to serve as fertilizer. This was the case in many large cities in Europe through most of the nineteenth century. At John Snow's instigation and with John Simon's subsequent executive action, London was better protected in the later part of the century (also see Chapter 8). Chadwick's sewers as modified by engineers were an advance. Both solid and liquid wastes were first collected in cesspools. The water to carry detritus away from the cesspools to sewerage farms was to be supplied at enhanced velocity, to be led off to sewerage farms and disposed of advantageously and even profitably as fertilizer (Finer 1970; Hamlin 1998). In Europe, at this time and through much of the nineteenth century, Sydenham's theory of miasma and disease remained the dominant causal doctrine, despite the work of Pasteur, Budd, Semmelweiss, Lister, Snow, Virchow and others. Chadwick's *Report for the Sanitary Commission*

of 1842 rested on his newfound modifications of that causal sequence—in which poverty was the result of disease engendered by local or general miasma—and guided the organization and interpretation of the mass of data collected. As secretary he had dominated the inquiry with dogmatic force (Hamlin 1998). Several leading medical men of the time brought to bear long-standing standard but contrary views about the ill effects of poverty on health, together with much contrary evidence. Nonetheless, Chadwick was not moved. Later, William Farr as Chief Statistical Officer countered with his own cogent analyses of the relationship of mortality with poverty in the national vital statistics. Still Chadwick fiercely fought off the threat that any new results might pose to his theory of miasma and poverty (Lambert 1963; Susser and Adelstein 1975; Eyler 1979; Hamlin 1998). Chadwick was always combative in defense of his theories, regardless of current medical understanding, and even of technically superior analyses that challenged his own positions. He was quite ready to dispute Farr's adoption of a new cause of death classification for the national vital statistics register. This added zymotic disease (a classification borrowed from the German chemist Liebig), which implied fermentation and the production of inanimate and poisonous microparticulates each causing specific disorders with characteristic expression. In Farr's classification zymotic disease could be epidemic, endemic, or contagious. Other diseases in the classification also were thought not to conform; they were presumed to multiply within the body of an affected person and then be transmissible. That is, it comprised largely those conditions now known to follow specific infections. The classification conformed to convention in that disease arising from miasmata was not transmissible person to person. Farr had offered a favorable view of the arguments of Jakob Henle's and others for the existence of infectious disease and the possibility of proving the case (Henle 1840; Susser 1994a, 1994b).

And yet for years to come, Farr did not himself entirely abandon the idea of miasma. Had he done so, he could hardly have maintained for long his amicable association either with the forceful Florence Nightingale (1820–1910), or his somewhat less intimate association with her consistent ally, Edwin Chadwick. Thus, when Farr in the 1850s merely went so far as to introduce the zymotic classification in his national registry for classifying disease, Nightingale upbraided him for the transgression from miasma doctrine. She and Chadwick dominated the Sanitary Movement—along with many in the medical profession—and both vehemently rejected the developing science of the Germ Theory. Near the close of the nineteenth century, both Chadwick and Nightingale died unrepentant believers in miasma and dismissive of contagion. In an often quoted statement, Florence Nightingale (Figure 6.4) wrote: "For diseases, as all experiences show, are adjectives not noun substantives. The specific disease doctrine is the grand refuge of weak uncultured

Figure 6.4 Florence Nightingale (1820–1910). Credit: Courtesy of the National Library of Medicine.

unstable minds, such as now rule the medical profession. There are no specific diseases: there are specific disease conditions" (Dubos 1965, p. 249).

Despite Chadwick's lack of comprehension of scientific developments and his supreme confidence in his own judgment, the closed sanitary system he devised, with water and sewage circulating entirely separately, had considerable beneficial health effects. These were not immediately apparent; indeed, some were untoward and not benign. In the beginning, as we shall see, water supplied from the Thames and also from other rivers in other cities, was not always drawn only from outlets upstream from sewerage discharges. This particular problem was already understood and noted by John Snow toward the mid-nineteenth century. Before the authorities were moved to act, however, a painful interval passed while Snow's work slowly gained currency by the force of the existing circumstances. Sanitary experts undoubtedly recognized the need for official public operation and surveillance of water supply systems. But in England in the mid-nineteenth century, strong forces protected private interests and resisted government intervention. The transfer from commercial to public hands in London had to wait until the twentieth century had begun.

Thomas McKeown (1912–1988) was a leading and theoretically fertile British epidemiologist in the post–World War II period. In a pioneering

analysis of British mortality statistics through the nineteenth century, he detected no beneficial effects of the existing sanitary system on national mortality. By exclusion, for lack of other apparent causes, McKeown attributed the decline to improved nutrition (McKeown 1976). Doubtless nutrition contributed, but little direct evidence exists to show its effects, other than assumptions based on national food production. McKeown's analysis is perhaps misleading in one respect (Susser 1993; Szreter 2002). McKeown did not allow for the lag in implementing the sanitary system in cities and towns before it covered the whole country (also see Chapter 15 for more on McKeown and more recent critiques). After the law was put in place (Hamlin 1998), installation of the whole sanitary system seems not to have reached completion for something like 25 years. By then, a steady decline in national mortality rates had begun. No less important, at the outset of the newly installed sanitation system, was the lag in understanding of the pathways of infection during the slow transition from miasma to germ theory. The blinders of false confidence in theory sustained the mistaken idea of miasma and blocked recognition of obvious bacterial contamination. Nothing makes clearer the effects of such misplaced confidence in a theory than the difficulties experienced by John Snow in gaining acceptance for his work. Snow did not stop short at the theoretical postulates he derived from his reading of the literature. As noted later, in his several investigations he went on to demonstrate the likelihood of the transmission of infectious disease by contaminated water (Vinten-Johansen et al. 2003).

Thus in the final quarter of the nineteenth century, mortality rates did at last begin the steady decline that has continued since. In particular, mortality from enteric infections borne by contaminated water—cholera and typhoid in particular—began to decline. So too did the dysenteries and other enteric disorders carried by flies from street refuse, cesspools, cesspits, and animal ordure. Among these, not the least were the annually recurring and mortal epidemics of summer diarrhea of infants. By the end of World War 1 another major change—when in many cities motor vehicles displaced horses and brought about great reductions both in droppings of fly-breeding manure and in the carrier population of flies—undoubtedly contributed to the decline of enteric infections. Nonetheless, it is safe to say that before the new sanitary system was in full function, on the health front no major intervention other than vaccination for smallpox had yet had any notable effect on population mortality (Davey Smith and Lynch 2004).

Further large-scale successes waited on the discoveries of the late nineteenth century in bacteriology, immunology, and epidemiology and on the surprisingly deferred acceptance of germ theory. Thomas McKeown's analysis of mortality rates in Britain up to the 1970s cannot of course be ignored. The results led him to a healthy skepticism about the effects of any kind

of mass intervention on these rates. He could perceive no effects, even in the era of antibiotics and medical care. By now, it is plain that McKeown was mistaken, in the main probably because he died before the direction of the accumulating data could become entirely clear. Extensive vaccination programs and the advent of chemical and antibiotic therapeutics have undoubtedly curbed mortality from both epidemic and endemic infections. Data accumulating on scales large as well as small support this conclusion (Susser 1993; Mackenbach 1996).

References

Chadwick E (1842; republished 1965). *Report on the Sanitary Condition of the Labouring Population of Gt. Britain*. Edinburgh: University Press.

Davey-Smith G, Lynch J (2004). Commentary: social capital, social epidemiology and disease aetiology. *Int J Epidemiol* 33:691–700.

Dubos RJ (1965). *Man Adapting*. New Haven: Yale University Press.

Engels F (1844; republished, 1993). *The Condition of the Working Class in England*. McLellan D, ed. Oxford, England: Oxford University Press.

Eyler JM (1979). *Victorian Social Medicine: The Ideas and Methods of William Farr*. Baltimore: Johns Hopkins University Press.

Finer SE (1970). *The Life and Times of Sir Edwin Chadwick*. New York: Barnes & Noble.

Hamlin C (1998). *Public Health and Social Justice in the Age of Chadwick: Britain, 1800–1854*. Cambridge: Cambridge University Press.

Henle J (1840). *On Miasmata and Contagie*. See 1938 transl. by George Rosen. Baltimore: The Johns Hopkins Press.

Jeal T (1973). *Livingstone*. New York: Putnam.

La Berge AF (1988). Edwin Chadwick and the French connection. *Bull Hist Med* 62:23–41.

La Berge AF (1992). *Mission and Method: The Early Nineteenth-Century French Public Health Movement*. Cambridge: Cambridge University Press.

Lambert R (1963). *Sir John Simon, 1816–1904 and English Social Administration*. London: MacGibbon & Kee.

Lewis RA (1952). *Edwin Chadwick and the Public Health Movement, 1832–1854*. London: Longmans, Green.

Mackenbach JP (1996). The contribution of medical care to mortality decline: McKeown revisited. *J Clin Epidemiol* 49:1207–1213.

Malthus TR (1826). *An Essay on the Principle of Population: A View of Its Past and Present Effects on Human Happiness; With an Inquiry into Our Prospects Respecting the Future Removal or Mitigation of the Evils Which It Occasions*. London: John Murray.

McKeown T (1976). *The Modern Rise of Population*. London: Edward Arnold.

Susser M (1993). Health as a human right: an epidemiologist's perspective on the public health. *Am J Public Health* 83:418–426.

Susser M (1994a). The logic in ecological: I. The logic of analysis. *Am J Public Health* 84:825–829.

Susser M (1994b). The logic in ecological: II. The logic of design. *Am J Public Health* 84:830–835.

Susser M, Adelstein A, eds. (1975). *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr.* Metuchen: The Scarecrow Press, Inc.

Szreter S (2002). Rethinking McKeown: the relationship between public health and social change. *Am J Public Health* 92:722–725.

Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip MR (2003). *Cholera, Chloroform and the Science of Medicine: A Life of John Snow.* New York: Oxford University Press.

7

Vital Statistics: William Farr and the Creation of a System

In the latter half of the nineteenth century, both the concept of environment and the numerical approach to the understanding of related public health problems were firmly entrenched. Yet for much of the century, most British epidemiologists and many elsewhere (with a few remarkable exceptions like William Budd and John Snow) were guided largely by Sydenham's theory of the interaction of miasmata with the "epidemic constitution" of seasons. Accordingly, they had followed a general line of research into environmental effects. Among them (if chiefly in their earlier years) were William Farr and later John Simon, and a number of others among the brilliant group of physicians Simon had selected and deployed as epidemiologists (Simon 1890; Snow 1936; Lambert 1963; Brockington 1965; Susser and Adelstein 1975b).

William Farr (1807–1883)

William Farr merits more than the passing comment he has so far received in this account. Indeed, he can be properly assigned a major role as a founder of epidemiology in its modern analytic form. Born into humble circumstances in the west of England, he was fortunate in being informally adopted

at a young age by a local gentleman, who saw him as a youth with aptitudes worthy of support. He acquired his education—wide-ranging as it was in the classics, philosophy, and mathematics—largely in reading and private study.

At the age of 19, he began to study medicine under the joint guidance of a local medical practitioner and a surgeon at the local infirmary. Three years later, in 1829, aided by a small legacy from his benefactor, Farr left England to further his medical studies at the renowned University of Paris. Farr cultivated his mathematical bent at the lectures of Pierre Louis, for him probably the most important of the several luminaries at the podium from whom he learned. On his return to London two years later to enter practice, Farr began to write on statistical topics. He had the encouragement of Thomas Wakley, editor of the *Lancet* which he had founded in 1823. In 1837, Farr contributed a major article, "Vital Statistics," to McCullough's *A Statistical Account of the British Empire* (Farr et al. 1837/1974).

By the age of 31, Farr's writings and talent had made him the obvious choice for the post of Compiler of Statistics in the General Registry Office (GRO) newly created in 1839. The office was to maintain a national system, essentially Farr's creation, of the vital statistics of births, marriages, and deaths (Susser and Adelstein 1975a; Susser and Adelstein 1975b; Eyler 1979).

The British system of vital statistics is generally acknowledged as of outstanding historical value, even though initiated some time later than others in Europe. One can say without exaggeration that the system is Farr's. Excepting the national census initiated in 1811, he alone devised, developed, and maintained the system. The data on specific causes of death, which from the outset Farr collected and analyzed, were a key element in its utility both then and now. The specification and refinement of variables thus begun has been a basic process in the scientific advance of epidemiology. In population studies, specification of causes opened an entirely new field of investigation and understanding. Farr contributed many analyses of the distributions of causes of death, in which he related specific endemic and epidemic manifestations to environmental and other independent variables. During his 40-year period in office, he became an influential guide in the international conferences held to plan and devise systems of vital statistics.

One interesting result on cholera, however, illustrates the hazards to which epidemiologists are still subject in analyzing mass data. The example shows how an important result, reported by an epidemiologist as great and gifted as Farr, was later refuted by John Snow's work.

In the analysis at issue, Farr set out a direct ecological relationship in the form similarly presented by Chadwick. As the independent study variable under test, he related the hypothesized environmental cause of altitude to the dependent study variable of cholera death rates. Both the independent

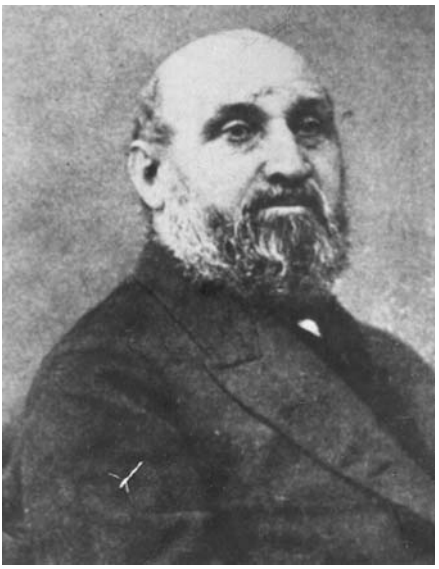


Figure 7.1 William Farr (1807–1883) in 1870. Credit: Courtesy of the National Library of Medicine.

and the dependent variables were characteristics of administrative areas, and are thereby units at the same level of organization. From this analysis Farr concluded quite reasonably but without qualification, that "elevation of habitation reduces the effects of cholera to insignificance."

The relationship Farr observed was undoubtedly present, but his interpretation was not as assured as he had then concluded. The result he reported held for a local but not a general phenomenon. In its lower reaches the Thames, as it reached and flowed through London, became increasingly contaminated by sewage. The contaminated water supply that caused most of the cholera in London was thereby brought into association with low altitude. Farr had developed an appropriately testable hypothesis, but he had not searched far enough for instances to contradict it.[1] In later pages, we shall consider Farr's observation at greater length.

Francis Bacon, the first Englishman and perhaps the first European to develop and systematize the methods of scientific inference, wrote: *"But the induction which will be useful for the discovery and proof of sciences and arts, should separate out a nature, by appropriate rejections and exclusions;*

[1] Snow and Farr's work on the London cholera outbreak has been considered by many scholars. See Susser and Adelstein 1975b and Frost 1936, and for recent elucidations, Eyler 2001; Halliday 2002; Vinten-Johansen et al. 2003; Bingham et al. 2004.

and then, after as many negatives are required, conclude on the affirmatives (Bacon 1620, First Book of Aphorisms).

In other words, Bacon posited that the most cogent test of a hypothesis available to scientists is the rigorous attempt at disproof. When only one hypothesis survives, it stands as the best approximation to truth until it too is challenged by new alternatives. If one disregards the grounds for Chadwick's ideas that the twentieth century has found erroneous, one must yet conclude that Chadwick the dogmatist was right in his assertions that sanitation underlay high mortality rates. In this one instance, one must allow that Farr, the sophisticated analyst, was wrong in attributing high mortality from cholera to elevation. Although Farr's analysis was much the technically superior, given only knowledge then current from observations of ecological units, the hypotheses each proposed were equally legitimate. Yet as long as the unit of observation was not of individuals but of defined populations, neither could go further in establishing direct and immediate causes of individual deaths. Ironically it was left to the inimitable John Snow—in his studies of epidemic cholera mortality data reported in 1849 that Farr had himself assembled and provided—to clarify these contradictory interpretations (Snow 1855; Frost 1936). In this work Snow sagely reflects on the evidence he collected of the local outbreaks and successive waves of epidemic cholera that in the nineteenth century devastated and terrified England.

Perhaps Farr's greatest accomplishment was to institute and write the *Annual Reports of the Registrar General,* the first appearing in 1839. These described and analyzed the health status of the country in terms of the database he had himself devised to enumerate births, marriages, and deaths. These continued uninterrupted until his last report in 1879 and are still maintained. That work ensures his place as a major innovator in both descriptive demography and analytic epidemiology. From the outset Farr had a clear vision of what he was about. Much of that task was presaged in his first *Annual Report.* His genius lies precisely in a range of strengths: Thus he conceptualized the nature and the means of analyzing the daunting problems of public health as population problems; he had the capacity to forge the statistical means to approach those problems; and finally he applied those means unremittingly on a national scale. His aim was to use the national data system he created both to describe and to explain (Susser and Adelstein 1975b). Farr did more than create a system of national statistics. During his 40-year incumbency, he greatly enhanced their value and their meaningful use across what was then the developed world. He derived denominators by ingenious resort to the antecedent decennial census. This device enabled him to calculate rates and trends and make comparisons. His standard classification of disease by diagnosis provided the basis for comparisons of mortality trends at local,

national, and ultimately, at the international level.[2] As needed, Farr called on sophisticated methods of analysis including life tables, analysis of cohorts, and statistical modeling. Among many others, Edwin Chadwick, Friedrich Engels, Florence Nightingale, and John Simon, each aiming to promote a personal cause at national level, made use of and indeed depended on the cogent data Farr's office provided.

Farr did not shrink from addressing issues of social and even of political importance, as when he entered the Malthusian fray. On the population question an acute dispute of the times raged between Friedrich Engels, the closest ally of Karl Marx, and the Malthusians (Susser 1974) (see Chapter 6). In a demographic critique, Farr dismissed the Malthusian argument that rising fertility would outrun food production and impoverish the nation. In 1844 he wrote: "In order, therefore, to understand the rate of increase by birth, it will be necessary to inquire how many persons are married, by how many marriage is foregone, and how long marriage is delayed after puberty" (Susser and Adelstein 1975b).

In reporting on the 1851 census, Farr illustrated further the play of demographic trends: "The numbers, and consequently the increase or decrease of the people in a civilized country, depend upon the age at marriage and the age of the parents when the children are born—the numbers who marry, the fertility of the marriages—the duration of life—the activity of the migration flowing in and out of the country. These facts more or less influence each other and in the present state of statistical observation, the precise effect of change in any of them involving others cannot be determined. It will be sufficient to indicate the extent of a change in each element, while the others remain constant" (Graham et al. 1852, The Census of Great Britain, 1851).

Farr also contributed his statistical skills to what has been described as a therapeutic revolution in the treatment of mental disorder. To remedy the conditions of those confined in asylums, as in the barbarous Hogarthian spectacle provided for the amusement of onlookers at the Bethlem Asylum in London, Dr. John Conolly, superintendent of the Hanwell Asylum just north of London, joined the movement for *moral treatment* of the insane originated by the Tukes in Yorkshire in the late eighteenth century.[3] We have noted above the growing overcrowding that followed the urban migration of

[2] Farr's work was the impetus for the international classifications of disease initiated in the second half of the nineteenth century; the first International Statistical Congress (1853) called upon Farr and Marc D'Espine, of Geneva, to create a classification of causes of death that could be universally applied.

[3] Conolly published between 1830 and 1856; (republished, Conolly and Hunter 1964; Conolly 1968; Conolly and Middlesex Lunatic Asylum at Hanwell 1973).

displaced rural peasantry in the early nineteenth century. Displaced poverty-stricken masses filled the workhouses and the streets of London and other cities. Many among them were mentally afflicted and incarcerated in large asylums. Farr, possibly at Conolly's request, analyzed patient mortality in a variety of institutions for the insane, including the Hanwell asylum. A sentence excerpted from his report of these results to the Statistical Society in 1851 concludes that "annual mortality of both male and female paupers in the licensed homes was nearly twice as great as mortality of paupers at Hanwell, and twice as great as the mortality of other lunatics in the licensed houses" (Susser and Adelstein 1975b).

The belief was then current, as has been argued by some even in recent times, that the incidence of mental disorder was increasing. Farr observed in passing that an increase in numbers in institutions for the insane was no indication of an increased frequency, since more humane treatment prolonged the lives of inmates and increased the numbers of survivors. He constructed a nosometric table. This was precisely the cohort life table now so heavily used to analyze matters such as mental hospital usage, or duration of survival from cancer, or shifts over generations in a variety of disorders. Perspicaciously Farr noted that from such a table "the lives of lunatics can be insured." He saw the significance for treatment as if with contemporary eyes: "The table is an instrument by which the effects of treatment on the mortality, the number of recoveries—and the duration of all diseases—can be accurately measured. It enables us to compare two or three different plans of treatment, and to determine their effects upon the principal results at which all medical treatment aims—the reduction of the mortality, and of the duration of the disease" (Susser and Adelstein 1975b).

After completing 40 years of highly meritorious service in the office Farr himself had essentially created, he sought and had surely embellished and earned the position of Registrar General upon the retirement of the long-term incumbent Major Graham. On not achieving this goal, Farr resigned his office. Within a year after his retirement in 1879, unhappily his mental faculties began to deteriorate, and he died in 1883 at age 76.

We have dubbed this long and full nineteenth-century period in the history of epidemiology the *Era of Sanitary Statistics*. Miasma was the dominant guiding causal concept of the era. Analysts sought out the undue clustering of morbidity and mortality as indicators of the assumed poisoning by miasmata. For that, the appropriate remedial and preventive action was drainage, sewage, and sanitation (Winslow 1943; Ackerknecht 1948). During the last half of the century, advances in the revolutionary new subjects of cellular pathology and of microbiology were pioneered respectively by Theodor Schwann and Rudolf Virchow, and by Louis Pasteur, Edwin Klebs, and Robert Koch. They were joined by rapidly growing numbers of

other biologically oriented scientists. The resulting steady scientific progress at last began to undermine miasma theory.

Before the end of the nineteenth century, we shall see epidemiology vindicate the theoretical insights of Ludwik Fleck in his study of the history of syphilis (Fleck 1935) and, more recently and specifically, those of Thomas Kuhn (who takes physics rather than medicine as the exemplary science for his theoretical stance) (Kuhn 1962, 1970). That is to say, shifts in dominant concepts (*paradigms* in Kuhn's terms) from one era to another entail radical parallel shifts—*scientific revolutions*—in the interpretations of what the same set of phenomena mean to the observer.

The era of sanitary statistics, guided by a miasma theory of broad, nonspecific, ecological causes, was superceded by germ theory. The new theory guided a new era of specific, organically based causes each giving rise to specific identifiable outcomes of disease. The scientific shift in epidemiology and consequently in public health and medicine was indeed revolutionary, even tectonic. This shift in microcosmic theory, as with the macrocosmic theories of Copernicus, Newton, and Einstein, was revolutionary and seemingly rapid. Underlying the apparently sudden demise of miasma as an idea, however, we shall see that the ferment of a theory of contagion had been brewing through centuries.

References

Ackerknecht EH (1948). Anticontagionism between 1821 and 1867. *Bull Hist Med* 22:562–693.

Bacon F (1620). The New Organon. Jardine L and Silverthorne M, eds. (2000). Cambridge, UK: Cambridge University Press. See The First Book of Aphorisms.

Bingham P, Verlander NQ, Cheal MJ (2004). John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. *Public Health* 118:387–394.

Brockington F (1965). *Public Health in the Nineteenth Century.* Edinburgh: E. & S. Livingstone.

Conolly J (republished 1968). *The Construction and Government of Lunatic Asylums and Hospitals for the Insane.* 1st ed. London: Dawsons.

Conolly J, Hunter RA (republished 1964). *An Inquiry Concerning the Indications of Insanity with Suggestions for the Better Protection and Care of the Insane.* No. 4. London: Dawsons of Pall Mall.

Conolly J, Middlesex Lunatic Asylum at Hanwell (republished 1973). *Treatment of the Insane Without Mechanical Restraints.* Reprinted. Folkestone: Dawsons.

Eyler JM (2001). The Changing Assessments of John Snow's and William Farr's Cholera Studies. *Soz. Praventiv Med.* 46:225–232.

Eyler JM (1979). *Victorian Social Medicine: The Ideas and Methods of William Farr.* Baltimore: Johns Hopkins University Press.

Farr W, Ratcliffe H, McCulloch JR (1837/1974). *Mortality in Mid 19th Century Britain*. Farnborough, Hants.: Gregg. Originally published in 1837 as part of J.R. McCulloch's *A Statistical Account of the British Empire.*

Fleck L (1935). *Genesis and Development of a Scientific Fact*. Chicago: University of Chicago Press (see 1979 edition with comment by Kuhn).

Frost WH, ed. (1936). *Snow on Cholera*. New York: The Commonwealth Fund; Oxford University Press.

Graham G, Farr W, Mann H (1852). *Census of Great Britain, 1851: Population Tables. I. Numbers of the Inhabitants, in the Years 1801, 1811, 1821, 1831, 1841, & 1851.* London: Printed by W. Clowes and sons for H.M. Stationery off.

Halliday S (2002). William Farr, the lancet, and epidemic cholera. *Med Sci Monit* 8:LE18–LE19.

Kuhn TS (1962, 1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lambert R (1963). *Sir John Simon, 1816–1904 and English Social Administration*. London: MacGibbon & Kee.

Simon J (1890). *English Sanitary Institutions, Reviewed in Their Course of Development and in Some of Their Political and Social Relations*. London: Cassell.

Snow J (1855). *On the Mode of Communication of Cholera*. London: J. Churchill.

Susser M (1974). Ethical components in the definition of health. *Int J Health Serv* 4**:**539–548.

Susser M, Adelstein A (1975a). An introduction to the work of William Farr. *Am J Epidemiol* 101:469–476.

Susser M, Adelstein A, eds. (1975b). *Vital Statistics: A Memorial Volume of Selections From the Reports and Writings of William Farr*. Metuchen: The Scarecrow Press, Inc.

Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip MR. (2003). *Cholera, Chloroform and the Science of Medicine: A Life of John Snow*. New York: Oxford University Press.

Winslow CEA (1943). *The Conquest of Epidemic Disease*. Princeton, NJ: Princeton University Press.

8

Contagion, Infection, and the Idea of Specific Agents

Historically and colloquially, the word *contagion* has been and still is loosely and widely used as a synonym for infection in any form regardless of mode of transmission. We use it more narrowly, with deference to Richard Mead's work in the early eighteenth century as noted in Chapter 2, to describe aspects or mechanisms of person-to-person transmission of disease.

The notion of transmissible infection existed in vague form in the Hippocratic era of the fifth to fourth century BCE (Jarcho 2000). Some early texts, not least the Old Testament, recognize and call for measures to contain the spread of disease. In the fourteenth century, Venice was under the threat of bubonic plague carried from the East by the crews of disease-ridden ships (Rosen 1958). To ward off such epidemics, the city passed a law, the first, to isolate the population when under threat to any access from without for a period of 40 days (hence quarantine, from the Italian *quaranta*). Opposition was strong from the merchants of Venice and other Italian cities whose trade was hurt by quarantines, but the law prevailed.

Much time passed before the mechanisms of transmission of specific diseases began to be understood. This long gestation suggests the extraordinary force and institutionalized consensus that sustain dominant prevailing

concepts and theories. The consensus seems to have been capable of overriding even direct observation of the direst circumstances, as regularly occurred with smallpox and sometimes with plague. Early on, inoculation against smallpox with scabs taken from infected persons was well understood to confer protection against infection. The practice was known in India even before the first millennium CE, although it had reached Europe only early in the eighteenth century. In 1798 Edward Jenner published his much acclaimed proof of the benefits of the immunity conferred by vaccination with cowpox against infection by smallpox. Even that accomplishment seemed to bring an understanding of infection and its spread no nearer. For much of the succeeding nineteenth century, Western thought made no more than a perceptible dent in the armor of the miasma theory.[1]

Girolamo Fracastoro (1478–1553) was a typical Renaissance man. Renowned as a great physician—held by some to be the greatest of his time—he was also a considerable poet. Both medicine and poetry are served by his long poem on syphilis, the work for which he is most remembered. He followed the poem, published in 1530, by a significant work on contagion in 1546. In this work, his remarkable description of the character and the transmission of infectious disease comes close to meeting present-day ideas, although Fracastoro did not recognize that the causative material is living matter. He departed from the notion of the ancients that a disease itself could be communicated from one person to another, and speculated that what passed between persons was a germ or seed.

Syphilis was the name Fracastoro gave to the disease in his allegorical poem *Syphilis sive morbus Gallicus* ('Syphilis, or the French Disease'), in which Apollo punishes the shepherd Syphilis with a disease of bodily sores. Syphilis is now thought to have been brought from the New World in 1492 by Columbus of Genoa, although the issue is debated. The first report of an outbreak was in Naples in 1494. In the annals of the kingdom of Naples, the disease reached Italy with the invading army of Charles VIII. Ludovico Moscardo in 1672 gave an apposite account of the origin and spread of the disease, as translated in the following passage:

> Not knowing whom to blame, the Spaniards call it the French disease, the French the Neapolitan disease, and the Germans the Spanish disease. (Gould 2000)

Being Italian, Fracastoro named the disease (in Latin) *sive morbus Gallicus*, in other words, the French pox.

[1] Winslow provides an excellent account—one fuller than suits the purpose of this work—in his book on the history of epidemic disease (Winslow 1943).

Fracastoro was a close observer of the features of several different infections. His 1546 book on contagion long stood virtually alone as a manual of infectious disease. In the book, he postulated three modes of contagion by supposedly imperceptible and inanimate particles. All three modes are familiar in the modern construct of infection, namely, direct contact, intermediary fomites, and infection without intermediaries (as through the air). In the next century a number of physicians followed this lead on contagion.

Athanasius Kircher (1602–1680) was a wide-ranging German Jesuit scholar at the court of the Pope. In a treatise on contagious disease, he described seeing under the microscope a multitude of corpuscles that invaded the body. He identified these as plague germs, living things. He assumed they arose by spontaneous generation, in accord with the common wisdom of the time. Internally they caused putrefaction of the body; externally they corrupted the air.

Fracastoro's near contemporary Francisco Redi (1626–1698) proved the idea of spontaneous generation false by experiment. The putrefaction in meat was assumed to give rise to the generation of the organisms. When he sealed a jar containing the meat, putrefaction still occurred but without any organisms appearing. The blowflies previously presumed to have been generated by the putrefying meat, thus prevented from laying their eggs in it, instead laid their eggs on the cover of the jar.

Redi's contemporary, Antonie van Leeuwenhoek (1632–1723) of the Netherlands, was a lens maker who reportedly made hundreds of microscopes (the first compound microscopes are generally credited to Zaccharias Janssen in 1590). Leeuwenhoek was also a biologist of note. With the aid of his lenses, he could give detailed descriptions of many microorganisms and could differentiate bacteria and protozoa. With an open mind (though seemingly unable to grasp the astounding significance of his own observations) he sent copious scientific letters to the Royal Society in England over a period of 50 years.

In the late eighteenth century, developments in the field of infection began to tail off. As noted in earlier chapters, the great and influential physician Thomas Sydenham gave a new lease of life to miasma with a theory that gave currency to a refinement: the "epidemic constitution" he posited arose from the emanations of miasma. The theory invoked vernal and autumnal entities of disease attributed to changes in season and atmosphere in the spring and fall respectively. The theory was plausible to the degree that epidemic infections do in reality vary by season. In consequence, the concept of contagion lost ground both in theory and practice.

We have seen that only in the mid-1800s, with a few notable exceptions (for instance, Richard Mead), did intimations of contagion again resume and sustain a strong challenge to miasma theory. In this period of the dominance

of miasma theory Jakob Henle in Germany had in 1840, as noted above, formulated the preconditions needed to prove a theory of infection by microorganisms.[2]

Henle argued that contagious diseases were caused by microscopic living organisms, probably of a plant variety. Infection by each type of organism incurred specific symptoms, including a time course with an incubation period that pointed to their replication within the body before the signs and symptoms of the infection could become manifest. William Farr—then beginning his career as Compiler of Abstracts—was impressed enough to open a correspondence with Henle (Rosen 1938; Susser 1987). Henle's theory drew on the observations of a number of early theorists. He drew, too, on the recorded reports of microscopic organisms, reports that grew in number once Leeuwenhoek's microscope became available. After Fracastoro, a scattering of writers over the years had postulated the existence of minute particles conveying disease from person to person (Stallybrass 1931). While miasma flourished as orthodoxy, little significant scientific advance had followed, however, until in 1840 Henle's influential treatise set out desiderata that a scientific proof of contagion must meet (Rosen 1935).

Within the following decade, other seminal works indirectly fortified the likelihood that infection was transmissible from person to person whether by direct contact, inhalation, or ingestion. In this regard, two young investigators, the Dane Peter Ludwig Panum and the Hungarian Ignác Semmelweis, each designed and executed brilliant prototypes for modern epidemiological studies. Both clearly establish infection transmitted between persons as the route of spread of the epidemic diseases of measles and of puerperal fever respectively. In England likewise within the next decade, as noted, John Snow and William Budd each produced good epidemiological evidence respectively of the transmission of cholera and of typhoid (the percipient Budd had made similar unpublished observations on typhoid even earlier).

In these classic studies, all four researchers had recognized specific disease entities arising from various sources, capable of transmission by various means, and assumed to be caused by some sort of living form. The identity of the postulated specific causes, however, perforce still rested on assumptions. The assumptions derived from consistent features of the clinical effects and pathology of the conditions shown to be transmissible. In disputing the

[2] Henle, born Jewish, was a physician and a liberal who had been imprisoned as a political activist. The Revolutionary wars had freed Jews of the Rhine province from the ghettos. In 1815, however, restrictions were restored. To cite the great German poet Heine, a contemporary of Henle's who shared Henle's experience, for Jews of the ghetto, baptism was a passport to European culture.

orthodoxy of miasma then current, Panum alone gained immediate understanding from any but a few of their senior contemporaries.

Peter Ludwig Panum (1820–1885)

Only recently graduated in medicine, in 1847 Panum published his painstaking and highly perceptive report of a devastating measles epidemic in the isolated Faroe Islands. Sent in 1846 to investigate by the Danish government at the age of 26, Panum carried out an exceptionally complete if mainly retrospective multilevel field survey of the total surviving population of 7,781 (another 6000, about 45%, had died). His studies took into account geography, demography, and climate as well as individual case tracing and mortality. He demonstrated person-to-person transmission, and the existence both of an incubation period and of immunity. These conclusions were derived by inference. They followed respectively from the time-order of infection in successive cases in contact with each other—namely, the now well-known two-week delay before the transmitted illness became manifest—and from the fact that the rather few elderly survivors of the previous 1781 epidemic all escaped a recurrence of infection (Panum PL 1847, republished 1939). For this work, Panum gained the approbation of his near contemporary in age, the politically enlightened and reform-minded pathologist Rudolf Virchow. Virchow himself entitled as epidemiology the papers in which he reported his early extensive studies of cholera in 1848 and typhus in 1849. Panum did not persevere with epidemiology, to its loss, but became a leading physiologist of his time.

In his report (in a translated version commissioned by Wade Hampton Frost), Panum wrote:

> When a physician is called upon to work in a place where climatic and dietetic conditions are different from those to which he has been accustomed, his first problem is to study the hygienic potentialities which affect the state of health of the inhabitants. It is, in fact, these hygienic conditions which contribute to the development and frequency of some diseases and the rarity and improbability of others, and which, more or less, modify the symptoms of every disease, and it is indeed on these conditions that the geography of disease, the special study of which will soon, perhaps, elevate it to the status of an independent science, is based. [Of course, we now know that science as epidemiology].

He continues:

> I shall, then, try to set forth the hygienic forces proceeding from the conditions on the islands, and as far as the observations I have been able to make permit me to do so, I shall attempt to show the influence which each of these forces in particular exerts on the general state of health of the inhabitants and on the frequency, development and method of propagation of different diseases. Together with the

> mortality rates of the country, I shall also seek to illustrate this further by statistical data collected during my sojourn on the islands. In another section I shall then present some observations in regard to the measles, inasmuch as they may be of general interest to the medical public.

In this research, Panum displays a rare vision of research thinking that does not come amiss in contemporary eyes. He probes multiple levels of organization and makes exemplary use of the time dimension (Panum PL 1847, republished 1939).

Alexander Gordon (1752–1799)

Alexander Gordon, a thoughtful and experienced obstetrician working with puerperal fever in Aberdeen published an important series of observations and inferences (Gordon 1795) 50 years before the influential review on this disease published by Oliver Wendell Holmes of Massachusetts in 1843, and before the experimental work of Semmelweis in Vienna, and the conclusive work of William Budd in Bristol, England. Moreover, Gordon's observations were fully documented. Analyzing the sequence of events of puerperal fever among his patients during an epidemic phase of that disease, he demonstrated that the order in which cases occurred followed the movement of a birth attendant, an *accoucheur*, from one patient to another, whether the *accoucheur* was a midwife or himself. "Every person who had been with a patient during the puerperal fever became charged with an atmosphere of infection, which was communicated to every pregnant woman who happened to come within its sphere" (Dunn 1998). Gordon rejected the notion of effects of the physical environment (home or hospital, freshly painted walls or not) but argued that the raw surface of the endometrium, from which the placenta had recently separated, made an excellent *nidus* for an infection. He analogized that surgical or other open wounds were vulnerable similarly to *erysipelas* (later to be understood as a *streptococcal bacterial* infection of the skin), and linked the two epidemics, noting that they coincided in time. His strong recommendation (very unpopular among accoucheurs, as can be imagined) was that it was their hands and clothes that carried the infection. He recommended that hands should be washed in chlorine and clothes washed between visits to women in labor, and that bedclothes and garments of women who had suffered from the fever should be burned.

Ignác Semmelweis (1818–1865)

Ignác Semmelweis was a Hungarian contemporary of Panum. In 1846, then a recent graduate of the School of Medicine in Vienna, he was appointed

Assistant Physician in the two free clinics of the maternity hospital. In the wards, he was at once confronted and shocked by the high death rate from childbed fever. His thought and work began to turn on puerperal fever. The facts as he saw them in the wards conformed neither with miasma nor any other current theory. Soon after Semmelweis had assumed his new duties, he was greatly affected by the death of a Dr. Kolechka, his teacher and friend. This death followed an incident in which he accidentally nicked himself while dissecting at autopsy. In great distress, Semmelweis could think only of the tragedy and how it might have come about. As he reflected on the autopsy of his friend, it came upon him that the rapid spread of the pathological lesions after the incident was identical with what he had seen at the routine autopsies of the many mothers and newborn infants who had died from puerperal sepsis soon after delivery.

From this set of events, Semmelweis drew a then distinctly unorthodox inference, namely, that the identical pathology found at his friend's autopsy and in puerperal fever must have a common source. In agonized reflection he wondered: might not the pathology in either circumstance be transmitted by decaying matter acquired in the course of conducting autopsies? If so, this death and also that of the many women in childbirth might have followed from contagion. The idea of such a mortal contagion led him to consider what the source could be. Autopsies could not but have been very much in mind during this series of events, and he decided there was good reason to suspect the autopsies so frequently conducted at the hospital. That hypothesis of course contradicted the theory of miasma then prevailing in Vienna and elsewhere.

Autopsies had been introduced as a routine procedure less than a decade earlier, in 1840, an innovation that aimed to advance both the study of disease and teaching. Physicians and their students dissected cadavers to develop their comprehension of pathology and the associated processes. Semmelweis at once set about testing his hypothesis. He began by collecting systematic data on deliveries and on autopsy practices in the two maternity clinics. First he set out the data for the years 1776 to 1822, years that preceded any regular practice of conducting autopsies. In 25 of those years mortality varied between 0.025% and 1%. In the public clinics he himself now directed over the years from 1833 to 1840, still before autopsy had become routine, at 1.25% the mortality rate was unchanged. In 1840, however, autopsy became routine and thereafter the mortality rate had increased more than fourfold, to 5.30%.

Semmelweis then turned his attention to current experience. In this period, the obstetric service had been split into two units. Circumstances now differed between the two units and, he suspected, mortality was much higher in one than the other. Analysis of the large number of deliveries confirmed

his suspicion. In the most severely affected unit, male medical students and obstetricians, all of whom routinely dissected cadavers, delivered the patients, and mortality was a horrific 5.58%. In the less affected unit, only midwives, none of whom dissected cadavers, delivered patients, and mortality was 3.38%, still a very high mortality rate, but clearly lower.

Consumed by his problem, Semmelweis then devised an intervention. He based it on an antiseptic measure practiced by English obstetricians of the time. They believed, he reported, that puerperal sepsis was contagious. When proceeding from one delivery to another, they therefore used a chlorine wash and donned clean garments to prevent spread by the contagion. He arranged that in both clinics all those conducting deliveries would beforehand do likewise, and scrub hands thoroughly in a chlorinated wash. He then assembled the clinic data needed to derive mortality rates over several years, both before the intervention and after it. Once again, his hypothesis was confirmed. The mortality rate in the unit served by medical students declined from 6.56% to 3.57%; the rate in that served by midwives changed merely from 3.38% to 3.06% (see Table 8.1).

In testing the intervention in both clinics, this innovative experiment elicited the effects of different modes of practice, tested in two dimensions well-defined by an intervention and changes over time. The work of Semmelweis presaged other features of the work of early epidemiological contributors to establishing the paradigm of infection, in this instance by contagion. Like his contemporaries John Snow and William Budd, he matched his epidemiological data on the occurrence of disease in clearly defined populations under specific environmental conditions, and against both the clinical and the pathological features of the disease under study. In these now classic studies, each earned his place as a pioneer of modern epidemiology.

Semmelweis reported his first results in 1847. Although letters and editorials in the medical press at first responded with excitement, the Vienna faculty was soon sharply divided. The entrenched concept of miasma did not permit ready entry for the notion of contagion as a guiding concept in the spread of disease. Moreover his interpretation impugned the practice of this leading medical center. The department chair, always powerful in those times, although impressed at first, later suppressed the results despite support for Semmelweis from such leading figures as Rokitansky (a founder of modern pathological science). In 1849 Semmelweis lost his appointment in the first clinic without stated cause.[3]

[3] In correspondence in the *New York Review of Books*, Sherwin Nuland attributes Semmelweis's difficulties to his disputatious character, argues that there is no good evidence that he was Jewish, and scouts the worth of his scientific contributions. Richard Horton (current editor-in-chief of the *Lancet*) sharply disputes Nuland, and affirms his confidence in the work of Semmelweis as described by K. Codell Carter (1983, also translator), the foremost Semmelweis historian of our day (Horton 2004).

Table 8.1 Deliveries and Deaths in Two Viennese Public Clinics, 1833–1858

A. Deliveries and deaths, 1833–1840 in two public clinics, conducted by male medical students and female midwifery students.

	Deliveries (*N*)	Deaths (*N*)	Rate (%)
Before the introduction of routine autopsies	71,395	897	1.25
After the introduction of routine autopsies	28,429	1,509	5.30

B. Deliveries and deaths, 1847–1858.

B.1. Male medical and female midwifery students deliver in both clinics.

	Deliveries (*N*)	Deaths (*N*)	Rate (%)
First clinic	23,059	1,505	6.56
Second clinic	13,097	731	5.38

B.2. Male medical students (first clinic) and female midwifery students (second clinic) separated by clinic:

1. Before chlorine wash

	Deliveries (*N*)	Deaths (*N*)	Rate (%)
First clinic	20,042	1,989	6.56
Second clinic (midwifery students)	17,791	691	3.38

2. After chlorine wash

	Deliveries (*N*)	Deaths (*N*)	Rate (%)
First clinic	47,938	1,712	3.57
Second clinic	40,770	1,248	3.06

Deliveries before and after autopsy became routine, conducted by male medical students (autopsy participants) and female midwifery students (nonparticipants in autopsies), and before and after the regular use of chlorine washes. (After Semmelweiss, as reported by Carter 1983).

In the Vienna of the late nineteenth century, it had not helped Semmelweis that, like Rokitansky, he was both Hungarian and, at the least, assumed by several to be Jewish. Denied other appointments thereafter and discouraged, in 1859 Semmelweis returned to Budapest. In the face of the many hostile critics on the Vienna faculty, for several years in Budapest Semmelweis had remained silent on the topic that had obsessed him. At last in 1860 he published his fully developed treatise on childbed fever (Carter 1983). Not long after, depressed and unhappy, he was admitted to a state-run mental hospital. He died within weeks, not improbably it is thought, from inhumane treatment there (Nuland 2003).[4]

[4] A contemporary autopsy report, plus exams, photographs, and x-rays taken at a disinterment and reburial in 1963, point to Semmelweis having endured a severe beating at the hands of the asylum personnel (Nuland 2003).

He was little cited thereafter until John Simon [pronounced as in the French of his immigrant father, Seemoan] took note of his work. Simon was a major figure in the public health of the second half of the nineteenth century (see Chapter 9). Winslow (1943) traces Simon's thinking year by year through his Annual Reports to the General Board of Health in England. Simon had long sustained a belief in the miasma theory. In 1854 Simon had not yet accepted Snow's postulate that a specific organism was the proximate cause of waterborne cholera epidemics. 1863 marked the beginning of his definitive shift from the miasma theory, however. That Annual Report acknowledged the seminal work Semmelweis had reported in his book of 1860. In 1874, after a 20-year lapse, Simon at last announced unqualified acceptance of contagion theory (Winslow 1943). This was apostasy that cemented the widening rift with his onetime supporter and friend Florence Nightingale.

Joseph Lister (1827–1912)

Joseph (First Baron) Lister might well be assigned a place beside Ignác Semmelweis as a successor and perhaps joint founder of clinical epidemiology. The son of a wine merchant, he was trained as a surgeon. Always interested in microorganisms, he devised technical improvements to the microscope. Later in life he wrote several scientific papers on the subject. In 1864, Lister was referred to an 1863 paper by Pasteur which reported his observations on fermentation and putrefaction, and on the presence of microorganisms in the air. Like all surgeons of the time, Lister was familiar with suppuration and the very high accompanying mortality often observed with amputation wounds. This unwelcome surgical phenomenon, it occurred to Lister, could be of the same order as Pasteur's report showing putrefaction to be the product of those same microorganisms. Phenol was in use at that time as a hygienic treatment for sewage and excrement. Logically enough, Lister selected it as an *antiseptic* likely to destroy inimical microbiological invaders.

In the operating theatres and the surgical wards in which Lister was attending surgeon, he began to treat surgical wounds with phenol, and the dressings with carbolic acid (later he applied carbolic acid with a spray). He applied creosote against possible vehicles for organisms—such as the air, instruments, hands, and clothes—with which the wounds from surgical incisions might come into contact. Not unreasonably, Lister was soon fully convinced that such antisepsis was effective, although in modern eyes the work would surely be criticized as experimentation without adequate controls. Such understanding was still in the future. In 1867, Lister published

a long paper on his work. Without the application of phenol to the wound, in 35 amputations 16 died (approximately 46%). With the above measures and the addition of phenol applied as an antiseptic to the surgical wound, in 40 amputations 6 died (15%), only one third of previous expectation (Lister 1867).

Lister's papers elicited huge interest. Surgeons, physicians, and obstetricians visited the hospitals in which he operated and then adopted his procedures. Across Europe and America, wherever surgeons applied his antiseptic method the results were dramatic. Almost immediately, hospital mortality rates fell sharply for amputation wounds. In obstetric units, too, in vindication of Semmelweis, rates for childbed fever (puerperal sepsis) fell likewise. Robert Koch, who came on to the scientific scene in the 1870s, used his great authority to criticize the use of carbolic acid on the grounds that it failed to destroy anthrax spores. Although such criticism was scarcely relevant to an operating surgeon, Lister switched to the even more toxic mercuric chloride.

Yet, in the face of these dramatic reports, miasma theory retained its dominance through another three or four decades. However, Lister, firmly convinced of the as yet developing germ theory, several times wrote personally to Pasteur to express his debt to him. He had admired the work of Semmelweis no less, although it had not directly influenced his own work. At the time Lister developed his methods of antisepsis, he explained in 1906, he had not yet learned of Semmelweis. Oddly, even perversely, among many colleagues who visited and admired Lister, and even among those who in their own hospitals followed his precepts on antisepsis, many nonetheless rejected germ theory. This contradiction probably had its roots in Lister the canny Scot himself. Regardless of professed beliefs, he played down germ theory. It was sufficient to satisfy him that his methods were followed. Later in the century Pasteur, invited to lecture and visit by Lister in England, observed the application of Lister's methods. Pasteur readily agreed that in the hospital wards microorganisms were transmitted by instruments, dressings, and surgeons' bare hands, although he doubted that organisms carried by air alone as the vector were sufficient to convey them and cause infection. Pasteur also proposed that the procedures to prevent infection be described as *asepsis* rather than as *antisepsis*: sufficient remedy would follow if doctors and nurses "flamed" their hands (passed their hands briefly through a flame) and then washed them, and if instruments and dressings were incubated.

It is perhaps ironic that Lister, energetic, enterprising, and resourceful as he was, contributed nothing new to epidemiological theory, nor indeed to scientific methods. Yet in the field of surgery and dramatically in following the very logic of Semmelweis's investigations of puerperal fever, wherever accoucheurs followed his example the results in medical practice were

spectacular. He died in 1912, appropriately honored and acclaimed for his contributions.

John Snow (1813–1858)

Vinten-Johansen et al. (2003) have provided the first fully researched account of Snow's life and career.[5] Snow qualified in London and entered practice in the north of England, in his native environment, in and around Newcastle-on-Tyne. Early on he revealed his rigorous and imaginative scientific bent. Local outbreaks of disease of uncertain provenance in his practice seemed always to stimulate his interest and to lead him to investigate. He habitually tried to test hypotheses and innovations further by experiment as well as in observational studies (Figure 8.1).

Snow left Newcastle and Tyneside for London to improve his opportunities for further study and research. He set up there as a general practitioner. A consistently diligent scholar, however, he attended the scientific meetings

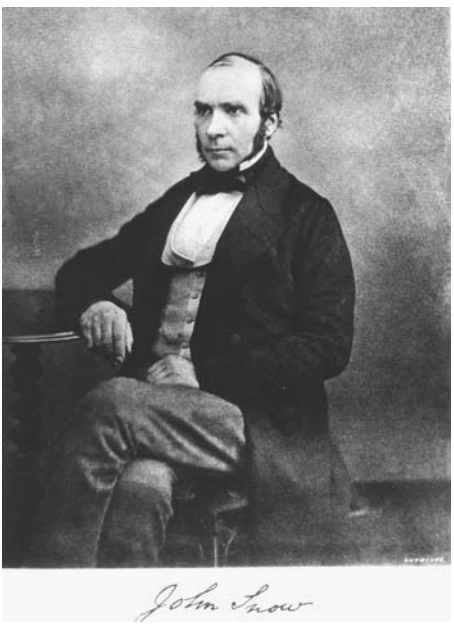


Figure 8.1 John Snow (1813–1858). Credit: Courtesy of the National Library of Medicine.

[5] This text may be profitably supplemented by earlier works—for example, Frost's foreword to the re-publication of Snow's 1854 book on Cholera (Frost 1936) and George Davey Smith's essay of 2001 (Davey Smith 2002)—as well as a recent retelling by Steven Johnson (Johnson 2006).

of medical societies, maintained his habit of systematic reading of journals and other medical literature. At the same time, he kept detailed records of the cases he saw and precise notes of his scientific investigations. Before long he was in heavy demand as the pioneer and foremost practitioner of anesthesia in London and indeed in the world. He did not flag in maintaining these studious routines even as his practice grew and flourished.

In Boston in October 1846, the Boston physician William Morton had stirred great excitement when for the first time he successfully anesthetized a surgical patient by administering ether by inhalation. When news of the dramatic event reached London, Snow at once entered on scientific studies of anesthesia with ether. Not long after, when the use of chloroform as the inhalant was first reported in Edinburgh, he immediately added research on chloroform to his systematic studies. Having otherwise only the obligations of his practice, Snow relentlessly pursued his successive investigations. He adopted an experimental approach to anesthesia by applying existing scientific knowledge of the behavior of gases and of human physiology.

As anyone can well understand, major surgery without anesthesia was a terrifying experience for patients, and one that could be traumatic for surgeons as well. General anesthesia was thus an enormous practical advance that served both patient and surgeon, and soon ether was being used in London. Snow's imagination was at once taken by the new practice. By December 28, 1846, armed by his own scientific studies of ether and having witnessed its use as a dental anesthetic, he entered without hesitation on the practice of anesthesia. Within a month, he had discovered and calculated the variable anesthetic strength of the effects of ether at different temperatures, produced an apparatus to control the variable flow, and then demonstrated his results at a medical society meeting. In 1847, he published a detailed volume on his observations (Snow 1847).

In November 1847, James Young Simpson, Professor of Obstetrics at Edinburgh, reported on his use of chloroform as an anesthetic. Within 10 days Snow was once more at the Westminster Medical Society and comparing the new "letheon" agent (in Greek legend, Lethe was the river of forgetfulness) with ether. He found chloroform both more efficient and efficacious than ether. As he had done with ether, he at once entered on systematic experimental studies. Snow studied the properties and effects of the new agents—not excluding himself as subject—and also the means he had devised to accommodate those properties and ensure their safe delivery.

His own practice in that area of his expertise grew rapidly. When in 1853 Queen Victoria demanded anesthesia for the delivery of her eighth child, there was no question that anyone other than Snow would be called on. Again he published a number of reports of his results, and in 1858 he assembled his research in a book, of which one might abbreviate the long title to *On*

Chloroform and other Anaesthetics. Snow's experimental studies of ether and chloroform—both of their properties and effects as anesthetic agents—still stand as the foundation of sound and safe anesthetic practice.

As modern-day epidemiologists know, Snow's repute as a founder of their discipline has nothing to do with his anesthetic contributions, excepting only an intense commitment to scientific rigor in both fields. His equally significant epidemiological contributions derive from his investigations of the cholera epidemics then threatening Europe. By the time he began to build a case for specific infection as the source of cholera, he had already shown himself to be a consummate scientist. Concurrently with his work on anesthesia, Snow had picked up the thread of his early research interest in cholera. He had experienced the 1832 epidemic when practicing among coal miners in his native Tyneside.

Snow was to that extent prepared when in 1848 another cholera epidemic threatened. Now in London, in his characteristic methodical manner he began with a consideration of the pathology of the condition. His friend Sir Alfred Baring Garrod (1819–1907)[6] and others had reported a marked increase in the viscosity of the blood in cholera victims. Garrod thought that some poison in the blood was responsible for fluid loss. Conversely, Snow inferred from his own work—and in accord with modern understanding—that such viscosity was not a cause of cholera but the result of fluid loss. Further, he challenged the idea put forward by some authorities that inhalation was the route of infection. For Snow the pathology in the gut, and its concordance with the main clinical symptom of watery diarrhea and so-called "ricewater" stools, was conclusive: the route of any agent must have been by ingestion.

Snow's first three London studies were all of localized outbreaks. On the negative side, these convinced him that no such thing as promiscuously distributed miasma, whether general or local, could be the cause of the outbreaks. On the positive side, his investigations led him to three central conclusions. They convinced him first that the patterns of distribution and the nature of the disease itself were owed to ingesting contaminated water; second, that the agent of the contamination was a specific organism. Third, he speculated that the agent might be microscopic bodies, inanimate but nonetheless capable of replication and transmission.

In all three studies, he had located the common source of the outbreak. In the first two outbreaks, both in Surrey south of London at Albion Terrace and Horsley Down, cesspools abutted on the incoming water supply and could readily contaminate the water (Vinten-Johansen et al. 2003). The third outbreak was the famous Broad street pump episode of 1854. Here again, a

[6] Father of Sir Archibald Edward Garrod (1857–1936), who studied alkaptonuria and described this and other metabolic conditions as genetic diseases.

large cesspool draining into the Thames was close enough to the much-used pump to contaminate that water supply. Finally, Snow capped these circumscribed studies with an investigation of epidemic mortality that stemmed from the sewage-contaminated Thames water supplied to a large population of South London by the Southwark and Vauxhall Water Company. This brief account suffices to convey the beginnings of the evolution of Snow's fundamental epidemiological contributions. They are described in detail in a valuable recent biography of Snow (Vinten-Johansen et al. 2003).

Snow supported his hypothesis with his findings in the first localized cholera epidemics he investigated. These two investigations, some done after the fact in 1849, were limited outbreaks in "court" or row houses, first at Horsley Down and then at Albion Terrace south of the River Thames. Here he was already contending with the report of a member of the Sewer Commission, who resorted to the standard explanation of miasma arising from a stream about a quarter of a mile away. Snow first investigated the distribution of the cases, and then showed that the water supplies of the row houses they had occupied could well have been contaminated by fecal dejecta. Both rows had cesspools built in close enough proximity to the water supply system of the households to allow their contamination by sewage.

These observations were at the heart of Snow's rejection of miasma theory. The conventional assumption, that the source of cholera was a miasma inhaled into the lungs, was at odds with Snow's observations. Given Chadwick's modification of miasma as sometimes occurring in limited local form, miasmatists were armed to dispute Snow's hypothesis that befouled water was a source of cholera. Snow's direct challenge was too much for convinced miasmatists, however. What they could not accommodate—given their prevailing idea that miasma penetrated the body but by means of the inhalation of vague and undefined atmospheric poisons—was that the disease was caused by ingestion of some specific if still unknown agent.

The strength of Snow's early challenge resided in the several direct associations of cholera with the pathological expression of the disease in the gut, and thus with ingestion rather than inhalation. Snow's findings, almost but not quite unique until then, provided a much better fit to contagion. William Budd likewise argued, in roughly concurrent studies of typhoid outbreaks in his rural general practice, that the ingestion of water open to contamination by sewage and the discharge of potentially infected feces, together with the pathological manifestations of Peyer's patches in the ileum, were consistent not with inhaled miasma but with a specific intestinal infection. Budd generously championed Snow's published results, having already reached similar conclusions about his own studies of cholera as well as of typhoid.

In 1849, long before John Simon on reading Semmelweis in 1863 had made known his quiet conversion from denial of infection to doubt about miasma

theory, John Snow had published a first version of his retrospective studies of cholera mortality in the London epidemics of 1832 and 1849. In 1854 on Broad Street in central London, another localized but serious epidemic broke out, one now familiar in epidemiological history for its drama and Snow's challenging interpretation of the mode of spread. Snow turned his attention at once from the investigation of the large-scale cholera epidemic he was already engaged with in South London. A Reverend Whitehead, together with Snow and others, acted on behalf of a committee of inquiry set up by the local Vestry Committee of St James and Dr. Fraser acted on behalf of John Simon (by then Chief Medical Officer at the General Board of Health). The Board of Health had already concluded that the source was a foul stream some 400 yards away, a conclusion in accord with miasma doctrine.

Golden Square was the site of a pump on Broad Street much favored by the neighboring populace for the quality of its water.[7] Here Snow meticulously conducted his own independent investigation. Painstakingly, he sought out and recorded the timing and location of all cases, their water usage, and its source. He also mapped the distribution and the timing of the cases among the houses on these streets. Once more, when he followed the route of the disposal of waste waters, he found a source consistent with his hypothesis. The ostensibly pure water fed into the pump was open to contamination by a discharging sewage pipe abutting on the pump well.

Snow's work was innovative in several respects. Notable, in addition to his fundamental contributions to the theory of contagion, are his use of the detailed mapping of the distribution of the Broad Street cases, and also his sense in this and other work of the necessity for controls. In 1855 Snow published a second edition of his completed studies of cholera (discussed later in a different context, Frost 1936). From the assembly of larger scale studies of cholera epidemics in England, he concluded that the source of the widespread cholera epidemics in London was to be found in sewage-contaminated water supplies drawn from the Thames. In its own time, this elegant model for epidemiological work met a mixed reception. A few percipient supporters like Benjamin Ward Richardson kept Snow's work alive. Snow had concluded further that the cause of cholera epidemics was a transmissible parasitic microorganism, and suspected that the vibriones he had detected microscopically were the likely proximate cause.[8]

[7] In Chapter 19, we describe the exceptional case, famously recorded by Snow, of the death from cholera of Mrs. Eley. Daily she had Broad Street water trundled to her home in far-off Hampstead.

[8] In 1854 Fillippo Pacini of Florence, a prominent microscopist, reported postmortem vibriones in the mucous membrane of the gut, but apparently this discovery remained unknown to British medicine (Vinten-Johansen et al. 2003).

John Snow was first spurred to investigate cholera, we have seen, by a number of localized outbreaks. Thus, when later he followed on with larger scale studies that related cholera deaths to administrative areas as Farr had done, Snow had already tested his hypothesis at a more refined level of observation. Taken together with his studies of the pathology of the gut, he thus had the advantage of a greater degree of precision in tests of his hypothesis. Each higher level of the units of observation adds new dimensions and greater complexity, with increasing numbers of relationships capable of confounding analysis.

Snow could proceed with greater confidence in the face of prevailing opinion, having supported his initial analyses and carried forward tests of his hypothesis with larger units of observation. With these data, he could comfortably follow Henle. Thus, about 30 years before Koch and Pasteur described the Vibrio as the agent in cholera, he postulated an invisible, self-reproducing living agent as the cause of the manifestations of cholera seen in individuals. So precise an hypothesis about individuals implied a linked sequence of events between data for areas and for individuals and could not be definitively tested at the level of organization of administrative areas alone.

Early in his studies, Snow tabulated in rank order the cholera mortality rates of the 1849 outbreak for the districts of London, and set against these rates the company responsible for the water supply (see Tables 8.2–8.4). Table 8.4 shows an obvious association of the water supplied by the Southwark and Vauxhall and the Lambeth Waterworks with the high rates of death from cholera. This was a direct relationship at the ecological level.

Five years later, Snow pointed out that in the 1853 outbreak a change had come about in the cholera distribution in South London: the Southwark and Vauxhall Company water supply was still associated with high cholera mortality, but the Lambeth supply seemed not to be. It turned out that the Lambeth Company had moved their waterworks to a point higher up the Thames, thus obtaining a supply of water free from the sewage of London. William Farr in his weekly return of births and deaths from the General Registry Office had provided this lead for Snow.

Now Snow again narrowed his focus downward to a lesser level of organization, namely, households. He produced thereby his most convincing evidence of the relationship of water supply to individual cholera deaths. As soon as the dependent variable is observed and specified at a level of organization higher or lower than that of the independent variable, the postulated causal relationship is made indirect. Thus as soon as Snow might try to relate the water supply of areas to deaths in households, he was dealing with an indirect relationship. Between area water supply and household deaths there must be intervening factors at the household level, and the next logical step in investigation was to search out these factors.

Table 8.2 Showing the Mortality from Cholera, and the Water Supply, in the Districts of London in 1849

District	Population mid-1849	Deaths from Cholera	Deaths by Cholera to 10,000 Inhabitants	Annual Value of House and Shop Room to Each Person in £	Water Supply
Rotherhithe	17,208	352	205	4.238	Southwark and Vauxhall Water Works, Kent Water Works, and Tidal Ditches
St. Olave, Southwark	19,278	349	181	4.559	Southwark and Vauxhall
St. George, Southwark	50,900	836	164	3.518	Southark and Vauxhall, Lambeth
Bermondsey	45,500	734	161	3.077	Southwark and Vauxhall
St. Saviour, Southwark	35,227	539	153	5.291	Southwark and Vauxhall
Newington	63,074	907	144	3.788	Southwark and Vauxhall, Lambeth
Lambeth	134,768	1,618	120	4.389	Southwark and Vauxhall, Lambeth
Wandsworth	48,446	484	100	4.839	Pump-wells, Southwark and Vauxahall, river Wandle
Camberwell	51,714	504	97	4.508	Southwark and Vauxhall, Lambeth
West London	28,829	429	96	7.454	New River
Bethnal Green	87,263	789	90	1.480	East London
Shoreditch	104,122	789	76	3.103	New River, East London
Greenwich	95,954	718	75	3.379	Kent
Poplar	44,103	313	71	7.360	East London
Westminster	64,109	437	68	4.189	Chelsea
Whitechapel	78,590	506	64	3.338	East London
St. Giles	54,062	285	53	5.635	New River
Stepney	106,988	501	47	3.319	East London
Chelsea	53,379	247	46	4.210	Chelsea

East London	43,495	182	45	4.823	New River
St. George's, East	47,334	199	42	4.753	East London
London City	55,816	207	38	17.676	New River
St. Martin	24,557	91	37	11.844	New River
Strand	44,254	156	35	7.374	New River
Holborn	46,134	161	35	5.883	New River
St. Luke	53,234	183	34	3.731	New River
Kensington (except Paddington)	110,491	260	33	5.070	West Middlesex, Chelsea, Grand Junction
Lewisham	32,299	96	30	4.824	Kent
Belgrave	37,918	105	28	8.875	Chelsea
Hackney	55,152	139	25	4.397	New River, East London
Islington	87,761	187	22	5.494	New River
St. Pancras	160,122	360	22	4.871	New River, Hampstead, West Middlesex
Clerkenwell	63,499	121	19	4.138	New River
Marylebone	153,960	261	17	7.586	West Middlesex
St. James, Westminster	36,426	57	16	12.669	Grand Junction, New River
Paddington	41,267	35	8	9.349	Grand Junction
Hampstead	11,572	9	8	5.804	Hampstead, West Middlesex
Hanover square & May Fair	33,196	26	8	16.754	Grand Junction
London	2,280,282	14,137	62		

The districts are arranged in the order of their mortality from cholera (Snow as reported in Frost 1936).

Table 8.3 Cholera Death Rates by Chief Water Companies Supplying Districts, Grouped by Sources of Company Water, London 1854

Aggregate of Districts Supplied chiefly by the Respective Companies				
Water Companies	Sources of Supply	Population	Deaths by Cholera in 13 Weeks Ending November 19	Deaths in 100,000 Inhabitants
(1) Lambeth and (2) Southwark and Vauxhall	Thames, at Thames Ditton, and at Battersea	346,363	211	61
Southwark and Vauxhall	Thames, at Battersea	118,267	111	94
(1) Southwark and Vauxhall (2) Kent	Thames, at Battersea; the Ravens-bourne in Kent; ditches and wells	17,805	19	107

Snow as reported in Frost 1936.

Table 8.4 Cholera Death Rates by Company Supplying Household Water, London 1854

Company	Number of Houses	Deaths from Cholera	Deaths in Each 10,000 Houses
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423	1,422	59

Snow as reported in Frost 1936.

As is often the case in epidemiological studies, to obtain the data relating to this smaller and more refined unit of observation, Snow had to conduct a special investigation. It so happened that the overlap of the water supply provided by the two companies was so intimate that an ideal opportunity for observation was created. Snow described the situation:

> As there is no difference whatever, either in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised, which would not thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer.
>
> The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolk down to the very poor, were divided into two groups

> without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity. (Frost 1936)

It might already be evident from the story of several crucial discoveries that epidemiology is an opportunistic science which rewards those gifted with imagination. To turn this grand experiment to account Snow had at once understood that, in any individual house where a fatality from cholera might occur, all that was needed would be to establish the source of the water supply. The disadvantage of the Southwark and Vauxhall Company consumers was large. Compared with those of the Lambeth Company, their relative risk of cholera deaths was more than 8 to 1.

The conviction arising from this study is greatly strengthened by the smaller unit of observation and the more direct relation shown between independent and dependent variables. In epidemiology despite the necessary insistence of a focus on populations, the ultimate concern of the discipline must also encompass health, disease, and death as it occurs in individuals. In the analyses by district previously shown, we could not know if all individuals were exposed to the same water supply. This analysis brought Snow close to that understanding. In this instance, the persons grouped in this or any households within the same area and exposed to different water supplies are specified with much greater precision and almost in imitation of random assignment. The room for action of confounding variables, if not entirely eliminated, is greatly narrowed.

Snow's work stands as a founding classic in epidemiology. By his procedures of method and inference, he solved a major problem and established a model for all who followed him. He was careful to seek out the intervening links of indirect causal chains, including the indirect relationships created by moving from one level of organization to another.

In its time the ironic significance of Snow's detailed theory of a specific molecular cause of disease was that it thoroughly violated the precepts of miasma theory. It was far from dislodging the miasma theory, however, although colleagues and patients sustained their high respect for Snow. Aside from his contributions to anesthesia, as the years passed after his death in 1858 at the age of 45, memory of his epidemiological achievements slowly faded. It was only in the late 1920s and 1930s that Wade Hampton Frost[9] revived Snow's work in his teaching of epidemiology at Johns Hopkins University School of Public Health. In 1936, Frost edited and rehabilitated

[9] The American public health expert William Sedgewick (1855–1920) in his textbook of 1902 devoted 13 pages to the Broad Street pump episode. See Vinten-Johannsen et al.

Snow's book on cholera and restored him to his due position as a founder of the discipline (Frost 1936). Today, every epidemiology student is expected to be familiar with Snow's work. His contributions continue to be explored in many works by epidemiologists, historians, and anesthetists (Vinten-Johansen et al. 2003).

William Budd (1811–1880)

William Budd of Bristol was a contemporary of John Snow and a country doctor gifted in practice and research insights. In time and significance, his thinking and observations on transmissible gastrointestinal infection are in parallel with Snow's work on cholera. Budd had himself published a paper on cholera only a month after Snow's work on the "Mode of Communication of Cholera." There he had concluded that cholera was an infection, owed to a fungus in fecal matter and acquired by ingesting water contaminated by it. Budd was generous enough to come to the support of Snow's work and conclusions. Before Budd published his paper on cholera, from studies of outbreaks of fever in his country practice he had already reached his own unpublished theory of the contagious spread of typhoid. Typhoid fever had only recently been segregated from typhus by Pierre Louis (the two infections do have similarities in the accompanying fever and rash). Budd's work on typhoid had begun in 1839, but he first published it much later, in a series of Lancet papers from 1856 to 1860. Eventually, on his retirement in 1873, Budd collected his work in one volume (Budd 1931, 1977).

Budd's studies are apt forerunners both in establishing the specificity of infectious diseases and the basis of clinical epidemiology. His first experience with typhoid—a disease that did not yet carry that label—was as medical attendant to students in a military school: 109 were affected in an epidemic. He noted that among those sent away to avoid the epidemic, after an interval of several days more boys began to droop with an infection they must have carried home. Over the next two decades, he reinforced his conclusions by close observation of the living conditions of his rural patients. He recorded the distribution of cases in his practice, and his careful follow-up of the spread of the disease among families and their contacts. Some instances were especially striking: to map the spread of the disease, Budd traced the subsequent movement of five affected persons across neighboring villages; each ignited outbreaks as they left one village for another. The typical open sewers on the farms of the time carried transmissible fungus, he concluded, and were properly characterized as mere "extensions of the infected intestines."

Budd clearly possessed great clinical acumen. With typhoid, he recognized a variable latent period of 7 to 21 days before the fever took hold (now known as the incubation period), and also the freedom from further attacks of survivors of previous attacks (now understood to be acquired immunity). He credited the first report of this then remarkable phenomenon, however, to Bretonneau, "one of the greatest authorities, living or dead." As a further characteristic of typhoid he noted the raised lymphoid tissue in the lining of the small intestines now known as Peyer's patches. Budd took the fact of such gut lesions seen with typhoid, as Snow did with cholera, to indicate that the route of infection was by ingestion and not inhalation.[10]

In considering and dismissing the possibility that miasma explained the epidemic phenomena he had so closely observed, Budd was well aware that, excepting a few distinguished supporters, he faced skeptics in the miasmatists who dominated the General Board of Health and much current thought. Budd argued from a typical Baconian negative. The epidemic eruption or quiescence of the disease, he asserted, occurred regardless of conditions thought to favor miasmata. He derided the idea of pythogenic fevers emanating from filth and corruption. The episode of London's affliction in 1858 and again in 1859, "when the Thames stank so badly" for weeks on end, buttressed his argument. A notoriously oppressive stench had risen from the sewer-contaminated Thames—unmistakably a miasma if ever there was one. The newspapers and many in the city daily predicted a terrible toll of death from the pestilence (see below). In the event, the outcome confounded all forecasts. Mortality from all causes, including deaths attributed to fevers and diarrhea, was actually lower than expected in normal times.

> The occasion was no common one. An extreme case, a gigantic scale in the phenomena, and perfect accuracy in the registration of the results—three of the best guarantees against fallacy—were all combined to make the induction sure. For the first time in the history of man, the sewage of nearly three millions of people had been brought to seethe and ferment under a burning sun, in one vast cloaca lying in their midst.
>
> The result we all know. Stench so foul, we may well believe, had never before ascended to pollute this lower air…For many weeks, the atmosphere of the Parliamentary Committee rooms was only rendered barely tolerable by the suspension before every window of blinds saturated with chloride of lime, and by the lavish use of this and other disinfectants. More than once, in spite of similar precautions, the law-courts were suddenly broken up by an insupportable invasion of the noxious vapor. The river steamers lost their accustomed traffic, and travelers, pressed for time, often made a circuit of many miles rather than cross the city bridges.

[10] Both Budd and Snow were founding members of the London Epidemiological Society formed in 1850.

> For months together, the topic almost monopolized the public prints. Day after day, week after week, the Times teemed with letters, filled with complaint, prophetic of calamity, or suggesting remedies.... "India is in revolt, and the Thames stinks," were the two great facts coupled together by a distinguished foreign writer, to mark the climax of a national humiliation. (Budd 1931, 1977)

Members of Parliament, and noble lords, dabblers in sanitary science vied with professional sanitarians in predicting pestilence. But alas for the pythogenic theory, when the returns were made up, "the result showed, not only a death-rate below average, but, as the leading peculiarity of the season, a remarkable diminution in the prevalence of fever, diarrhoea and the other forms of disease commonly attributed to putrid emanations" (Budd 1931; 1977).

Although these words might seem to proclaim the end of the miasma era, the protagonists of the idea of infection as the cause of specific disease would see it flourish for at least another 25 years before the general view of disease causation dispensed with the notion of miasma.

References

Budd W (1931; reprinted 1977). *Typhoid Fever: Its Nature, Mode of Spreading, and Prevention*. New York: Arno Press; Reprint of the 1931 ed. published by American Public Health Association, New York.

Carter KC (1983). *The Etiology, Concept, and Prophylaxis of Childbed Fever by I Semmelweis (1860).* KC Carter. Trans. Madison: University of Wisconsin Press.

Davey Smith G (2002). Commentary: behind the Broad Street pump: aetiology, epidemiology and prevention of cholera in mid-19th century Britain. *Int J Epidemiol* 31:920–932.

Dunn PM (1998). Dr Alexander Gordon (1752–99) and contagious puerperal fever. *Arch Dis Child Fetal Neonatal Ed* 78:F232–F233.

Frost WH, ed. (1936). *Snow on Cholera*. New York: The Commonwealth Fund; Oxford University Press.

Gordon A (1795). *A Treatise on the Epidemic Puerperal Fever of Aberdeen*. London: G. G. & J. Robinson.

Gould SJ (2000). Syphilis and the Shepherd of Atlantis. *Nat Hist* 109:38–42, 74–82.

Horton R (2004). The Doctors' plague: germs, childbed fever, and the strange story of Ignac Semmelweis. *New York Rev Books* 51:9–11. Also see exchange of letters, vol. 51(5).

Jarcho S (2000). *The Concept of Contagion in Medicine, Literature, and Religion*. Malabar, FL: Krieger Pub.

Johnson S (2006). *The Ghost Map: The Story of London's Most Terrifying Epidemic—and How It Changed Science, Cities, and the Modern World.* New York: Riverhead Books.

Lister J (1867). *The Collected Papers of Joseph Lister (1979).* Special ed. Birmingham, AL: Classics of Medicine Library.

Nuland SB (2003). *The Doctors' Plague: Germs, Childbed Fever, and the Strange Story of Ignâac Semmelweis.* 1st ed. New York: W. W. Norton.

Panum PL (translated and republished 1939). Observations made during the epidemic of measles on the Faroe islands in the year 1846. *Med Class* 3:802–866 [Hatcher AS, trans.].

Rosen G (1935). Social aspects of Jacob Henle's medical thought. *Bull Inst Hist Med* 5:509–537.

Rosen G (1938). *On Miasmata and Contagie (Henle 1840).* George Rosen (Trans.). Baltimore: The Johns Hopkins Press.

Rosen G (1958). *A History of Public Health*. Expanded ed. Baltimore: Johns Hopkins University Press. Reprinted 1993.

Snow J (1847). *The Inhalation of the Vapour of Ether in Surgical Operations.* London: John Churchill (reprinted in 1953 by *B J Anaesth* 25(1):53–67 [part 1], 25(2) 162–169 [part 2], 25(3) 253–267 [part 3], 25(4) 349–382.

Stallybrass CO (1931). *The Principles of Epidemiology and the Process of Infection*. London: G. Routledge & Son, Ltd.

Susser, M (1987). The Work of George Rosen: Observations on a Very Small Sample. In Susser M, *Epidemiology, Health and Society.* New York: Oxford University Press, pp. 57–60.

Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip MR (2003). *Cholera, Chloroform and the Science of Medicine: A Life of John Snow.* New York: Oxford University Press.

Winslow CEA (1943). *The Conquest of Epidemic Disease*. Princeton, NJ: Princeton University Press.

9

Origins of a National Public Health System

Under the 1848 Public Health Act, Edwin Chadwick succeeded in setting up under his own direction the first General Board of Health. In the early nineteenth century, the sustained contempt for medicine he manifested was not difficult to justify; he looked instead to engineering for technical support. Medical practice was then entirely unregulated, excepting the mild self-regulation of their members by academic bodies, the Royal College of Physicians and, later, the Royal Colleges of Surgeons and of Obstetrics and Gynecology. Quacks could hang a shingle without hindrance.

It was near mid-century before professional organizations had formed and assumed the power to set standards for themselves, and later still before John Simon himself established training for public health professionals. With Florence Nightingale, Chadwick stood in strenuous opposition to any domination by the medical profession. Its nascent organizational strength was then becoming manifest in the newly formed British Medical Association.

Harnessing Epidemiology to Public Health: John Simon (1816–1904)

The public roles assumed by John Simon during his highly effective career, over the 27 years before he felt obliged to resign in 1875 at the age of 59,

are significant for the history of public health. The question may arise, for readers of this text, whether it was as significant for the history of epidemiology. About this we shall see that there is little if any doubt of that. Simon was a scientist first, a socially effective and politically skilled health activist second. Indeed, Rudolf Virchow (a founder of cellular pathology, a dominant figure in European medical science of the mid-nineteenth century, and at the same time a puissant and enlightened social reformer) had classed Simon among the cleverest of pathologists. In a telling instance, Simon had been confident enough to challenge the tumor growth of Rokitansky, accredited founder of modern pathology. Simon, grandson of a French émigré whose son had done well in the city, was a cultivated man. His wife Jane did much to sustain his interest in literature and the arts amid a distinguished social circle of artists, writers, and poets, John Ruskin among them.[1]

Like all his predecessors in taking the epidemiological approach to public health and medicine, Simon was largely an autodidact. He was perhaps not a thinker of sweeping vision, which is no detraction in the light of his considerable practical achievements. Rather he was consistently ingenious in elucidating and devising means to remedy or control the worst of the problems that faced him. Many ills, new and old, beset the health of the rapidly urbanizing population of nineteenth-century England. In the main, his singular contributions were practical achievements: in particular, ensuring the systematic role of epidemiology in public health, as well as devising official and properly regulated medical structures and roles to give effect to the appropriate public health measures.

Simon demonstrated the methods and applied the results to the ultimate social goal of maintaining the public health. The goal was to unravel public health problems by establishing the relevant facts about disease in terms of its distributions, its sources, and causes, and its effects on populations. In this arena, the contributions of Simon and his team to a variety of empirical population and social survey methods are impressive. Indeed, his team of young medical officers (largely recruited ad hoc) developed methods for the population survey approach in Britain that precede the priority commonly accorded to Charles Booth and Seebohm Rowntree around the turn of the nineteenth century.

[1] Pronounced as in French, Simon's name was accented on the last syllable. Aside from medical colleagues, many distinguished persons frequented Simon's home. Among those now still famous were the scientists Charles Darwin, and from abroad, von Pettenkofer and Claude Bernard; among writers and artists with names still known were John Ruskin (a close companion), the artist Burne-Jones, Charles Kingsley, William Morris, G.H. Lewes (long-standing partner of the novelist George Eliot), Gabriel Rosetti with his daughter Christina, and Algernon Swinburne. The already well-known novelists Charles Dickens and W.M. Thackeray and others were also included in his social circle.

In the United States, a contemporary survey, widely discussed and probably known to Simon, was Edward Jarvis's report on *Insanity and Idiocy in Massachusetts*, published in 1855 (see Jarvis 1855, republished 1971). As with the studies devised or inspired by Simon, Jarvis elicited data from physicians, clergymen, hospitals, and whichever informants he could reach, and published careful, systematic tabulations of his findings. Simon, at around the same period, directed, for example, the 1856 statistical studies of water quality, the 1857 questionnaire on smallpox vaccination addressed to more than 500 European doctors, and the 1866 statistical estimates compiled from the data of sanitary and venereal disease surveys.

No one before Simon had so consistently, clearly, and fully demonstrated the importance and the effectiveness of epidemiology for the whole range of public health policy as it existed in his own time. Excellent historical material is available on Simon and his contributions (Simon 1890; Lambert 1963; Brockington 1965; also see Wohl 1983; Hamlin 1998; Worboys 2000). Indeed, few have matched his record since in these respects. Simon's career provides the first fully developed instance of the essential and permanent role of epidemiology in the pursuit of the public health grail. John Simon began his medical career at St. Thomas' Hospital in London as a surgeon and lecturer in pathology, and always kept abreast of new medical developments. Under Chadwick's new Public Health Act in 1848 the City of London, nerve center of the rapidly growing metropolis, appointed Simon as its first Medical Officer of Health (his was the second such appointment in England overall, Duncan of Liverpool holding the first). For this appointment Simon had had Chadwick's initial support. Simon's seven-year tenure was remarkably successful. His charge was to oversee the public health of the square mile that comprised the City. While he always carefully retained his teaching hospital consultancy in pathology, Simon proved himself to be an extremely effective official. The new medical officer quickly gained the influential support of the Times (as also the friendship of an important journalist on the paper), as well as that of the *Lancet* and the *British Medical Journal*.

The *Times*, bellwether of the press and the English governing classes, regularly gave space and applause to Simon's programs. Thus it published in their entirety the annual reports he initiated as a requirement of good practice. These reports, if "ostentatiously scientific" to some eyes, were well-written, data-based as far as they could be, synthetic, informative, and persuasive. They were also the harbingers of plans, policies, and innovations. Among many such innovations, Simon persuaded William Farr (of the General Register Office) to publish weekly reports of mortality in the various parishes of the City. The Boards of Guardians in each parish were responsible for the local administration of the Poor Law, and Simon arranged for those data to be assembled by the medical officer assigned to the parish.

To begin with, Simon adhered to the conventional and dominant miasma paradigm that poverty led to ill health. He cleaned the streets, addressed the gross overcrowding of tenements in dilapidated buildings, agitated for a constant and clean water supply, instituted a Sewer Commission, and much else. In the face of cholera, he recognized the import of Snow's demonstrations that the source of the 1854 South London epidemic was fecal contamination of the water supplied by the Southwark and Vauxhall Company. But not yet convinced by Snow's conjecture of transmissible microorganisms, Simon qualified Snow's conclusion by reducing it to only one cause among such others as the inhalation of the fecal contaminants. Still, he made efforts to bring the water companies under public control but failed; it was only after the turn of the nineteenth century that such control was achieved. Nonetheless, during Simon's tenure in the City post in Royston Lambert's words (Lambert 1963), "…the City of London….in these years…from being a national scandal became a model of enlightened health government." The mortality data of those years conform with the claim.

In 1855, well before the Great Stench of 1858 had called forth Budd's irony, John Simon had resigned his City appointment and accepted the new national position as Chief Medical Officer of the second General Board of Health, now reconstituted without Chadwick but in practice carrying forward much of what Chadwick had accomplished. Simon's commanding performance in the City of London had made him the obvious candidate. The constant support of the leading medical journals, as well as of the *Times*, consolidated his candidacy for the new national post. Simon accepted despite misgivings about likely Parliamentary obstruction in the developing two-party system of conservative Tories (led by Disraeli) and liberal Whigs (led by Gladstone). Now, his canvas was national.

Chadwick by then had lost favor and much influence, although he never gave up his sanitary crusade. The Parliamentary forces of the Tory party, hostile to any centralizing tendencies, had brought about the dissolution of Chadwick's first General Board in 1854. Much distressed by his grand plans and his overriding of all opposition, his forceful sanitary innovations, and the accompanying devolution of powers to local health authorities had distressed them even more. Benjamin Hall, a libertarian Tory and Welsh nationalist, opposed to central government control and anathema to Chadwick, was the new parliamentary chief in the Privy Council to whom Simon was beholden. To the surprise of many, Simon soon converted Hall from his libertarian views to the cause of public health. Not least surprised were Edwin Chadwick and Florence Nightingale, paired allies inveterately hostile to Simon.

A man of foresight, Simon in his new office was at the outset loath to seek powers of coercion. He was a master of organization, administration, suasion, and effective action as well as of scientific inquiry. His contributions

were at the time unique for any country. In brief, he had devised and developed public health ideas and action based on the best scientific thought in contemporary medicine; he had then applied that theoretical base to create a public health system. The system rested on a structure of national and local programs throughout the country; as the last step he learned to sustain vigilance by monitoring their performance. He aimed to create a corps of doctors trained to these purposes, the foundation for future public health physicians. Chadwick's Poor Law had established local Boards of Guardians, each with a medical officer, to deal with the defeated and deprived populations of the workhouses. While serving the City, Simon adapted their functions to some extent to accommodate broader local public health needs, and solicited from them regular morbidity reports.

Simon did all this with no more permanent assistance from the Privy Council and Parliament than a single clerk. He circumvented that handicap with help from his political superiors in the Privy Council whose confidence he had soon won. From them he obtained departmental allocations to carry out many broader public health tasks, and cajoled the funds for each particular undertaking. His approach to a recognized problem was first to hire a specialist to conduct a thorough scientific investigation, then to frame the policy that could prevent, eliminate or alleviate it, and, finally, to deploy the data in a call for an appropriate Parliamentary Act. For his several individual investigative endeavors, Simon recruited and trained a quite brilliant group of young doctors, most engaged in teaching and research. He would call on them for temporary paid assignments. Some of these assignments he was able to stretch over several years.

His annual reports continued to lay the basis for future action in the form of legislation. The Public Health Act of 1858/9 was a parliamentary Bill inspired and largely crafted by Simon himself, as was all the legislation coming out of his office. Under a broad interpretation of his powers, he established himself in the Privy Council as the equivalent of a civil service secretary responsible to the Minister; in effect the formal equivalent of the directors of all other civil service departments. From that time on he was able to report autonomously in his own right as Chief Medical Officer. He thus achieved control of the National Vaccination Program enacted in 1857, which made vaccination compulsory up to age 13. Able now to hire staff to execute the program, he had consolidated it by 1861. He had also gained the right to initiate inquiries into any health matter, although not the right to enforce the law. He had in effect created, under the aegis of the Privy Council and in another guise, the first national department of health. It was not until 1919 that an appropriately titled and structured Ministry of Health was actually created. Central elements of Simon's organization and approach, however, can be seen to persist in the twentieth and even into the twenty-first century.

Notable new endeavors were several, many summed up in his 1864 Annual Report. Industry and occupation were a focus from 1861 to 1863 (Brockington 1965). Greenhow, investigating health in various districts in England and Wales, concluded that industrial workers' lives were shortened due to lung diseases, including in mining and textile, metal and earthenware manufacture. Greenhow observed differences between men and women, took into account women's occupations, and also noted the health effects on employed children (Greenhow 1860). As a result, Simon recommended government inspection of industries; also the extension to all industry of Nuisance Laws beyond the mines, to which alone they had applied.[2] In 1863, John Bristowe looked at preventable jaw diseases with phosphorus exposure in match factories. William Augustus Guy examined the effects of arsenic and phosphorus exposures in general; Simon's report condemned the use of such toxins and called for inspection and regulation. George Whitley took on lead and mercury. Edward Smith carried out nutritional surveys—presaging modern methods and unparalleled for half a century thereafter—that demonstrated malnutrition first among factory workers and then nationwide; Smith also reviewed the state of printing and tailoring in London, and William Ord that of dressmaking. Overall, it was a tale of hideous conditions and overwork. Mortality rates of printers and tailors were double those of farm laborers of the same age.

This list is only partial, and remains partial even when one adds several other surveys: the cotton famine in Lancashire, resulting from the devastation of the South during the American Civil War; Bristowe's investigation of the as yet uncontrolled druggists; Greenhow's study of the causes of infant mortality in six cities. The rates ranged over 180–220 per 1000 (as against the average for the country of 154 per 1000). These high levels Greenhow reasonably attributed to the conditions he found there: poor sanitation, factory employment of mothers, artificial feeds, indifferent nurses, and dosing with opiates. Hunter in 1864 found "housing for the poor a reproach to the civilisation of England." Each investigation evoked a call for legislation.

One might mention too the investigations of localized outbreaks—conducted from 1863 onward not only in the United Kingdom but abroad—to

[2] Simon's initiatives passed indirectly through the Privy Council, not a major governing entity, to become Bills and then Acts of Parliament. These included the following: 1863 Nuisance Act, giving power to condemn bad meat and other food; 1866 Sanitary Act, opening the second era of public health in England; 1866 Factory Acts requiring adequate cleanliness and ventilation to control industrial diseases; 1867 Gangs Bill, regulating farm gangwork by women and children; Merchant Shipping Act, authorizing medical inspection aimed at abolishing scurvy; Sewage Utilisation Bill giving power, previously limited to large cities, to all Local Authorities to enforce adequate drainage and sewage; Rural Housing Act, shifting authority to condemn unfit housing from Parochial authority to Poor Law Unions; 1875 Sanitary Act.

prepare the country for possible epidemic spread. Whitley studied typhus in St. Petersburg, John Scott Burdon-Sanderson cerebrospinal meningitis in Russia (as yet then unknown in Britain), and Netten Radcliffe possible cholera in the Mediterranean. John Gamgee found that up to 20% of the meat in the country came from diseased animals, and later Thudicum spent eight months observing the slaughterhouses, markets, and butcher shops to produce information barely added to for another century.

The foresight that led to these alert responses marked Simon's recognition of Semmelweis's work in 1863, and his deferred acceptance of contagion rather than miasma as the agent of such epidemic diseases. The significance of Simon's conversion from miasma in the climate of those times is made plain by the exacerbated hostility toward him that it provoked on the part of both Chadwick and Florence Nightingale. William Farr, by contrast, regardless of his early approval of Henle's hypotheses regarding infectious diseases, was still able to give important and friendly support to Florence Nightingale in the form of data and analysis.

Simon had a flexible mind, and he underwent still another political and practical conversion: he was wooed away from the niceties of suasion toward the forcefulness of sanctions. Pressed by experience, he became reconciled to the necessity for enforcement. The failures of local authorities to execute their assigned duties under the new laws persuaded him to change course. It became clear that if his plans for an effective public health program were to be realized, both central authority and central direction of implementation were essential. This set Simon on the path to his new concept of State Medicine, which in modern terms translates into broad central responsibility for the public health.

An early intimation of this change followed from George Buchanan's study of 25 towns, each of which had implemented the 1866 Sanitary Act. The object was to test the effects of sanitary improvements before and after their introduction. Although improvements in mortality rates in typhoid and so-called filth diseases had occurred, he was surprised, at the same time, by the marked decrease in the rates of pthisis (tuberculosis)—in some towns the only important change that had occurred. "True," wrote Simon, "the results which have been elicited are not in all cases equally intelligible—but on the whole can be well understood." This cautious note indicates that by that time, in his Ninth Report in 1867, he was aware that pthisis was unlikely to be associated with sanitation, but wanted to emphasize the positive nature of the results.[3]

Simon's successive inquiries into local health conditions across the country revealed to him the depressingly negligible efforts and negligence of

[3] Robert Koch published his finding of the tubercle bacillus in 1882.

local health authorities to perform their assigned tasks under the new laws. Particularly galling was the failure, even the active obstruction, of local agencies in fulfilling their obligations of compulsory vaccination for smallpox. By the late 1860s, he had established a corps of trained and licensed public vaccinators with an emphasis on the newborn and infants. Despite those efforts, in 1871 the smallpox epidemic claimed 21,000 lives and, from 1871 to 1873, 44,000 in all. At this stressful time, having recognized that immunity acquired by vaccination in infancy was not lifelong, he instituted additional obligatory vaccination at puberty.

During this period, although John Simon was politically a liberal reformer and a supporter of the Gladstone government, ironically a hostile alliance formed against him. This consisted of Stansfield, the Minister to whom he was responsible, Chadwick, Nightingale, the Board of Guardians, and various other protagonists or defenders of the Sanitary Movement and the old Poor Law. Step-by-step they succeeded in stripping Simon of the powers to act centrally that he had acquired over the years. Ironically again, the successor Tory government led by Disraeli was returned in the winter of 1874 and stood against centralized powers, yet it nonetheless restored centralized health administration. Extensive legislation on health also restored both central powers and local obligations.

Although Simon did not regain all, he still had a major role in the 1875 Public Health Act, which consolidated 30 years of legislation and experience. There was nothing comparable elsewhere at that time, and in England it stood untouched for the 60 years until 1936. Yet the department suffered a loss of funds that threatened Simon's vaccination and other programs. Simon, uncharacteristically confrontational in these later years, issued an ultimatum that was rejected. In consequence, he resigned at the age of 59. Although his 1875 Act stood, this untoward outcome dismantled a good deal of both the administrative structure of his department and its scientific approach to public health and legislation.

Major Greenwood, professor of epidemiology in the London School of Hygiene from its founding in 1928, dubbed Simon a great opportunist. It has also been said that his acts outstripped his thought. The description could perhaps be justified by contradictions between his ideas and his actions. These are exemplified by his strong faith in individualism as against his adoption of centralized legislation and administration. Thus the medical department, which Simon established as Chief Medical Officer of the Privy Council, has been seen not only "as its own progenitor" but also, in essence, as the nidus of a national Ministry of Health. It was the means by which Simon created a national model of scientifically based and centrally organized public health. In creating his department, he had to work against a strong English tradition of local autonomy and limited central powers. Yet he

strengthened public health at the local as well as the national level. Simon's Sanitary Act of 1866 together with "the great" consolidating Sanitary Act of 1875 had brought about a major shift. The acts together both secured central government powers over local authorities, and entrenched local authority powers over individuals.

In the result, Simon's notion of state medicine was embodied in his creation of a national public health program administered by local authorities. His program proved to be a guide to future public health movements and official organization in many other countries. The scope of his legacy, based as it was on faith in scientific medical and epidemiological principles, goes well beyond matters of legislation, organization, administration, and implementation. In executing such programs, Simon created a now familiar epidemiological and public health model for investigating the extent and nature of defined problems, devising appropriate remedies from the observed conditions, and implementing them through legislation and further investigation to determine the effects of regulatory intervention. While executing his encompassing public health functions as Chief Medical Officer, moreover, he had sustained his scientific interests and his grasp of modern advances in bacteriology, pathology, and medicine generally, all of which he brought to bear in developing policies for prevention and control. He left an indelible mark on the content of the modern discipline of an epidemiologically based public health.

References

Brockington F (1965). *Public Health in the Nineteenth Century.* Edinburgh: E. & S. Livingstone.

Greenhow, EH (1860). On the different prevalence of certain diseases in different districts of England and Wales. In Levi, ed. *Annals of British Legislation: Being a Classified and Analysed Summary of Public Bills, Statutes, Accounts and Papers, Reports of Committees and of Commissioners, and of Sessional Paper Generally, of the Houses of Lords and Commons*. London: Smith, Elder, & Co., pp. 80–150.

Hamlin C (1998). *Public Health and Social Justice in the Age of Chadwick: Britain, 1800–1854.* Cambridge: Cambridge University Press.

Jarvis E (1855, republished 1971). *Insanity and Idiocy in Massachusetts: Report of the Commission on Lunacy.* Cambridge: Harvard University Press.

Lambert R (1963). *Sir John Simon, 1816–1904 and English Social Administration.* London: MacGibbon & Kee.

Simon J (1890). *English Sanitary Institutions, Reviewed in Their Course of Development and in Some of Their Political and Social Relations.* 2nd ed. London: Cassell.

Wohl AS (1983). *Endangered Lives: Public Health in Victorian Britain.* London: J.M. Dent.

Worboys M (2000). *Spreading Germs: Disease Theories and Medical Practice in Britain, 1865–1900.* Cambridge, UK: Cambridge University Press.

10

Germ Theory, Infection, and Bacteriology

In the 1860s, medicine saw major developments at the biochemical, cellular, and molecular level. For the early preliminary work and the later revolutionary and transforming discoveries we look primarily to Europe. The tangible beginnings of the new and revolutionary developments arose first in France, and were then followed up in Germany. From 1835 on, a number of microorganisms had been recognized in association with various pathological conditions (Table 10.1), but no firm causal attributions to disease had been made (Evans 1993). Evans reported that in 1837 Donné had observed the *trichomonas* organism in vaginal infections, in 1854 Pacini had observed the *vibrio* in cholera, and later Virchow and Zender had observed the worm of *trichinosis*. At the time, however, none of these observations were assigned a causal role in disease.

Louis Pasteur: Founding Germ Theory

Louis Pasteur (1822–1895) was trained as a chemist. Early on, however, his talents and interests had engaged him in problems with a biological cast. In France and elsewhere, his name is justly revered as a founder of our understanding of infection. Pasteur went on to become a commanding figure in

Table 10.1 Some Early Discoveries in Bacteriology, 1835–1889

Year	Disease	Cause	Discoverer
1835	Diseases of silkworm	Fungi	Bassi
1837	Vaginal infections	*Trichomonas vaginalis*	Donné
1844	Skin diseases	Fungi	Gruby
1850	Anthrax in animals	Anthrax bacillus	Davaine, Rayer
1854	Cholera	*Vibrio cholerae*	Pacini
1855	Anthrax in animals	Anthrax bacillus	Pollender
1860	Trichinosis	Trichinae	Virchow, Zender
1862	"Diseases" of wine	Bacteria	Pasteur
1865	Diseases of silkworms	Bacteria	Pasteur
1868	Relapsing fever	Spirilla	Obermeier
1872	Septicemia	Bacteria	Coze, Feltz, Davaine
1876	Anthrax in humans	Anthrax bacillus	Koch
1878	Wound infection	Bacteria	Koch
1879	Gonorrhea	Bacillus	Neisser
1880	Typhoid fever	Bacillus	Eberth, Gaffky
	Leprosy	Bacillus	Hansen
	Malaria	Parasite	Laveran
1882	Tuberculosis	Tubercle bacillus	Koch
	Glanders	Bacillus	Löffler
1883	Cholera	Bacillus	Koch
	Erysipelas	Bacillus	Fehleisen
1884	Diphtheria	Bacillus	Klebs, Löffler
	Tetanus	Bacillus	Nikolaier, Kitasato
	Pneumonia	Bacillus	Fraenkel
1887	Epidemic meningitis	Bacillus	Weichselbaum
	Malta fever	Bacillus	Bruce
1889	Soft Chancre	Bacillus	Ducrey

Adapted from Evans 1993.

several aspects of the science of the nineteenth century. A special citation on his election to the French Academy lists six major original achievements (Vallery-Radot 1960; Dubos 1986; Debré 1998).

His first major contribution related to fermentation, a biological process critical for the industries of brewers, bakers, and food handlers. Long-standing theory held that microorganisms originated as the *de novo* products of spontaneous generation resulting from fermentation; this would explain why maggots characteristically appeared in meat allowed to rot. As noted previously, in Italy the experiments of Francesco Redi had first challenged the theory by experiment (Farley 1977). Redi's simple experiment led him to reverse the then accepted causal order that fermentation produced microorganisms: ferments were produced by living organisms. Redi later concluded that all ferments were similarly produced by living organisms, and were not products of the ferment. A century after Redi, the Abbé Spallanzani (1729–99) demonstrated, in a series of experiments, that Redi's finding held

true for organisms appearing with the putrefaction of meat. Yet the idea of the spontaneous generation of organisms in fermentation persisted unquenched well into the nineteenth century (Farley 1977).

Thus, by mid-century, a number of scientists had provided evidence against the theory of spontaneous generation of microorganisms. That is not to say that the proponents of the contrary illusion were silenced. For one reason or another none of the proponents of living germ theory had quite closed the case (Winslow 1943). Finally, in 1861, Louis Pasteur reported his first major finding in what would be a fruitful and indeed a spectacular research career. In experiments he demonstrated the consistent absence of fermentation once organisms were removed to produce thoroughly sterile conditions. He also proved the converse of the existing theory: fermentation did not generate microorganisms, but was itself the product of specific microorganisms. Pasteur's famous experiments with sterilized flasks decisively refuted the theory of spontaneous generation. Despite his work, it was several decades before consensus was achieved.

Pasteur's refutation of spontaneous generation and his demonstration of the autonomous existence of living microorganisms was the foundation for a second major accomplishment. France was then as now a major

Figure 10.1 Louis Pasteur (1822–1895). Credit: Courtesy of the National Library of Medicine.

wine-producing country. In 1864, his interest was engaged in the growing problem of wines souring or otherwise going off. Trade under a new commercial treaty with England was threatened. On the assumption that bacterial invasion of the casks was the cause, Pasteur resorted to the microscope (the instrument was evidently primitive even by those early standards). He soon showed that the deteriorated wines contained bacteria, and that the simple process of heating to 50 to 60 degrees centigrade eliminated the pathogens (Dubos 1986; Debré 1998).

Surprisingly, a discovery that could and did save an important industry was given little credence or credit at the time. The neglect was corrected in 1867 when he was awarded the Grand Prix exhibition medal for saving the wine industry. In the year following the epidemic in the vineyards, another economically destructive agricultural epidemic arose to affect the equally important silk industry. On microscopic examination, Pasteur again traced the unquestionable source to microorganisms specific to the epidemic manifestation: the silkworms were seemingly unaffected, but the moths that emerged from the chrysalis were clearly infected and quickly died (Dubos 1986; Debré 1998).

In respect of these discoveries, Pasteur was much more than a savior of industry. This work on fermentation and on infection provided a sound factual basis for a theory of disease caused by bacteria—"germs" in the

Figure 10.2 The Rabies Laboratory in the Pasteur Institute, Paris, France. Apparatus was the same as used by Louis Pasteur. Credit: © Bettmann/Corbis. Reproduced by Permission.

common parlance of the time—and shaped new and directly applicable concepts of infection by microorganisms. Thus, beginning with work applied to agriculture and animal husbandry, Pasteur's discoveries proved equally applicable to the health disorders of human beings. He gave force and direction to two sets of ideas fundamental to medicine and public health. These ideas centered on disease processes and their immediate antecedents, and they retain their force and centrality undiminished to the present day.

The first instance had to do with his work on the unwanted fermentation that afflicted major industries in the maturing of wine and the brewing of beer. In the course of this work, Pasteur established by microscopy and experiment a first basic principle: a specific microorganism had effects specific to it. This key causal concept of the germ theory, first formulated as an hypothesis by Jacob Henle and subsequently at least implicit in the work of Panum, Semmelweis, Snow, and Budd, had never before been demonstrated in actuality. Pasteur's unequivocal result had set biomedical science on a totally new course, with momentous consequences for health and sickness.

Yet the hold of the entrenched concepts of spontaneous generation along with miasma was made manifest once more by the delay in applying the work of contagionists (also see Morabia 2007; Oppenheimer and Susser 2007). Many doctors and their teachers continued to adhere to outmoded ideas and the dangerous practices arising from the neglect of the new knowledge. At first, Pasteur's path-breaking discoveries went largely unheeded, much as had the earlier epidemiological work as well as the demonstrably dramatic effects of strict antisepsis in surgery as first argued for and as practiced in Britain by Joseph Lister. As witness, in Pasteur's native France, in the mid-1870s high mortality from puerperal sepsis persisted unabated.

A second set of Pasteur's ideas fundamental to biology, medicine, and epidemiology related to host immunity as a defense against disease. No significant advance in this area had occurred since Jenner who, in his treatise of 1795, had shown that cowpox vaccination protected against smallpox. Pasteur was the first to develop Jenner's creation of a vaccine into a generalization applicable to many microorganisms. We leave the account of this second fundamental development and Pasteur's contribution to Chapter 11.[1] Following the antecedent handful of epidemiological revelations of infectious disease transmission described in Chapter 9, however, Pasteur had confounded his critics and opened a new way forward. With his discoveries, bacteriology and then immunology had begun their development as sciences.

[1] As is only to be expected from the theories of Fleck and Kuhn, the now revered Pasteur was not without opposition (Hellman 2001). In this field and others, virtually all early nineteenth-century pioneers and major innovators who challenged miasma theory faced opposition.

The Germ Theory: Koch, Klebs, and Mycoplasma Tuberculosis

Following on from the work of Pasteur, the center engaged in the most productive pursuit of pathological organisms moved to Germany, then only a collection of principalities lately unified into a powerful Bismarckian state. The contemporaries Edwin Klebs (1834–1913) and Robert Koch (1843–1910) were especially prominent and prolific among bacteriologists. Each isolated a fistful of microbial pathogens, sometimes almost concurrently for the same diseases. Klebs did not garner as many garlands of fame as did Koch, although the bacterial genus *Klebsiella* was named for him, and as noted below he was scarcely less productive. These developments are covered in some detail by Evans and also by Carter (Carter 1985, 2001; Evans 1993).

Robert Koch (1843–1913) isolated the anthrax bacillus in 1876 as noted in Table 10.1. Famously, this marked the first human disease that could firmly and definitively be attributed to a specific microorganism. Koch was a brilliant prize-winning student of Jacob Henle, and not surprisingly he adopted Henle's theoretical speculation that every disease might be caused by a specific organism. By 1876 Koch was District Medical Officer and country practitioner in Wolsztyn (formerly in Prussia, now Poland), and soon was equipped with a thriving laboratory in which many new organisms were isolated. Among his many discoveries of pathogenic bacteria, Koch's greatest renown stemmed from his most dramatic report and discovery of the mycoplasma for tuberculosis, the cause of the disease known up to that time as phthisis (Winslow 1943; also see Koch 1882). In doing so, he had announced the proximal cause of the most widespread, chronic, debilitating, and often fatal disease of nineteenth-century Europe. Its distribution in populations has varied through time, but in that period, phthisis was the most dreaded fatal disease of younger women. For this discovery and others, Koch won the Nobel Prize in 1905 (Brock 1999).

Koch was invited to announce his discovery before a highly distinguished audience, including Rudolf Virchow among others, at a gathering of the Berlin Physiological Society on March 28, 1882. In the course of describing his research and its outcome, he spelled out a first version of four postulates required to confirm a specific organism as cause of a specific disease. The postulates demand a one-to-one causal relationship and thus conform to the Galilean formulation of necessary and sufficient cause. "Koch's postulates," in a later refinement of his several previous versions, are set out in the following passage:

> "first, the organism is always found with the disease, in accord with the lesions and clinical stage observed; second, the organism is not found with any other disease;

Figure 10.3 Robert Koch (1843–1913). Credit: Courtesy of the National Library of Medicine.

> third, the organism, isolated from one who has the disease and cultured through several generations, produces the disease" [i.e., in a susceptible experimental animal]. Koch then relented from excessive rigor and added a fourth criterion: "Even where an infectious disease cannot be transmitted to animals, the 'regular' and 'exclusive' presence of the organism" [i.e., the first two postulates are satisfied] "proves a causal relationship."[2,3]

It is an irony that the tubercle bacillus, the discovery that first made him famous, does not conform to the postulates: the relationship between the bacillus and the symptoms of the disease do not relate so consistently in one-to-one fashion.[4] It soon became clear that practical considerations stood in the way of fulfilling all the postulates. Human pathology cannot always

[2] Kindly translated by my former student and colleague, Holger Hansen.

[3] One might translate the original postulates, as in the revised restatement below, using language for causal criteria consonant with modern epidemiology and in a way that is more general than for infectious causes alone (Susser 1991).

[4] A causal attribution requires an invariable association of bacterial organism with a disease manifestation (postulate 1) that is wholly specific, unique, and unconfounded by manifestations of other diseases (postulate 2). The association must meet the criterion of biologic coherence in maintaining consistency on prolonged bacterial culture and in its effects in repeated animal experiment

be replicated by experiment. For instance, Koch himself, when he had first linked the *vibrio* of cholera to that disease, was unable to transmit the organism in the laboratory and thus could not replicate the signs of the disease experimentally. Organisms pathogenic to human beings like Hansen's leprosy bacillus and Noguchi's syphilis spirochete among others, although in ever diminishing numbers, have not yet been cultured nor transmitted to experimental animals (with one natural exception, for leprosy in the armadillo) (Scollard et al. 2006).

This degree of indeterminacy does not invalidate any causal contribution. Other than animal experiment, the grounds remaining for attributing a causal role have many times proved valid by consistent replication in continuing experience. Indeed, such approaches are part of the basic armamentarium of science. From a modern perspective, a more serious shortcoming is that Koch's formulation leaves no room for a chain or complex of multiple causes antecedent to manifestations of disease. The logic of Koch's words, taken literally, allows for only a single outcome of any particular infection. Causes can be plural, as John Stuart Mill made clear in the nineteenth century. Indeed, causality is virtually always plural depending on the frame of reference and the concept applied. Causes of a single outcome can be simultaneous, or proximal but antecedent to the outcome, and may act in one or several dimensions.

By 1890, Koch was making a less mandatory statement of his postulates (translated by Rivers):

> However, if it can be proved: first that the parasite occurs in every case of the disease in question, and under circumstances which can account for the pathological changes and clinical course of the disease; secondly, that it occurs in no other disease as a fortuitous and nonpathogenic parasite; and thirdly, that it, after being fully isolated from the body and repeatedly grown in pure culture, can induce the disease anew; then the occurrence of the parasite in the disease can no longer be accidental, but in this case no other relation between it and the disease except that the parasite is the cause of the disease can be considered. (Rivers 1937)

These words ascribed to Koch notably relax the demand for a causality of strict singularity, in requiring only that the organism "occurs in no other disease as a fortuitous and nonpathogenic parasite." Beside this less stringent criterion, here Koch insists on only two essential conditions for a causal organism; first, that it is always present in cases with all the clinical features

(postulate 3). Failing animal experiment, in repeated observations the association must meet the criteria of consistency and predictive performance (postulate 4).

of the disease; and second, that once isolated in pure culture from an affected person, it can in experiment repeatedly induce the disease anew.

By 1891, however, Koch had relaxed the postulates still further. It had become clear to him that diseases could conform in every respect excepting only that a number of distinctly pathogenic organisms, some of which he listed, could not be experimentally transmitted. The implicit and mistaken demand that the organism not be found associated with or responsible for any other disease remains. He was not yet aware of the existence of nonconforming carrier states, in which infection with potentially pathogenic organisms does not produce the expected disease. An additional weakness clear to the modern eye, given that Koch must surely have been aware that Pasteur had recognized immunity and had framed the idea as a theoretical construct, is that the postulates take no account of it. Immunity seems not yet to have become a regular part of the causal equations affecting the occurrence and distribution of infectious disease.

For at least three-quarters of a century after Koch had formulated the postulates, innumerable students and graduates of epidemiology and medicine learned and recited the celebrated if flawed postulates carrying his name. Belatedly, medical historians and epidemiologists have come to recognize predecessors unacknowledged in Koch's triumphant initial presentation and report. He made no mention of the original formulation of his teacher Jacob Henle's criteria for the attribution of disease to a specific microorganism. Fiftten years thereafter, these criteria seem first to have been filtered through an 1877 paper by Edwin Klebs on tubercular disease (Klebs 1877; Henle 1840/1938), which Koch also neglected to cite. In that paper Klebs, in reporting his successful transmission of tuberculosis organisms to animals from the milk of infected cows, had also cited Henle's criteria for specific infection of a half-century before. At a presentation a month later, however, Koch acknowledged Klebs's contribution in a gracious exchange. Klebs rose to speak to his prior work with no less grace and without detracting from Koch's achievement.

Klebs was scarcely less fertile than Koch in bacteriological discoveries, among them the pathogenic organisms of typhoid and diphtheria in 1884 (Baumgartner 1935), and he followed closely behind Koch in priority for some other attributions of causality to microorganisms. An emergent consensus designates the postulates by the eponym Henle-Koch or, clumsily but justly, Henle-Klebs-Koch (Table 10.2) (Evans 1976; Carter 1985, 2001).

The postulates served well to establish specific microorganisms as the causes of specific diseases, and they met Galileo's prescription of "necessary and sufficient cause." In later times, however, epidemiologists have recognized that the postulates are vulnerable to fallacy. One example arose with

Figure 10.4 Edwin Klebs (1834–1913). Credit: Courtesy of the National Library of Medicine.

the "carrier states" mentioned earlier, a condition found with several bacteria. Epidemiologists had learned before the end of the nineteenth century that pathogenic organisms such as corynebacterium diphtheriae and mycoplasma tuberculosis can exist in the body without causing the disease with which they are typically associated (Evans 1993). Once understood, the existence of unaffected carriers illuminated many obscurities of epidemic spread.

Among multiple outcomes of a single cause, one must allow for the absence of any effect. In the early twentieth century in New York, "Typhoid Mary" earned her soubriquet as a notorious carrier of the bacillus. Despite warnings, she persisted in taking employment as a cook and thereby infected many diners. Exasperated health authorities finally saw no recourse other than keeping her isolated in quarantine (Leavitt 1992).[5]

[5] Mary Mallon was forced to live for more than 26 years in what became nearly lifelong isolation. Sadly, officials never focused on educating Mallon on methods of hygiene that would have drastically reduced her infectivity (for instance, stringent hand-washing) or ensuring that she work in an appropriate occupation away from food preparation (Leavitt 1992).

Table 10.2 Contributions of Edwin Klebs and Robert Koch

Year	Klebs	Koch
1872	Bacteriology of gunshot wounds	
1873	Produced tuberculosis in animals by injecting milk from infected cows	
1876		Treatise on cause of anthrax
1877	Treatise on cause of anthrax Studies on tuberculosis Postulated on causation	
1878	Inoculated primates with organism of syphilis	Studies of septic infection of animals
1880	Description of typhoid bacillus	
1881		Introduction of solid culture media
1882		Treatise on etiology of tuberculosis
1883	Description of diphtheria organism	
1884	Studies on cholera	Isolated cause of Asiatic cholera
1885	Worked on etiology of cholera	
1889	Studies on malaria	
1890		Tuberculin discovered
1891		Human and bovine TB are different
1892	Pathology and treatment of cholera	
1893		Studies on Rhodesian red water fever, recurring fever, trypanosomiasis
1896		Investigates rinderpest, sleeping sickness
1905		Awarded Nobel Prize
1908	Use of attenuated TB organism to produce immunity in guinea pigs and humans	

Adapted from Evans 1993.

By now, we have learned that many pathogens capable of producing dire effects commonly exist also as silent infections in human carriers. Among the many such are the polio virus in poliomyelitis, entamoeba histolytica in amoebic dysentery, the Klebs-Loeffler bacillus in diphtheria and a great many others. These organisms are able to exist in the body without causing any disease. Sometimes the host is immune; sometimes the organisms themselves belong to a strain less virulent than those capable of producing disease; at other times the organisms are present in insufficient numbers. Such silent infections are nonetheless capable of transmission to other persons. As long ago as 1884, Friedrich Lóffler first demonstrated the existence of silent carriers in his laboratory studies of the diphtheria bacillus

(both cocci and bacilli had been identified by Edwin Klebs in 1883, hence the eponym Klebs-Lóeffler bacilli). Lóeffler was able to show that only the bacillus produced a diphtheria-like illness; he went on to describe a "carrier state," whereby healthy children carried that bacillus in their throats without succumbing to illness (English 1985). Koch also had recognized the danger of convalescents transmitting disease. Aware as well that healthy carriers existed, however, Koch did not consider the organisms they carried to be similarly as dangerous as those in the persons affected by the same organisms (Rosen 1958, 1993).

The fact is that causality involving living organisms is a complex matter readily obscured by unexpected vagaries of behavior. Max von Pettenkofer (1818–1901), occupant of a chair in Munich, was the leading German hygienist of his time. Like John Snow, Pettenkofer was impelled to study cholera by the epidemic of 1854. He is credited among much else with introducing experimental methods in hygiene. A famous experiment he conducted on himself bears narrowly on the question of silent infection by otherwise virulent organisms, but also bears broadly on the history of changing paradigms in epidemiology. The story long remained an enigma. It unfolded only recently, providing a fresh appreciation of the protean manifestations of pathological microorganisms.

Pettenkofer had rejected John Snow's conclusions in his reports of 1859 bearing on cholera outbreaks. His vision was closer to that of the miasmatists than the developing demand for specificity of effects found allowable. His was not a narrow view, however, and he had a sense of multidimensional causality (Morabia 2007; Oppenheimer and Susser 2007). In voluminous writings through the end of the century Pettenkofer continued to reject all such specific and limited mechanisms of disease transmission. With good reason, he insisted on the need to take account of environmental states and changes in them. In 1865 he had expounded his own theory at length, and he elaborated it over the years in terms of complex relations between seasons, humidity, and the physical state of soils. In disputing the spread of cholera by contagion in its literal sense, he cited incidence data among doctors and others caring for affected patients that were no greater than among people at large (Figure 10.5).

Pettenkofer's alternative theory explaining cholera was a hybrid that reduces to two critical variables, the organism and miasma. Manifest disease, he argued, arose from the poisons of a localized kind of miasma (echoes of Chadwick's supposition to account for local outbreaks). This miasma emanated from the conjunction of two main elements: specific *vibriones* carried by excrement and the like, and the conditions under which cholera organisms sustained virulence, namely, only in soil left moist by declining groundwater level. In fact, his theory was virtually an extension of Snow's

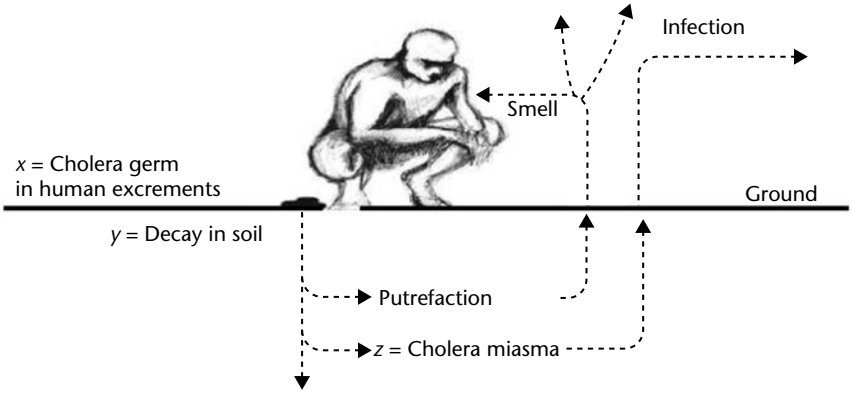


Figure 10.5 Depiction of Max von Pettenkofer's groundwater level theory. From: Morabia A (2007). Epidemiologic interactions, complexity, and the lonesome death of Max von Pettenkofer. *Am J Epidemiol* 166(11):1233–8. Reprinted by permission of Oxford University Press.

imputation of the source of infection, which was to say, the ingestion of water fouled by excrement and thereby contaminated by living and multiplying microorganisms.

Although Pettenkofer in 1892 was not yet ready to concede all to Koch's germ theory of 1882, his own suppositions relating cholera outbreaks to season, rainfall, humidity, and moisture of the soil were not that distant from Snow's interpretation of his own studies. Doubtless unaware of the hazards of ecological studies (Susser 1994a, 1994b), some of his followers produced seemingly supportive statistics. To disprove the causal claim specific to the *vibrio*, Pettenkofer devised a brave and dramatic personal experiment (Winslow 1943). He and his assistants each drank a glassful of water containing *vibriones*. Surprisingly, Pettenkofer and one assistant escaped with little ill effect, although another did not.[6]

The above historical sketch provides as good an example as any of a revolutionary shift in theory in accord with the Kuhnian paradigm described in

[6] A new study reported in 2002 appears to have solved the long-lasting mystery of Pettenkofer's impunity in this experiment (Merrell et al. 2002). Surprisingly, the new study found that cholera vibriones, collected from contaminated river water and cultured, lost virulence and seldom affected volunteers. Vibriones collected directly from the stools of cholera patients in Bangladesh, by contrast, exhibited high levels of virulence. When fed to mice and collected from the small gut, these vibriones were 10 to 100 times as many, and up to 700 times as virulent, as those recovered from river water. They were transformed (probably in rapid mutations) during their passage through the human gut. The result points not only to the extraordinary precision attained by molecular techniques, but equally to the large variations that can be induced by refined interaction under different environmental conditions.

Chapter 2. Within a mere decade after Koch following Pasteur propounded the criteria for germ theory, consensus had been reached among the generality of medical scientists. As expected from a new concept, the theory guided the pursuit and discovery of new kinds of facts. In an arena of free thought, such large changes in the theory and practice of science, and consequently in public policy, have often been much slower to eventuate.

The demise of miasma theory for instance, survived the succession of early challenges from Budd, Panum, Semmelweis, and Snow. Even the experimental and practical support of germ theory, consummately demonstrated by Pasteur and then Koch, did not at once demolish miasma theory. The new causal theory clearly demanded a course of public health investigation and intervention quite different from any required by entrenched miasma theory. Considerable resistance came from its adherents among both physicians and public health officials. The nineteenth century had turned before the new science of bacteriology had finally dispersed the metaphorical miasma of miasma theory and quickly came to rule the ideas of medicine, public health, and their faculties both in Europe and the United States.

Bacteriology flourished and grew powerful in its pursuit of one new organism after another, each linked to a particular infectious disease. The eponyms attached to each new organism were many, comfortingly assuring each discoverer that his name (virtually all were male) would long be known. For medicine the hunt became a major preoccupation. Among those pursuing the health sciences in the halls of academe, the concentration and narrowed focus of scientific endeavor downgraded and crowded out a more contemplative mode. In the previous era a broad-gauged primacy of humane concerns had seemed to predominate or was at the least more evident in the practice of medicine. Tested and specific scientific knowledge of human psychological function and emotional need, still scarce in our time, was not yet in the gift of medicine, psychiatry, or psychology. Tested knowledge at the societal level was even less so.

In the epidemiology of the half century before World War II, microbiology grew dominant as a basic scientific underpinning of the discipline. In the epidemiology of that period, methods of investigation concentrated on the search both for sources of outbreaks of infectious disease and for the modes of transmission of microorganisms. Some notable exceptions include, among others, the work of Goldberger and Sydenstricker (Kraut 2003; Hardy 2004), Major Greenwood (Greenwood 1935); and Donald Budd Armstrong (Susser and Stein 2005). Familiar with the fast-growing pace of current developments, all in medicine today are familiar with an evolutionary process that absorbs one new revelatory technique after another. The process compels both education and practice constantly to adapt to new modes of thought, concepts, and discovery.

References

Baumgartner L (1935). Klebs: a centennial note. *New Eng J Med* 213:60–63.

Brock TD (1999). *Robert Koch a Life in Medicine and Bacteriology.* Washington, DC: ASM Press.

Carter KC (1985). Koch's postulates in relation to the work of Jacob Henle and Edwin Klebs. *Med Hist* 29:353–374.

Carter KC (2001). Edwin Klebs's Grundversuche. *Bull Hist Med* 75:771–781.

Debré P (1998). *Louis Pasteur.* Baltimore: Johns Hopkins University Press.

Dubos RJ (1986). *Louis Pasteur, Free Lance of Science.* New York, NY: Da Capo Press.

English PC (1985). Diphtheria and theories of infectious disease: Centennial appreciation of the critical role of diphtheria in the history of medicine. *Pediatrics* 76:1–9.

Evans AS (1976). Causation and disease—Henle-Koch postulates revisited. *Yale J Biol Med* 49:175–195.

Evans A (1993). *Causation and Disease: A Chronological Journey.* New York: Plenum.

Greenwood M (1935). *Epidemics and Crowd Diseases*. London: Williams and Norgate.

Farley J (1977). *The Spontaneous Generation Controversy From Descartes to Oparin*. Baltimore: Johns Hopkins University Press.

Hardy, A (2004). Methods of Outbreak Investigation in the 'Era of Bacteriology' 1880–1920. In Morabia A, ed. *History of Epidemiologic Methods and Concepts.* Boston: Birkhauser, pp. 199–206.

Hellman H (2001). *Great Feuds in Medicine Ten of the Liveliest Disputes Ever.* New York: Wiley.

Henle J (1840). *On Miasmata and Contagie*. See 1938 transl. George Rosen. Baltimore: The Johns Hopkins Press.

Klebs E (1877). Uber Tuberkulose. *Munich, Amtl Bericht ver Versamulung, Deutsch Naturforsch Aertze* 50:274.

Koch R (1882, 1982). Classics in infectious diseases. the etiology of tuberculosis: Robert Koch. Berlin, Germany 1882. *Rev Infect Dis* 4:1270–1274.

Kraut AM (2003). *Goldberger's War: The Life and Work of a Public Health Crusader.* New York: Hill and Wang.

Leavitt JW (1992). Typhoid Mary strikes back—bacteriological theory and practice in early 20th-century public-health. *Isis* 83:608–629.

Merrell DS, Butler SM, Qadri F, Dolganov NA, Alam A, Cohen MB, Calderwood SB, Schoolnik GK, Camilli A (2002). Host-induced epidemic spread of the cholera bacterium. *Nature* 417:642–645.

Morabia A (2007). Epidemiologic interactions, complexity, and the lonesome death of Max Von Pettenkofer. *Am J Epidemiol* 166:1233–1238.

Oppenheimer GM, Susser E (2007). Invited commentary: the context and challenge of Von Pettenkofer's contributions to epidemiology. *Am J Epidemiol* 166:1239–1241.

Rivers TM (1937). Viruses and Koch's postulates. *J Bacteriol* 33:1–12.

Rosen G (1958, 1993). *A History of Public Health*. Baltimore: Johns Hopkins University Press.

Scollard DM, Adams LB, Gillis TP, Krahenbuhl JL, Truman RW, Williams DL (2006). The Continuing Challenges of Leprosy. *Clin Microbiol Rev* 19:338–381.

Susser M (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am J Epidemiol* 133:635–648.

Susser M (1994a). The logic in ecological: I. The logic of analysis. *Am J Public Health* 84:825–829.

Susser M (1994b). The logic in ecological: II. The logic of design. *Am J Public Health* 84:830–835.

Susser M, Stein Z (2005). Commentary: Donald Budd Armstrong (1886–1968)—pioneering tuberculosis prevention in general practice. *Int J Epidemiol* 34:1191–1193.

Vallery-Radot R (1960). *The Life of Pasteur.* New York: Dover Publications.

Winslow CEA (1943). *The Conquest of Epidemic Disease*. Princeton, NJ: Princeton University Press.

11

The Concept of Host and Immunity

In the vocabulary of epidemiology of the past century and more, the common usage of the word "host" brings to mind either an affected subject or a carrier of transmissible disease. The idea seems to have evolved into its current form in the late nineteenth century. For medical scientists and epidemiologists in particular, the role of microorganisms in disease had moved to the center of attention. Pettenkofer's triad of host, organism, and environment made the affected person a necessary and indeed a fundamental element in any manifestation of disease. We noted a related idea in Hippocratic writings in earlier chapters. The "humoral constitution" was seen as a crucial element in the reaction to the environment that was thought to precipitate disease. Thus the notion ascribed to individual reaction an importance in the manifestation of disease equivalent to that of the external effects of environment and active agents.

Smallpox: Jenner and Mass Prevention Realized

Smallpox is a disease that, until the twentieth century, had long been among the greatest of human plagues (Hopkins 2002). In the last quarter of the eighteenth century, Edward Jenner (1749–1823) set public health on the path

to its eventual virtual eradication in the late twentieth century. Yet after Jenner first published the report of his work on vaccination in 1798, a century elapsed before the unprecedented large-scale success of vaccination to prevent smallpox was again attained with other organisms, first by immunization for diphtheria, and then for other infections. New concepts were a prerequisite for generalization from the particular practice of smallpox vaccination to immunization against a wide range of diseases. Until the concept of immunity was clearly formulated, effective immunization was confined to the one condition in which Jenner had discovered it to be effective.

Previous to Jenner, a primitive sense of immunity had existed for centuries. Some items of that history remain relevant. Thus the fact that survivors of smallpox were immune to its recurrence was occasionally noted. A quotation from Francis Bacon's journal in 1626 (cited under "contagion" by the Oxford English dictionary) is illustrative: "In infection and contagion from body to body, as the plague and the like, the infection is received many times by the body passive; but yet is repulsed." Inoculation with the scabs of smallpox itself, although it carried a risk of virulent infection and death of perhaps 1% to 3%, had surely saved lives. In India the practice was known before the Christian era; by the eleventh century the method was practiced in China, and by the thirteenth century in the Middle East and Africa. Inoculation reached Europe much later (Hopkins 2002). In eighteenth century England, the court physician Sir Hans Sloane and others promoted inoculation. This beginning with the Royal Family in 1721 was at the insistence of Lady Mary Wortley Montagu. As a young woman, she had herself had an attack of smallpox that left her face deeply pockmarked. As the intrepid wife of the British ambassador in Turkey, she had learned of the practice of variolation and applied it to her own family. Contemporaneously, the Reverend Cotton Mather in Boston learned of the practice from his African Berber slave, confirmed the report from other slaves, and then promoted it to stem a rising epidemic in 1722. Available mortality statistics point, somewhat roughly, to a distinct preventive effect (Miller 1981).

In early literature, a few intimations hinted that protection against smallpox was conferred not only by survival after smallpox but by cowpox. Jenner was a country doctor, although no ordinary one. Before he left London to practice in the rural southwest, his natural history studies on the nesting habits of the cuckoo had won him election as a Fellow to the Royal Society by its luminaries. Over the next quarter century and more, he patiently conducted a study of such cases of both smallpox and cowpox as came his way. One may reasonably infer, from Jenner's report of his research in 1798, that he was guided from the start by an *a priori* if not explicitly stated hypothesis.

In present-day terms, Jenner's reported procedures of observation and case reporting clearly imply the hypothesis that persisting immunity to smallpox

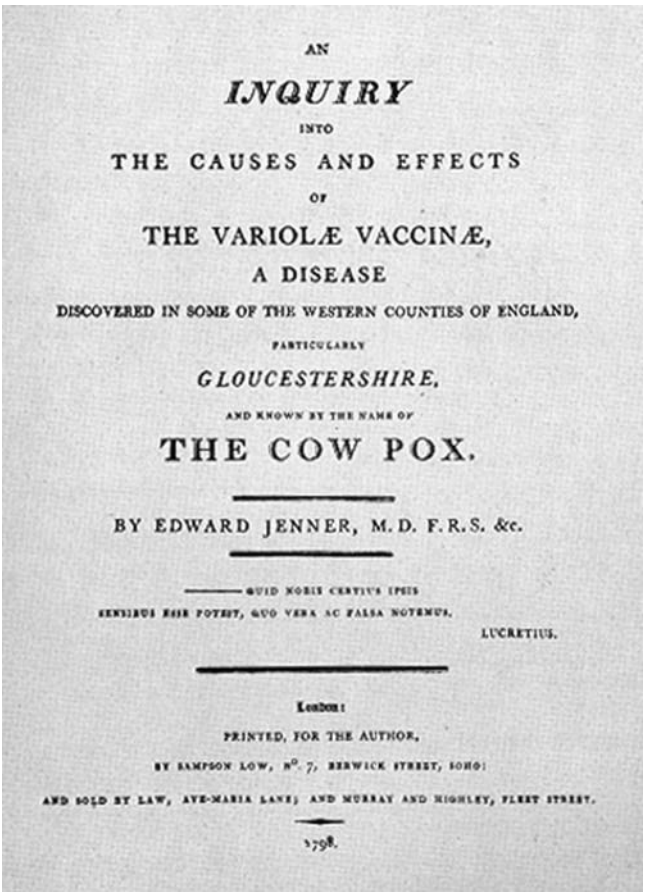
could be deliberately induced. The centuries-old practice of variolation (deliberate smallpox inoculation) to prevent subsequent inadvertent infection already gave grounds to surmise that the procedure could protect. Smallpox was a mass killer of the time and had been so over at least the previous millennium. In the common lore, cowpox was an infection seemingly related to smallpox. Jenner's insight, as others believed and indeed practiced, was that the human form of the disease could be prevented by deliberately subjecting those as yet uninfected by smallpox to variolation with cowpox.[1]

Beginning in 1773, over the next 25 years Jenner sought out every case that he could collect of either smallpox or cowpox. In every outbreak of the disease of each pox, he followed all the cases accessible to him, whether in his practice, or by report from elsewhere. He took special note of histories of exposure to both smallpox and cowpox. Thereafter, following either histories or direct observation of subsequent exposure in cases of either pox, he took note of the effects of each such infection. Jenner went beyond diligent observation. His plan with both cowpox and smallpox was first to observe, and then to compare and document reactions to exposure or to infection. These observations he then classified according to history of previous exposure, whether to one or the other type of infection or to previous inoculation with smallpox.

Once he had reached a tentative conclusion, he conducted a bold experiment (on our reading of a text that sometimes seems obscure about timing). In 1796, he inoculated a boy by the name of James Phipps with cowpox taken from the sore on the hand of Sarah Nelmes, a milkmaid recently infected. Six weeks later, he inoculated James with matter taken from the pustule of a patient affected by smallpox. Jenner inferred from the absence of any reaction to his inoculation with smallpox that the boy was now immune to smallpox. Believing that he had proof in hand, he submitted a brief paper to the Royal Society, summarily rejected because his evidence was slim, and his conclusion audacious and at variance with established knowledge.

His dissertation of 1798 (published under his own auspices since the Royal Society declined to do so) reports in some detail 23 such instances in individuals and groups of varying ages and both sexes (Jenner 1798) (Figure 11.1). Jenner's final test, carried out in several cases as opportunity arose, was to inoculate either smallpox or cowpox experimentally, always taking account of and depending on the previous history of these infections in the case recorded. After more than a quarter century of clinical and epidemiological

[1] In the eighteenth century, a great many surviving adults were characteristically pockmarked. Jenner's idea that infection with cowpox reduced the risk of further infection is said to have come to him in what seems to have been an epiphany. Allegedly, he overheard a fresh-faced milkmaid assert: "I cannot take the smallpox, because I have had the cowpox."


AN

INQUIRY

INTO

THE CAUSES AND EFFECTS

OF

THE VARIOLÆ VACCINÆ,

A DISEASE

DISCOVERED IN SOME OF THE WESTERN COUNTIES OF ENGLAND,

PARTICULARLY

GLOUCESTERSHIRE,

AND KNOWN BY THE NAME OF

THE COW POX.

BY EDWARD JENNER, M.D. F.R.S. &c.

——— QUID NOBIS CERTIUS IPSIS
SENSIBUS HIBI POTEST, QUO VERA AC FALSA NOTEMUS.
LUCRETIUS.

London:

PRINTED, FOR THE AUTHOR,

BY SAMPSON LOW, N°. 7, BERWICK STREET, SOHO:

AND SOLD BY LAW, AVE-MARIA LANE; AND MURRAY AND HIGHLEY, FLEET STREET.

1798.

Figure 11.1 Cover of Jenner's An Inquiry into the Causes and Effects of the Variolae Vaccinae from 1798, as published in 1801.

study of the collected cases, he began his crucial test of the possible immunity conferred by previous infection of one or the other type of pox. Some of the more notable results follow.

The first five cases reported had been inoculated with variolus (obtained from the scabs of smallpox patients). Three of the five had previous histories of cowpox; the other two had survived smallpox. None of the five produced either the expected and characteristic inflammatory or systemic reaction from an initial infection. Two further quite singular cases were each from a large family group. In both groups, each was the single person who had escaped infection during a previous outbreak, one of which was smallpox and the other cowpox. Subsequently both family groups had again faced an outbreak, this time of cowpox. In each family, a single person was infected

by smallpox, each of whom had alone escaped infection during a previous experience of an outbreak.

Jenner reported a number of other cases that either conform to the results described above or simply add information. For instance, having inoculated one five-year-old child with cowpox, he transferred the inoculum successively to five more children and evoked the appropriate reaction in all. Children, one should note, were preeminently the victims of smallpox, perhaps because a majority who reached adulthood were immune survivors of one of the two types of pox.

Jenner was well aware of the great significance of his undertaking. Elsewhere, he expressed rare confidence in his conclusions and the future significance of his studies. Indeed, he is credited with a remarkable prediction that vaccination would eradicate smallpox.[2]

There remain three footnotes to add to Jenner's story. One is to evaluate his own personal role, which clearly did not reside in the rigour of his field work or of his analyses. He was the first to use a human infected by cowpox, rather than an infected cow, to provide the inoculation for his cases; by doing so, he certainly modified and first created an effective vaccine, the role Hopkins (2002) assigns to him. Certainly, once the cowpox vaccine had been introduced, dramatic declines in death from smallpox followed. We applaud him, less for the originality than for the persistence with which he followed the possible affinity between cowpox and smallpox, but most of all for conceiving the idea that, as with James Phipps, one might assess the causal role in protection of cowpox infection by inferring the immunity that follows the response to variola inoculation as an appropriate surrogate.

Jenner, in 1798, coined the term *variolae vacciniae* to refer to the origin of the inoculum from *vache*, a cow. Another ironic tale, still perhaps to be further clarified one day, is that it seems likely that the lymph Jenner distributed as his stock was in fact not cowpox, but an attenuated lymph of smallpox. This stock derived, apparently, from an unintentionally contaminated source Jenner had obtained from his colleague, Dr. William Woodville, who had charge of the Smallpox and Inoculation Hospital in London. Dr. Woodville first gave a patient in his hospital inoculum from an infected cow and then, a few days later, inoculum from a smallpox-infected patient. Certainly, it is well established that the vaccine now used contains vaccinea, a virus

[2] In a search for a printed source of these words of Jenner, none of the three versions traced are identical, and none gives a reference. Hopkins (2002) gives two versions (both indexed); the third, however, is in a reprint of Jenner's privately printed dissertation. His vision was in fact fulfilled 200 years later. In 1980, the Centers for Disease Control in the United States announced that the work of eradication led by D.A. Henderson was done.

distinguishable from those of smallpox and cowpox (Razzell 1977; Baxby 1981). In fact, Derrick Baxby suggests it derives from horsepox!

A last comment concerns the later history of smallpox, part of which Jenner himself would have appreciated. In 1966, the World Health Organization (WHO) adopted a resolution calling for the worldwide eradication of smallpox. This was feasible, given both the effectiveness of the vaccine, and the knowledge that man is the only known host. A 10-year plan was adopted and virtually completed within the specified period. The last case in England occurred in a laboratory-related incident in 1973; the last case in North America was carried by an affected person from Brazil to Toronto in 1962. In 1971, after 22 smallpox-free years, vaccination was discontinued in the United States. The question of reintroducing vaccination has arisen, however, in a bizarre form Jenner could never have anticipated. The point at issue is whether to shield the population by vaccination from the threat of terrorists using smallpox virus as a weapon.

Beginnings of the Science of Immunity: Pasteur

Louis Pasteur, in his second major phase of thought and research, carried forward, consolidated, and generalized the themes of immunity and host resistance to specific organisms (Vallery-Radot 1960; Debré 1998; Dubos and Brock 1998). His work was the necessary prelude to the general application of immunization for many diseases in domesticated animals and in human beings. Pasteur carried forward the Hippocratic idea of the host beyond conceptual recognition, important as that had been, to practical realization in inducing immunity. As we have learned in the longer run, the spread and distribution of infectious disease could not be fully comprehended without also recognizing the determining role of the immune state of the individual host and of the host population at large. Account must be taken of those roles in all their complexity at macro- as well as micro-levels. Infection requires specific organisms, each with its own biology, viable environment, favored hosts, and modes of transmission. Without the interaction of available and susceptible hosts and a viable environment for the infecting organism, there can be no infection.

Drawing on Jenner, Pasteur began his active campaign by successfully immunizing hens against chicken cholera. In the face of this quickly fatal epidemic infection, hens keeled over like bowling pins. To make his vaccine, Pasteur first attenuated the cholera organisms. This step followed a chance observation, one of the many in the sciences that have proved fruitful for "the prepared mind," as Pasteur's great contemporary Claude Bernard sagely observed. Pasteur had noted to his surprise that old vials of chicken cholera

were no longer fatal to chickens. Searching for the reason, he found that exposure to oxygen attenuated the cholera organisms. Moreover, after injection with the attenuated organisms, the chickens proved immune to further exposures of virulent organisms.

Certainly, Pasteur will have been well aware of the risk-laden use of inoculation by smallpox itself to produce immunity. When he addressed the problem, he mentioned Jenner, but one can only infer that probably he had in mind Jenner's use of the milder form of pox that afflicted cows to protect against the more virulent smallpox. Having succeeded with chicken cholera, he turned his energies to anthrax (in the literature of the time also described as "splenic fever"). This disease was destroying 25% to 30% of sheep and cattle across much of Europe. To attenuate this organism was no simple matter.

In its spore form anthrax retained virulence for years and could not be cultured. However, the organism also exists in an active form of rods or filaments. Pasteur found a way around the difficulty with spores by culturing the rods or filaments of the organism before they could form spores. Again a problem arose, in that injecting rod forms into animals apparently killed them, although in the dead carcasses no organisms could be found. Once more Pasteur undertook the attenuation of virulence by exposure of the rods to oxygen at the 42/43 degree centigrade temperature range. In 1880 he announced that he had produced a vaccine.

Although triumphant, he was still not without the harassment of critics. Hippolyte Rossignol, editor of the Veterinary Press, cuttingly greeted the result of "the Pontiff, the learned Mounsier Pasteur" as "microbiolatry." Thus challenged by skeptics, Pasteur devised a controlled experiment. He vaccinated a number of sheep, left the controls unvaccinated, and put them out to graze in a small anthrax-infected enclosure. Many skeptics and local luminaries came to see the outcome, which demonstrated complete efficacy of the vaccine.

In 1885, Pasteur essayed the first known instance of the application of any vaccine to a human host already infected. This was the celebrated case of Joseph Meister.[3] This nine-year-old child, having been bitten by a rabid dog, faced the desperate threat of rabies. It was common knowledge that the outcome of rabies was virtually always fatal. Pasteur responded to the appeal for help from the distressed father. Pasteur had already made a rabies vaccine by attenuating the organisms but had not yet tested it in a human being. Not without doubts, he administered the vaccine. The feared hydrophobia

[3] Loyal to the end, Meister became a concierge at the Pasteur Institute (founded in 1888). In 1940 during the occupation of Paris, he refused to open Pasteur's crypt for the German invaders. In despondency, he committed suicide (Debré 1998).

Figure 11.2 The Pasteur Institute, Paris, France. Credit: Courtesy of the National Library of Medicine.

of rabies did not ensue. Although this was only a single case in an uncontrolled experiment, he had sustained his theory that attenuated organisms could induce immunity. Everyone now knows that the remedy has proved itself many times as a defense against rabies (Figure 11.2).

Aside from the case of smallpox, as previously noted a few other instances of immune states induced by antecedent infection had been inferred and reported. One of these was Panum's observation 60 years before of lifelong immunity conferred on individuals after an attack of measles (Panum 1847). Since Jenner, however, little progress had been made in formulating any general theory of immunity. One of Pasteur's most significant contributions was to produce such a theoretical base, one that could apply to the generality of infectious diseases. Moreover, in the exceptional case of Joseph Meister, he extended theory into therapeutic practice. The case suggested that immunity could be mobilized not only to protect against future attacks of disease, but could take on a virtually therapeutic function in blocking infection that had already invaded the body.[4]

[4] During the latter half of the twentieth century, we have come to understand, step by step, that the immune principle is not limited to infection; it applies in reverse to autoimmune diseases and even to hormonal aberrations in obesity and diabetes and other conditions not necessarily expressed as disease.

Toward the turn of the nineteenth century, other researchers furthered the study of immunity and preventive immunization. Notable among them were Emil von Behring and Paul Ehrlich in the case of diphtheria, and early in the twentieth-century Almroth Wright and Richard Pfeiffer in the case of typhoid. An early conceptual development was the toxin-antitoxin against diphtheria by von Behring and Ehrlich (winners respectively of the Nobel Prize for medicine in 1901 and 1908). The general idea of this treatment (Behring's law) was that the immune bodies provoked by a specific infection in one host could defend against the same infection when transferred to another host.

Successes against many infectious diseases followed: in many countries now, every child is the beneficiary. These developments stemming from the idea of immune response led to fresh concepts now applied, not only in developing vaccines, but also in such diverse fields as autoimmune diseases, homografting to replace malfunctioning organs, and a theory of stepwise carcinogenesis. The concept extends to the greatly refined level of genes interacting with environment: in both host and agent, gene mutations and polymorphisms bring about variations in immunity.

References

Baxby D (1981). *Jenner's Smallpox Vaccine: The Riddle of Vaccinia Virus and its Origin*. London: Heinemann Educational Books.

Debré P (1998). *Louis Pasteur*. Baltimore: Johns Hopkins University Press.

Dubos RJ, Brock TD (1998). *Pasteur and Modern Science*. Washington, DC: ASM Press.

Hopkins DR (2002). *The Greatest Killer: Smallpox in History, with a New Introduction*. Chicago: University of Chicago Press.

Jenner E (1798). *An Inquiry into the Causes and Effects of the Variolae Vaccinae, a Disease Discovered in Some of the Western Counties of England, Particularly Gloucestershire, and Known by the Name of the Cow Pox*. London: S. Low.

Miller G (1981). Putting Lady Mary in her place: a discussion of historical causation. *Bull Hist Med* 55:2–16.

Panum P (1847). Observations made during the epidemic of measles on the Faroe Islands in the year 1846. *Bibliothek for Laeger* 1:270–344. See Panum P (1989) *Med Classics* 3:829–886 for recent reprint.

Razzell PE (1977). *Edward Jenner's Cowpox Vaccine: the History of a Medical Myth*. Firle, Sussex, UK: Caliban Books.

Vallery-Radot R (1960). *The Life of Pasteur*. New York: Dover Publications.

12

Epidemiology Fully Harnessed to Public Health: New York

In the Western world in public health practice through the latter decades of the nineteenth century, ideological struggle with an old guard had brought little change. The sanitary movement and not the germ theory held sway and determined the nature of the policies of the time. In the last decade of the century, however, we have seen how a new era emerged. Since the Renaissance, Western Europe had been the locus of virtually all the preceding fundamental research, and had remained the locus for the accrual of the new understanding of infection and its role in disease. In the late nineteenth century, however, the scene for public health innovation shifted to New York City, then still the New World to many in Europe. There, innovative public health policy first fully implemented, in public health practice, the new knowledge of epidemiology and bacteriology.

During the nineteenth century in the postrevolutionary United States immigrants, many of them the poor of Europe, rapidly swelled the populations of New York and other American cities in the East. These cities endured an evolutionary growth cycle akin to, if in time a step behind, industrializing and urbanizing Britain. Through most of the century, the Eastern cities of the United States likewise took tentative and impermanent steps toward meeting the growing threat of epidemic infectious disease. Imported epidemics carried by trade and immigrant ships—of smallpox in the first instance, of

yellow fever, and then of cholera in 1832 and from 1849 through 1854—had a ready portal of entry through the busy East Coast ports. The opening of the Erie Canal brought infection from Canada to New York. Yellow fever traveled to the Gulf and Atlantic coasts, and spread along the river valleys, causing panic, disruption, and loss of life.

Despite the alarm elicited by each epidemic, action was at first desultory. As elsewhere, whether in medieval Italian cities beset from the East by plague or as in other European countries, city fathers at first resisted the inconvenient steps that might enhance protection because of a perceived threat to commerce. Typically, in the end the necessity prevailed for taking some kind of action by such means as were available. Whether they were effective or only thought to be so, at some point action became obligatory (Rosen 1958; 1993).

In New York, Boston, and other American cities too, city councils responded to epidemic onslaughts—of smallpox and cholera in the first instance—by setting up health boards (Rosner 1995). New York also appointed a port quarantine officer. As occasion arose, this officer would set up procedures to put quarantine into effect (Duffy 1990). At the outset such measures were often merely temporary expedients.

These plans and responses did not necessarily include physicians, whose repute was low through much of the nineteenth century. Through much of that century, indeed, a striking divide on health boards put lay representatives and the medical profession at odds. From common experience in their own lives, many of the lay public were of the conviction that epidemic diseases spread by contagion from person to person. In epidemics small or large, however, the orthodox teachings of the medical profession were quite contrary. In hindsight, the modern eye—unless schooled in the power of dominant paradigms—is bound to count it remarkable that throughout the nineteenth century many doctors remained committed to seventeenth century Hippocratic–inspired teachings on miasma theory. Not unreasonably, given the then current doctrine and the absence of any effective means of treatment, they gave primacy to the dispersal of miasmas as preventive sanitary measures.

Despite the faulty theory, these measures could often lead to sensible action. Doctors were quite properly convinced of the necessity for clearing streets and swamps of sewage and refuse, and of the virtues of living at higher altitudes. In New York City, these measures had the genuine but unanticipated benefit of reducing the risk of disease and death from two major killing diseases. Both yellow fever and malaria were prevalent mosquito-borne diseases. Malaria was a chronic warm weather hazard. Fatal yellow fever epidemics swept across the United States throughout the eighteenth century, and sporadic, severe epidemics persisted in cities on the Gulf Coast of Mexico through the nineteenth century. In 1878 a yellow fever outbreak

travelled up the Mississippi to cause havoc in Memphis. In New York, however, the last cases were recorded in 1822. Virtually no further yellow fever outbreaks occurred in the United States after 1905 (Patterson 1992). With the completion of the Panama Canal in that year, however, cholera traveled by way of the canals and rivers to Chicago, Wisconsin, St. Louis, and Pittsburgh, but again New York escaped.

In 1890, a new era of epidemiology and public health began that seems to have emulated—or, if not, to have unknowingly replicated—aspects of the model for public health practice Simon in England had developed in his later years at the Privy Council. The Board of Health in New York City, at the urging of its elite physicians, appointed as the director of a new Division of Bacteriology and Pathology Hermann Biggs, a graduate of Cornell University and of Bellevue Hospital Medical College in the city. Today the division still serves New York under the name of the Public Health Laboratory of the New York City Department of Health and Mental Hygiene.

Biggs, during a postgraduate experience in Germany late in the nineteenth century, had come upon the new ideas of the bacteriologists and perforce engaged with the new knowledge. The flurry of discovery and rapid fire productivity of Koch, Klebs, and the several other talented bacteriologists in Europe was at its height. Following Pasteur, they were elucidating the significant role of bacteria in disease and also exciting others to effort in the new direction. Biggs aimed to create an effective public health program that put to use what he had learned. It was not surprising then that Biggs, together with other leading physicians who shared his aims, saw a laboratory as central to his plans for promoting health in the city.

However, in the year 1892, public health officials of the City of New York faced crises such as confronted few other cities at the time. By the end of the nineteenth century, New York was not only one of the most populous cities in the world. It was also uniquely polyglot, with a crowded mass of newly arrived, poverty-stricken families living in unsanitary, primitive, and over-crowded tenements. At the healthier locations in the lower reaches of Manhattan, the more settled Americans, enjoying the Gilded Age, lived safe from malaria. The Health Department was heavily dependent on the political favors of a powerful few (Markel 1997). There was a Department of Sanitation, led by Dr. Cyrus Edson, a well-trained physician (and politically very well connected), and a Quarantine department. The Quarantine department was at first responsible to New York, but during the crises described below, the Federal government sought a much stronger role.

The first crisis was the appearance in the harbor in January 1892 of the SS *Massilia* (Markel 1997). This was a slow steamboat, carrying Jewish

refugees fleeing from their prior homes in Russia. Some had walked thousands of miles to reach the port of Odessa; then on being ordered to leave, found sponsorship among a well-off Jewish community to sail to the United States. However, they first disembarked at Constantinople, where legal difficulties and argument detained them. There they lived in unsanitary circumstances for three months. Finally permitted to sail, the ship stopped to collect Italian passengers, and after 28 days, arrived in New York. Allowed to disembark after a cursory physical examination, a Hebrew Reception society settled them in the crowded Lower East Side of the city. Three days later, some were discovered to have typhus. Edson immediately went into action, checked all the addresses in which the newly arrived residents had been placed, and moved all who were diagnosed with typhus, as well as many of their contacts and others from the tenements, into the isolation hospital on the East River. Many died there, the publicity was enormous, and although the epidemic was controlled within a very few weeks, the political fallout against the immigration of Russian Jews stimulated the development in the senate of a bill (which failed to pass) selectively to refuse them entry.

A new crisis arose only six months later. This stemmed from the cholera epidemic raging in Hamburg during the years 1891–1892. Hamburg was the site from which Russian Jews took ship to the United States. While awaiting passage, they were housed in a slum neighborhood with poor sanitation, and a probably polluted water supply from the Elbe. Robert Koch was called in to advise the authorities. Severely critical of what he found, he advised forcefully of the need for boiled drinking water, thoroughly cooked food, and improved housing.[1] Still, the Hamburg authorities were slow to move. Nonetheless, they were anxious that the temporary residents, waiting to take ship to the United States, should leave their shores as planned. They advised the United States representative in Germany that Hamburg was clear of cholera. Within the next three weeks, three ships had left Hamburg for New York. Each was loaded with infected passengers some of whom died at sea, while putting many fellow passengers at mortal risk. The poor passengers in steerage were hardest hit, although a few among the better-off passengers travelling in the cabin and luxury classes were also infected.

The first ship to sail, the S.S. *Moravia*, was almost completely occupied by Russian Jews, all in steerage; on arrival in New York, the ship was refused entry, and all passengers were quarantined. The next ship, the S.S. *Normannia*, also carried infected passengers, but in this ship some passengers were well-known and well-off. The quarantine regulations applied to

[1] See Morabia 2007 and Oppenheimer and E Susser 2007 for a fuller account of cholera in Hamburg, and how the authorities responded.

all, although enforced in a blatantly discriminatory fashion. The steerage passengers were all taken off and lodged in so-called lazaretto hospitals already overcrowded. The better-off were to be housed in a hotel on Fire Island (Markel 1997). Despite strong protests from the Fire Island inhabitants, they were eventually settled there. Here was an explicit example of the Health Department treating immigrants according to their social status.

Although this crisis too subsided as the disease was contained and kept from the mainland, the episode itself left its mark. Not only was there conflict between the New York and Federal authorities: there was also this new transgression of human rights on the part of the Health Department. For Biggs, the first and immediate concern was to stem the tide of diphtheria. Diphtheritic croup was at the time the most terrifying and perhaps the deadliest disease of childhood in the city (Hammonds 1999). In the first instance, the Board of Health rejected Bigg's requests for the necessary funds. Exceedingly cost-conscious, budgetary concerns rendered the Board impervious to his arguments. In the wake of those negotiations, however, in 1892 an epidemic once more compelled a response from the reluctant Board. Ironically, the Board was impelled, not by the constant and endemic child deaths from diphtheria, but by fear of a new explosive and acute epidemic, such as cholera that might threaten adults.

Nonetheless, the child deaths endemic among all classes built up steady pressure on the Health Board. Biggs made a strong case for his need for a laboratory in the Department, deploying the facts that in 1891, 4874 cases of diphtheria had been notified, of whom 1361 had died. Biggs emphasized the crucial need for laboratory findings with diphtheria, and most clinicians agreed that clinical findings were insufficient for accurate diagnosis. Diphtheria was a disease with which physicians were tragically familiar: the distinguished Drs. Abraham and Mary Putnam Jacobi lost a son to the disease in 1893. William Park was appointed to head the new laboratory, and in the event the fine work accomplished by Park and his team not only made a major contribution to the understanding of the disease, but also raised the status of New York City's Department of Health nationally and internationally (Fee and Hammonds 1995).

Park's first action was to create a procedure for swabbing the throat of the suspected patient; the laboratory would then search for the Klebs-Löeffler organisms known to be associated with diphtheria. He also carefully explored the meaning, for public health, of the carrier state of well individuals from whose throats the bacillus could be cultured. He sought to determine whether the role of carriers differed in the virulence of the bacilli they "carried"; or whether the carrier organism was not the Klebs-Löffler organism but another, nonpathogenic but similar in appearance. Later, in

1894, von Behring in Germany and Roux in Paris developed antitoxin to diphtheria, and Biggs secured funds for Park to set up a laboratory to produce this life-saving treatment. In spite of a somewhat rocky start, the antitoxin was soon approved for regular use.

During his studies abroad, Biggs seems very likely to have been apprised of Simon's program at the Privy Council. Simon's work was well known in Germany to such public health authorities as Pettenkofer and others (Lambert 1963). Certainly the model Biggs created of public health practice enhanced by field studies, the laboratory, and research was similar in its central features to that of Simon. His actual program was more akin to Simon's during his first public health appointment to the City of London, however, than to his national program when elevated to the Privy Council. Biggs was intensely engaged in his immediate locale and closely watchful of emergent problems (Fee and Hammonds 1995). However, neither Biggs, nor W.H. Park and Alfred Beebe—the young colleagues he had recruited for the laboratory—subscribed any longer to the supposed and still prevalent verities of Sanitary Era miasma.

The New York City approach to protecting public health was soon adapted to exploit the potential of the bacteriology laboratory. The primary and critical innovation in New York was a system of constant field surveillance for epidemic outbreaks and notification of cases. Suspect cases were confirmed by laboratory diagnosis, and if positive, isolated in quarantine thereafter (Fee and Hammonds 1995). The system, thus kept on the alert by monitoring and inspecting potential sources of infection, enabled quick response to early signals of potential outbreaks and epidemic spread. Quite soon, the division had demonstrated its capabilities for surveillance and control among the large immigrant population of the city.

As noted, the threat of cholera provided the ostensible reason for the funding of the New York Public Health laboratory. Paradoxically, it was not cholera that was the grounds for either the great repute for innovation the division and its laboratory quickly attained, or for the assured funding from the city fathers consistently obtained thereafter. To begin with, in the new division routine hygiene measures were introduced to deal with milk diarrhea and smallpox vaccination. The springboard for all these programs was in fact its work with diphtheria. Deploying a number of scientific as well as practical advances, Park and colleagues showed that curbing transmission of infection by silent carriers was a critical key to preventing the many fatalities from diphtheria epidemics among the children of the city. The New York program joined the earlier British program developed by Simon in serving as models for the new public health in Europe as well as the United States.

References

Duffy J (1990). *The Sanitarians: A History of American Public Health*. Urbana: University of Illinois Press.

Fee E, Hammonds EM (1995). Science, politics, and the art of Persuasion: promoting the new scientific medicine in New York city. In Rosner, ed. *Hives of Sickness: Public Health and Epidemics in New York City*. New Brunswick, NJ: Rutgers University Press, for the Museum of the City of New York, pp: 155–196.

Hammonds EM (1999). *Childhood's Deadly Scourge: The Campaign to Control Diphtheria in New York City, 1880–1930*. Baltimore: Johns Hopkins University Press.

Lambert R (1963). *Sir John Simon, 1816–1904 and English Social Administration*. London: MacGibbon & Kee.

Markel H (1997). *Quarantine! East European Jewish Immigrants and the New York City Epidemics of 1892*. Baltimore: Johns Hopkins University Press.

Morabia A (2007). Epidemiologic interactions, complexity, and the lonesome death of Max Von Pettenkofer. *Am J Epidemiol* 166:1233–1238.

Oppenheimer GM, Susser E (2007). Invited commentary: The context and challenge of Von Pettenkofer's contributions to epidemiology. *Am J Epidemiol*. 166:1239–1241.

Patterson KD (1992). Yellow fever epidemics and mortality in the United States, 1693–1905. *Soc Sci Med* 34:855–865.

Rosen G (1958; 1993). *A History of Public Health*. Baltimore: Johns Hopkins University Press.

Rosner D (1995). Introduction: "Hives of Sickness and Vice" In Rosner, ed. *Hives of Sickness: Public Health and Epidemics in New York City*. New Brunswick, NJ: Rutgers University Press for the Museum of the City of New York, pp: 1–22.

13

Evolution and Genetics: Darwin and Galton

In the nineteenth century a number of new ideas about host factors converged. Charles Darwin (1809–1882), in *The Origin of Species* in 1859, among much contention about his then sacrilegious ideas, revolutionized prevailing biological thought.[1] His theory left no room for biblical assumptions of a living world of static forms created at a stroke by a divine being.[2] His biological vision encompassed a reproductive process among the living individuals of a species, each newly and uniquely formed by sexual reproduction and always as members of populations interacting with others.

At the very center of his theory of evolution, Darwin placed the natural selection of traits, developed according to their survival value for a species. He proposed his theory as either primarily or solely a theory of evolution made possible by individual genetic change from generation to generation. Darwin postulated that species evolved over the course of time, a process of natural selection and survival, which in given conditions produced gradual

[1] Noteworthy books on Darwin's life and scientific contributions include: Gruber et al. 1974; Bowler 1990; Mayr 1991; Desmond and Moore 1992; White and Gribbin 1996.

[2] Darwin delayed publication for almost 20 years, fearful in those devout times of charges that he had flouted and nullified the hand of God; ultimately, he was prompted to publish by the imminent appearance of the similar theory independently arrived at by Alfred Russel Wallace.

change. In the face of the necessities imposed by environmental conditions, selection of populations would follow according to individual characteristics more or less favorable to survival.

Francis Galton (1822–1911), an admiring younger cousin, was greatly impressed by Darwin's ideas.[3] Although much taken by Darwin's theory, Galton's own perspective led him to shift away from the idea of natural selection and toward a purely hereditarian theory based on breeding alone (Galton 1888, 1894). In their correspondence, Darwin shows himself unhappy with such reductive hereditarian notions and also with the criticism of his work, especially because before publishing, Galton had neglected to discuss his differences with him (Gillham 2001).

Much later, in correspondence with Galton in 1875, Darwin makes the nature of his discomfort quite clear. His idea of pangenesis assumed that individual traits were transmitted sexually from both parents by *gemmules*. These miniscule particulate units, he suggested, accumulated in the sexual organs, replicated in cell division, and were dispersed throughout the structures of the body. To account for the obscure issue of reversion—the disappearance of a trait and its reappearance in later generations—Darwin proposed that some of the gemmules could lie dormant. Gemmules were not immutable; change might come about by chance, or from injury to reproductive organs either very early in development (a process that now we might call epigenetic) or by gradual change over generations (genetic mutation).

Galton accepted the hypothesis that inherited particles determined traits, but was disturbed by the implication that traits could be acquired (even in the particular sense that Darwin proposed). He set about testing and quantifying his own hypothesis in an experiment. His crossbred rabbits produced 24 litters and 124 progeny. Transmitted germs (Darwin's gemmules renamed), he concluded, were immutable. Moreover, germs could neither be affected nor transmitted either by blood transfusion or by alterations in somatic tissues. The four postulates he derived were only partially in accord with Darwin.[4]

[3] Galton had begun medical studies at the precocious age of 17. Being mathematically gifted, he had gone on to pursue the rigorous Tripos in mathematics at Trinity College, Cambridge. In the end, feeling himself ill and apparently beset by anxiety, he failed to complete these exams.

[4] To translate each in terms of modern knowledge, as Galton's recent biographer does (Gillham 2001), might not be too fanciful. Each postulate is followed by a restatement in parentheses; that is in alignment with current knowledge. 1. Each one of an enormous number of quasi-independent units of the body has a separate origin or germ ("each gene specifies a different protein"); 2. Germs greatly outnumber body structures; most are sterile, and only a few develop into body structures ("proteins encoded by genes do not contain hereditary particles, but with a possible hint that some are recessive"); 3. Germs in a latent state are transmitted to offspring but may not be expressed; 4. End results are expressed during development and depend, in Galton's words, "on the mutual affinities and repulsions of separate germs" (genes are expressed during development). He reaches the insight, also, that offspring can have only half as many as the germs from two parents combined.

Galton finally arrived at a theory compatible in essence with polygenic inheritance; traits are determined by a number of genes. That is to say, traits do not vary only binomially but also quantitatively and continuously as do height and many other characteristics. Galton had been sufficiently stimulated by Darwin to continue developing his own diverging theory. He set about the study of those families in England that he could identify as having exhibited distinction intellectually or in achievement. From these data, he put forward an ethnocentric and self-regarding theory. The theory explained familial resemblances in personal characteristics by the mechanisms of individual hereditary transmission. Inheritance explained the evident superiority of the fortunate and well-endowed classes and races and, likewise, the innate inferiority of those less-endowed and less fortunate (Galton 1894).

Galton's revision of Darwinian theory was the immediate stimulus for his studies of superior families, and the ultimate stimulus for the founding of the disciplines of biometrics and eugenics. To this end, Galton endowed the Chair in Eugenics at University College, London, first filled in 1911 by the mathematician Karl Pearson, his follower and nominee (of whom, more later). Galton developed firm convictions (shared by Pearson) about Anglo-Saxon racial superiority, the inferiority of black African peoples, people of Jewish descent, and even of the South African Boers whom he encountered as a young man on a trek in Southern Africa, (and who, as Galton did not realize, were after all blended Anglo-Saxons of Dutch, German, and French stock).

Biometrics can be defined as the study of the distribution, variation, and inheritance of natural characteristics. A key technique invented by Galton for such studies was the statistical technique of correlation (Galton 1888). Galton's chief intellectual heirs in the study of inheritance were three. Karl Pearson and W.F.R. Weldon both followed up his idea of gradual continuous variation embedded in the biometric concept. The thinking of William Bateson (1861–1926),[5] Galton's third protégé, took a different direction. He broke from the others because, in accounting for gradual evolutionary shaping, he saw the inheritance of definitive phenotypic characteristics—say,

Thus "… chance deficiency in the combination of either of them, and of any particular species of germ, will be supplied by the other." (Note, that contributions may be unequal: thus in Hereditary Genius, he attributed 70% of the genius to the father and only 30% to the mother). Therefore, Galton thought, bisexuality was inevitable. It is a reasonable claim that this observation predicts knowledge of meiotic division.

[5] Bateson was the first to use the term "genetics" in 1905. His in-depth understanding and promotion of Mendel's work led to important advances in the detection of inheritance patterns in human diseases (Harper 2005). Bateson also published his own translation of Mendel's papers on hybridization.

of carnations white or red—as a central problem. It was at odds with the continuous variation built into biometric theory.

By contrast, Darwinian theory, modified and as subsequently understood, requires a continuing history within and across populations of the interactive process between genes and environment. Given enough time, during which, in each specific environment, species adapt or survive selectively by the chance inheritance of favorable traits, the interaction produces a slow but steady evolution of life in all its forms. The process unfolds further in a continuously changing environment, which is in part the result of the activity of the evolving organisms themselves.

The Host and Genetics

Gregor Mendel (1822–1884) was an Augustinian monk and abbot. His epochal paper of 1866 described simply and clearly the genetic transmission of particular characteristics from one generation to the next. Mendel followed a line of thought entirely independent of Darwin; almost certainly Mendel was ignorant of Darwin's writings (though Darwin's *Origin of the Species* was first published in 1859). Mendel's paper languished unknown either to Darwin, or to Galton, or to many others interested in the mysteries of heredity.[6] Everyone now knows that in the seclusion of his monastery Mendel's elegant experiments involved the growing of sweet peas of different strains and colors. He observed and named dominant and recessive traits. He concluded that the individual variation among the successive generations he propagated could be explained by the ordered transmission of the specific traits of each variety. Such interest as his work had evoked lapsed for the next four decades.

Hugo deVries, Carl Correns, and Erich von Tschermak-Seysenegg are named today as the rediscoverers of Mendel's historic 1866 paper, though it was William Bateson who did the most to promote Mendel's work and place it in a meaningful context (Bateson 1909; Harper 2005). The journal in which Mendel's paper appeared was apparently quite widely read in Europe. Nor was its long obscurity owed to monastic isolation. The silence among scientists seems most likely to reflect an interpretation ahead of its time. It was a unique result in an experimental science that did not yet exist.[7]

[6] It is extraordinary to read the Darwin/Galton correspondence and papers in the period 1869–75 (Gillham 2001) postulating "patent" and "latent" heritable characteristics and reversion for both of which the 1866 paper of Mendel is uniquely explanatory.

[7] One should note further that the paper contained some anomalies of reporting. R.A. Fisher (1890–1963), a leading statistician and theoretical geneticist of the twentieth century, found the

Indeed, as the twentieth century opened, genetically-oriented biologists saw contradictions between the theories of Galton and Mendel. Even Galton's belief in the overriding power of his postulated hereditary germs did not readily resolve the contradiction. The revelation of clear-cut genetic transmission concentrated sharp attention on the nature of the host. Once the significance of Mendel's work began to be appreciated, spectacular discoveries of genetic transmission soon followed. A number of striking physical and metabolic abnormalities were reported by clinicians, beginning in Britain with Garrod's description in 1902 of alkaptonuria.[8] Such observations provided human examples of patterns of transmission that were clearly Mendelian, fortified Mendelian theory, and ultimately could be traced to their physical and biochemical source. The particulate carriers of inherited characteristics had acquired the name of *genes* in 1909.[9] For a given gene, variations in its expression may occur, these arising from slight variations in the alleles of which it is comprised.

Galton's protégé Karl Pearson was a trained mathematician. In parallel with these developments in genetics, he invented sophisticated techniques (primarily the correlation coefficient to express degrees of resemblance) to carry forward Galton's work on the statistical study of biological phenomena especially among human beings. Pearson gave shape to biometrics. He also adopted Galton's eugenic viewpoint holus-bolus. In 1911, he was appointed to the chair of eugenics endowed by Galton at University College in London and founded the first important school of statistics in Britain, namely, Biometrics. Pearson is a significant historical figure for epidemiology as well as for statistics, although one must allow that he could be no less dogmatic than was his sponsor Galton (Susser 1977).

R.A. Fisher (1890–1962) a British mathematician, statistician, and biologist, made at least two notable advances to the understanding of biometrics and evolution. At the turn of the nineteenth century, scientific dispute raged around the divergence of the two theories of inheritance, namely, the continuous variation of Galton's germs and the separate transmission of the singular entities of Mendel's specific characteristics. In 1915, Fisher finally

---

results too good to be true. To his dismay, his calculations of the standard deviations and coefficients in the experimental results showed no variation at all. He was unbelieving that the reported results could fully reflect the actual data. Instead, Fisher speculated that Mendel's earlier work had convinced him of the phenomena of the rules governing these phenomena (Box 1978).

[8] Alkaptonuria is an autosomal recessive amino-acidopathy in which lack of an enzyme leads to an accumulation of homogentisic acid in urine and causes it to darken on standing.

[9] The terms gene and genotype are attributed by the Oxford English Dictionary to W. Johannsen, Elem.d.exacten Erblichkeitslehre (1909). Genes, originally the ultimate units of mutation and recombination, came to be seen as sequences of nucleotides within nucleic acid molecules each determining the primary structure of a protein or polypeptide molecule.

resolved this problem. Using regression equations, he mathematically synthesized the process of polygenic and single gene transmission, showing that these were not mutually exclusive.

Fisher made a second and perhaps even more important contribution to the concept of evolution. In the post-Mendelian climate of discourse, he weighed the role given by Darwin to the environment and evolved for himself what he termed a *nondeterminant* approach to evolution. By this he meant that the characteristics of species were not immutable but could adapt to changes in the environment. For instance, in populations of grouse lice and of butterflies, he estimated the effects on fertility of both observed and manipulated changes in the environment. More or less as he predicted, changes to individuals occurring at random and favoring survival became the dominant pattern in the surviving population. He describes such changes as "creative" (Box 1978). The selective quotations below convey the trend of his thoughts.

"All laws of natural causation are essentially laws of probability, hence the predictability of a system has the same basis as 'natural' in the social sciences." The nondeterminant view recognizes the creative element in evolutionary change—that sense of bringing something to pass that otherwise would not. On the other hand, "determinate" would mean every cause must itself be an effect of an antecedent cause, which, in turn, was itself merely an effect. "Man can remember, but not foretell," he wrote to emphasize the difference. And again, "The selective value of chance must always be in harmony with the world around us." Chance must be able to utilize its advantages or penetrate its undiscovered possibilities. As organisms evolve, heritable changes take place as they mutate.

Next Fisher turned his mind to experiment and to testing his predictions. Polymorphic characteristics, like color in butterflies, provided a fine opportunity, since one could test color as a natural marker for change. The characteristics of a population, he theorized, are maintained in equilibrium by a balance of selective forces: for example, one could compare the frequencies observed in the field from those to be expected from random mating and so estimate the selective intensities that maintain the stable polymorphisms in the species. Thus among grouse-locusts, color pattern exhibited differential responses to environmental changes, whether natural or contrived (Box 1978). On a Sicilian island, Fisher, working with marked butterflies, observed changes in the population that were present at birth and was able to distinguish those resulting from environmental changes. Thus, he estimated which of the changes were related to microhabitation and how season affected the populations. Similarly with the peppered moth, he noted their different prospects for survival in unpolluted woodland and in urban parks.

Genetics and Evolution

Increasingly in recent decades the application of refined molecular analysis, combined with careful clinical observation has revealed unsuspected diversity even in seemingly specific genetic disorders. Such variability among disorders is attributed to either dominant or recessive single genes. The variant genes are heritable, but much of the variability is owed to mutant genes. The origin of such mutations is not always evident. Studies of the effects of radiation in such diseases as leukemia show that mutations can be induced by such an external agent. Frank MacFarlane Burnet speculated that, since millions of years ago environmental radiation was much higher than it is now, mutation could have exerted more rapid changes in evolutionary patterns than it would do today. We note later that there are organisms that divide with great rapidity and evince frequent errors in mutation. Such organisms have been found to respond rapidly to environmental hazards (see later discussion on HIV). It seems then that specific attributes of the host, acquired as well as inherited, could be determinants of their own disorders. In the twenty-first century, host factors in disease are being studied with ever growing intensity. Current familial and population models, both those of the host-oriented geneticist and the environmentally-oriented epidemiologist, have converged despite much mutual misunderstanding.

The Eugenics Movement

As will emerge from an account of the flow and counterflow of intellectual ideas well into the 1930s, intellectuals with the shared objective of improving the health, well-being and productivity of their societies, saw quite different ways of obtaining such an objective (we do not speak here of the aberrant and malignant policies of Mussolini, Hitler, and Stalin, although they too aspired to reshape their societies). The Eugenics movement began with Francis Galton, who was followed later by his designated successor Karl Pearson for whom Galton endowed a new chair at University College, London (Kevles 1995). At that time, it seems, such ideas were not irreconcilable with a progressive stance in other matters. Pearson not only invented many statistical techniques still in use. He was also a socialist who moved in intellectual circles. He counted among his friends Olive Schreiner (the most renowned South African novelist of the nineteenth century), Eleanor Marx (the daughter of Karl Marx), Sidney and Beatrice Webb, and George Bernard Shaw (all leading British socialists of the first half of the twentieth century).

Yet, on eugenic grounds, Pearson supported the Anglo-Boer War of 1899–1901, having classed the Boers as an inferior race. Moreover, from his observations of malnourished immigrant Jewish schoolchildren in the East End of London, in 1921 he "reluctantly" concluded that they were so stunted and backward that further admission of such immigrants could not be recommended. Ironically, not a few among that cohort, grown to adulthood over the subsequent decades, thoroughly repudiated Pearson's view and assumed significant positions in British society in politics, the universities, literature, and the sciences.

Fisher, too, was a eugenicist. Indeed, in his undergraduate years at Cambridge, he founded the University Eugenics Society. Fisher emphasized the advantages for society of more offspring emanating from more capable parents. However, his view was distinct from the crudities of Nazi ideology in the 1930s, in that he neither proposed attempts at imposing such policies, nor simplified the issues to a philosophy of race or politics. The environment should be so modified, he argued, that "the motivation of independent and self-reliant citizens to have children should be increased." On these grounds, he deplored contraception for educated people. The dedicated socialist George Bernard Shaw, in his late play "Man and Superman," countered and sided with Havelock Ellis in arguing that poverty prevented men and women from choosing genetically optimal partners. "To cut humanity up into small cliques, and effectively limit the selection of each individual to his own clique, is to postpone the Superman for eons, if not forever," and "…for the eugenic good, women should be permitted to become respectable mothers without having to live with the fathers of their children" (Shaw 1903). Havelock Ellis argued similarly that "The realization of eugenics in our social life can only be attained with the realization of the woman movement in its latest and completest phase as an enlightened culture of motherhood…" (Ellis 1912).

By contrast, Karl Pearson opposed the Factory and Education Acts, legislation designed to outlaw child labor, on the ground that it would discourage sturdy working-class families from bearing more children. On the other hand, Harold Laski, likewise a leading socialist intellectual, supported the legislation on the grounds that a higher minimum wage would permit parents to support more children.

For committed eugenicists, the other side of the coin was to reduce fertility in "undesirable" elements in the population, and especially among people classified as "feeble-minded." Genealogies published on both sides of the Atlantic roused public opinion in favor of sterilization, not a policy that would have won favor with earlier eugenicists. For instance in the United States, Goddard, in 1912, described the Kallikak Family (a pseudonym), whose supposedly typical members comprised not only those of low normal

intelligence, but also criminals, prostitutes, and alcoholics (Goddard 1912). In an earlier similar enterprise Dugdale, in 1877 in upstate New York, had described the notorious Jukes family (also a pseudonym), who had among their group large numbers of criminals and sociopaths (Dugdale 1975). In 1915, Estabrook followed up the Jukes family, with equally pessimistic findings (Estabrook 1916). Charles Davenport set up a Eugenics Record Office to identify all eligible individuals, and established his Institute in Cold Spring Harbor. He was committed to the view that "undesirables" should be discouraged from breeding and proposed legislation, enacted in several states, which permitted involuntary sterilization of "feeble-minded persons." In the post–World War I period, this type of Social Darwinism elicited strong reaction from such researchers in the field as E.O. Lewis and Lionel Penrose in Britain, Joseph and Helen Wortis in the United States, and many others, as described in Chapter 21.[10]

References

Bateson W (1909). *Mendel's Principles of Heredity.* Cambridge, UK: Cambridge University Press.

Bowler PJ (1990). *Charles Darwin: The Man and his Influence.* Oxford, UK: Blackwell.

Box JF (1978). *R.A.Fisher: The Life of a Scientist.* New York, NY: Wiley.

Desmond AJ, Moore JR (1992). *Darwin.* New York, NY: Warner Books.

Dugdale RL (1975). "*The Jukes": A Study in Crime, Pauperism, Disease, and Heredity, also Further Studies of Criminals.* Fifth ed. (first ed. 1884). New York, NY: AMS Press.

Ellis H (1912). *The Task of Social Hygiene.* Boston, MA: New York.

Estabrook AH (1916). *The Jukes in 1915.* Washington, DC: The Carnegie Institution of Washington.

Galton F (1888). Co-relations and their measurements, chiefly from anthropometric data. *Proc R Soc* 45:135–145.

Galton F (1894). *Natural Inheritance.* New York, NY: Macmillan.

Gillham NW (2001). *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics.* New York, NY: Oxford University Press.

Goddard HH (1912). *The Kallikak Family: A Study in the Heredity of Feeble-Mindedness.* New York, NY: The MacMillan Co.

Gruber HE, Darwin C, Barrett PH (1974). *Darwin on Man: A Psychological Study of Scientific Creativity.* 1st ed. New York, NY: E. P. Dutton.

Haller MH (1984). *Eugenics: Hereditarian Attitudes in American Thought.* New Brunswick, NJ: Rutgers University Press.

[10] See Marks (1993) for a historiography of eugenics; also Haller (1984) for a history of the eugenics movement from its origins through the first half of the twentieth century.

Harper PS (2005). William Bateson, human genetics and medicine. *Hum Genet* 118:141–151.

Kevles DJ (1995). *In the Name of Eugenics: Genetics and the Uses of Human Heredity.* Cambridge, MA: Harvard University Press.

Marks J (1993). Historiography of eugenics. *Am J Hum Genet* 52:650–652.

Mayr E (1991). *One Long Argument: Charles Darwin and the Genesis of Modern Evolutionary Thought.* Cambridge, MA: Harvard University Press.

Shaw B (1903). *Man and Superman: A Comedy and a Philosophy.* Westminster (London, UK): Archibald Constable.

Susser M (1977). Judgement and causal inference: Criteria in epidemiologic studies. *Am J Epidemiol* 105:1–15.

White M, Gribbin JR (1996). *Darwin: A Life in Science.* New York, NY: Dutton.

14

Furthering the Epidemiology of Social Gradients and Disease: Goldberger and Sydenstricker

Development of the theme of social determinants in the health of societies has been uneven. For our purposes, the advent in the early twentieth century of the partnership of Joseph Goldberger and Edgar Sydenstricker in the United States marks a notable advance in social epidemiology as a science.

We may remind ourselves, however, that as early as the fourth century BCE Plato commented on the different meaning of sickness for a rich man and a poor one; that the merchants of Venice in fifteenth-century Italy well understood the consequences and economic significance of an epidemic; that in seventeenth-century London, John Graunt first thought to distinguish the distribution of deaths from plague in residential areas of the wealthy from those of the poor, which he did by enumerating the location of deaths from the Bills of Mortality routinely collected and recorded by the city. Also in the seventeenth century, Bernardino Ramazzini (1633–1714) provided a systematic exposition of occupational diseases. In late eighteenth-century France, philosophers and revolutionaries first actively promulgated the idea that health be accorded the status of a human right. In post-revolutionary Paris in the early years of the nineteenth century, in nascent epidemiological studies both Villermé and Parent-Duchatelet implicated poverty—a descriptor assigned according to the physical conditions in each of the several defined

residential districts (*arrondissements*) of Paris—as in some way responsible for the differences in health among them.

In mid- and late nineteenth-century Berlin, Rudolf Virchow the founder of cellular pathology and the brilliant student of such distinguished teachers as Jacob Henle, Ernst Haeckel, Johann Lucas Schonlein, and Johannes Müller, had expounded the view that epidemics were the manifestation of social and economic maladjustment. It followed that social and economic determinants of health deserved study no less systematic and rigorous than did biological determinants. During the European-wide revolutionary movements of 1848, Virchow joined an activist group. The group pressed the view that societies had an obligation to protect and secure the health of their people. To meet that obligation, Virchow proposed and advocated the need for three elements: recognition of the key role of social factors in health; the scientific study of those social factors; and the legal entrenchment of a right to health. In pursuit of these goals, Virchow contributed a theoretical basis for much that followed. Important as were these early forerunners of the ideas of present-day epidemiology, in reality they were realized as somewhat disjointed explorations of social gradients in health. This work was not, however, the immediate stimulus for any sustained development of the theme.

As described in Chapters 6 and 7, in England in the 1840s William Farr added social class categories to age, sex, and location as descriptors of mortality, thereby adding a new dimension of analysis to demographics, mortality rates, and the public health. In the third quarter of the nineteenth century in England, as Chief Medical Officer Sir John Simon applied administrative practices and systems that could improve the health of the public at large, and particularly that of the poor.

Friedrich Engels, as noted in Chapter 6, was the son of a German textile manufacturer. He was an avid student of the growing body of socialist writings in Europe, and became the lifelong collaborator of Karl Marx. In 1842, his father had dispatched him to England to act as his agent in the rapidly growing industrial city of Manchester. Walking the streets and visiting the homes and factories, he made notes on all he saw; he also summarized the written statistics of the time. Within two years, he published his classic account, a passionate exposé: *The Condition of the English Working Class in 1844* (first published in German, and not available in English until 1877) (Engels 1844, 1993).

Engels' view was that the dreadful living conditions of the poor in British cities were part of an historical process, namely, the forces of the industrial revolution that drove agricultural workers into the cities. He recognized, as also did his friend and political guide Karl Marx, that the rising momentum of capitalism and industrialization underlay the emergence of a new industrial society. From this perspective, both capitalists (manufacturers,

mill-owners) and an urban proletariat driven from their rural habitat were part of the same system. In Lancashire where he made his observations, he saw plainly that the very sites and distributions of workers' homes were based entirely on the requirements of the capitalist mill-owners and took no account of the residential needs of the new working class mill-laborers.[1] Paradoxically, Engels argued that "wage-slaves" could gain relief from their plight only by uniting, the one objective made achievable by the geographic concentration of workers in the industrial cities.

Engels carried out his investigations at the peak of the industrial revolution in England. Its constant accompaniment was an influx of workers drawn from the countryside into the cities. There they would endure the inhuman conditions of life and labor inflicted on the workers in the "satanic mills," so described by the poet William Blake. Even more severely, Engels in his detailed studies excoriated a system in which workers perforce endured the deprivations of poor housing, nutrition and sanitation, and deplored the accompanying iniquities of child labor, crime, insecurity, alcohol and the deadening routines of factory work that governed their lives. From such data as were available, he also documented the evils of opiates, patent medicines, and lack of access to medical care. He scoured the national and local mortality statistics available in the annual reports compiled by William Farr (Chief Statistical Officer to the Registrar General), from which he drew mortality rates of the "new" urban populations serving the cotton mills and other growing industries. Much as the poor compared with the well-off, the rapidly growing towns they inhabited compared unfavorably with settled older towns.

Unlike Chadwick and other native sanitary reformers Engels, a foreigner in England, did not and probably could not immediately stimulate reform. Neither did his methods conform with the scientific approach Virchow sought (Engels 1844, 1993). The legal reform of the English health system that Engels hoped for, which would confer health rights on the poor, had to await later statistical surveys. Nevertheless, in the mid-nineteenth century the concurrent writings of Friedrich Engels and the convincingly dramatic novels of Charles Dickens (Marcus 1974) provided an overture to the social surveys initiated at the turn of the century, notably those of Charles Booth and Seebohm Rowntree. These were systematic studies of living conditions that exposed numerically and by direct observation, if less dramatically than previous accounts, the miseries of the poor in England.

These social surveys contain the elements of an embryonic social epidemiology. They cleave to the ideal of the revolutionary French humanists at

[1] The Minority Report of the Royal Commission on the Poor Laws in England 1909 (authored by Beatrice Webb; see 1974 ed.), argued strongly that pauperism was not due to personal failure, but to capitalist exploitation.

the turn of the eighteenth century, which asserts a right to health for people at all levels of society. So broad an objective calls for the study of social factors in health, disease, and the organization and evaluation of public health and medical care. By the beginning of the twentieth century in Britain, France, Germany, and the United States, such studies were being seriously undertaken.

To produce persuasive material was certainly the motivation for the work in London of Charles Booth (1840–1916), a founder and leader of the Salvation Army. He was a cousin of Beatrice Webb who, together with her husband Sydney Webb and the brilliant playwright George Bernard Shaw, were members of the influential and socialist Fabian Society. Booth carried out systematic surveys of the conditions of the families of the poor in London, extracted hard social and economic data and, over the succeeding decade, published his voluminous findings (Booth 1889, 1892, 1902).

At the turn of the nineteenth century, Seebohm Rowntree (scion of the wealthy Quaker family and chocolate manufacturer) formed the view that poverty was the root of ill-health, the reverse of the sequence assumed by Chadwick. Stimulated by the work of Booth, Rowntree proceeded to survey the living conditions of the poor in York, his native city. *Poverty: A Study of Town Life,* appeared in 1901 (Rowntree and Bradshaw 2000). In the interests of improving the conditions of the poor, Rowntree aimed to answer the question of the sources of poverty. He sought to discover how much was due simply to insufficiency of income, and how much to improvidence. Rowntree's inquiries thus bridge Engels' *Condition of the Working Class* and Sydenstricker's later surveys in the United States. Forcefully and systematically, his work made clear the relevance of wage levels to the health and quality of life of the population. Rowntree concluded, as Engels had done, that while wages set absolute limits to survival, even when resources were adequate to provide for food and shelter, man could not and should not live by bread alone.

In the United States, Sydenstricker seems not to have acknowledged Rowntree as a predecessor. Rowntree had emphasized that standards of living of the poor depended not only on wages but also on the number of dependents for whom the family needed to provide. With that realization, Rowntree developed a formula for nutritional requirements of family members, which he derived from dietary studies of Atwater and Woods in New York City in 1895–1896. Rowntree had studied the chemistry of nutrition with a view to directing the laboratory side of the family chocolate factory in York. Consequently, he was as knowledgeable about nutrients as anyone of his time.

To obtain measures of population health in York, Rowntree collected notices of births and deaths across place and over time, and further specified these vital statistics by age, gender, level of poverty, and causes of death. Among school children, he measured height and weight, and then made a

judgment of their overall health. For data on adults, he turned to War Office data on potential recruits for the South African War of 1899–1901.

Rowntree was fortunate in that so much of his work was put to good use in his lifetime. At the time of the South African War of 1899–1902 (the Boer War), his data on the effects of poverty on children reinforced the dramatic finding that a huge proportion of recruits was not sufficiently healthy and well-grown to meet the Army's standards and furnish its needs. Later, for the influential *Minority Report of the Poor Law Commission* (1909), he provided ammunition for the Fabian critics in the form of relevant data on the poor. Still later Rowntree's personal friendship and collaboration with Lloyd George, then Prime Minister, enabled the Liberal government to extend to families the health insurance provided for workers. In further socially oriented legislation, Lord Beveridge sought Rowntree's counsel in fixing monetary scales for National Insurance and National Assistance. "In Britain today the social assistance which determines the living standards of one in eight of the population owes its origin to Rowntree's 1899 poverty standard" (Rowntree and Bradshaw 2000).

In the United States, an extraordinary data-based series of investigations of the ills of poverty was begun in New York in 1917. Joseph Goldberger (1874–1929), a physician in the United States Marine Service (later, the United States Marine Hospital Service), was assigned to an investigation of pellagra.[2] In a historic partnership with the statistician and economist Edgar Sydenstricker (1881–1936), they set about the problem jointly. Sydenstricker introduced a systematic way of thinking about diseases, their social distribution and the methods to study them, and the implications for health policy (Sydenstricker 1974). Indeed, he pioneered rigorous survey methods not only in the United States but subsequently for the League of Nations.

Joseph Goldberger, in himself and in his career, personifies the emergence of epidemiology as a discipline built on its own logic from data founded on observation, experiment, and evidence-based action at the population level (Terris 1964). Goldberger had arrived with his family in New York in 1881 as a six-year-old Jewish immigrant from Hungary. He made his way through the educational and occupational opportunities available to him, to qualify as a physician. Ambitious, and excited by the prospect of combating disease, he managed to obtain an assignment with the team of the United States Marine Hospital Service then contending with yellow fever in Mexico (not without contracting the fever himself in the course of his work).

In a highly productive career, among the initial questions Goldberger studied and resolved was whether the larvae and offspring of yellow

[2] See Kraut (2003) for a recent biography.

fever–carrying mosquitoes could themselves be carriers of the often mortal disease. In careful experiments he rejected this postulate, thereby setting narrower limits to the potential modes of transmitting the infection. Turning to studies of typhoid, in a medical venue evidently ignorant of the work of William Budd in Bristol in the nineteenth century, Goldberger recognized as Budd had done the role of contaminated water in transmitting *B. Typhosis*. Again, in seeking the source of a form of typhus, and having found that he was able to transmit the infection to monkeys from lice that carried the organism, he thereby traced the source of infection to infestations of head and body lice. Later, when faced with finding the cause of an eruptive skin condition that afflicted sailors and other temporary residents in boarding houses in the surrounds of Philadelphia, he identified the responsible mite that infested the straw bedding that served the victims for mattresses.

Following these exemplary shoe-leather exercises, during a diphtheria epidemic in Detroit where he was busy seeking and mapping carriers as potential sources, Goldberger was summoned by the Public Health Service to turn to the investigation of pellagra, a condition then epidemic among poor rural laborers in the South. Goldberger, fresh from his several spectacular discoveries of infections as causes of disease, was a logical choice to head an investigation of the cause of pellagra. An official Pellagra Commission had been investigating this disease for some months. Shortly before Goldberger set about his new task, the commission had already dismissed a then current hypothesis that the ingestion of spoiled maize was a cause. In all probability, the commission concluded, pellagra was in their words, *"a specific infectious disease communicated from person to person by means at present unknown."*

Before Goldberger plunged into the challenge of this new field, he settled himself in the library to read everything to be found about pellagra, and made field visits to epidemic sites in prisons and mental hospitals. After three months of study, he proffered a paper rejecting the standard hypothesis that pellagra could be infectious. He rejected that hypothesis because no record showed any member of the nursing or administration staffs ever to have been affected by pellagra.[3] From his readings, and more especially from his visits to mental hospitals and prisons, he had discovered that it was in the spring that mental patients and prisoners regularly contracted the disease. In the days before electrical refrigeration, and after a winter with little fresh food was to say that pellagra could be the result of deficient diets.

[3] Typical of Goldberger's hands-on approach, he visited a mental hospital in Georgia where pellagra was rampant, and walked around the dining hall at mealtime. Although he had been assured that staff and patients ate the same meals, his personal observation gave the lie to these assurances. Staff were clearly favored in quality and quantity.

Although Goldberger's work reads as convincingly today as it did then, his conclusion met with much skepticism. In the early decades of the twentieth century, memories still lingered of the nineteenth-century battles over 50 years and more to establish infection as a source of epidemic disease. In that light, given a nutritional cause it was hardly surprising that his papers were received with skepticism and even hostility. Experimental evidence was called for.

This he provided, to begin with in rather rough and ready form later to be refined. He sought out and selected three orphanages where, among children who had developed pellagra in previous years, the condition had recurred in at least 50% in the subsequent year. Goldberger then supplemented the diet of those children with both legumes and fresh animal protein (meat, milk, and eggs). Conclusively, pellagra did not recur. Later he turned to work further with a prison population. There he recruited 12 white male convicts who, in exchange for an offer of pardon, volunteered for a dietary experiment. They were to limit their diet to a regime typical among the poor of the South. More than half the volunteers (6 of 11) developed pellagra (one had sickened and dropped out of the experiment).

Despite the success of Goldberger's creative experiments, his results did not at once dispel disbelief. Skeptics of that time were reluctant to accept that flagrant disease could be attributed to deficiencies in diet and foodstuffs. In the face of these two studies, and even after Goldberger had ruled out infection by injecting blood from pellagra patients into himself, his wife, and his friends without any one contracting the disease, opposition to the dietary theory persisted. The positive results of the subsequent trials in the field of such supplementary foods as milk, fresh vegetables, and meat, however, made a compelling case for prevention.

Goldberger turned next to two different approaches to test his conclusions. One of these was to return to the laboratory where, in experiments with dogs, he established to his own satisfaction that the disease of black-tongue in dogs was a model for pellagra. In another approach to the pellagra problem, together with Sydenstricker, he developed the population surveys in the mill-towns (described below). Although during Goldberger's lifetime no specific dietary deficiency was discovered, discovery of the effects of Vitamin B (later found to be niacin) ultimately rendered his conclusion definitive.

In his papers written almost a century ago, the logic of Goldberger's argument connecting pellagra with diet stands out as highly convincing. In his lifetime, however, his results seemed to challenge the dominant late nineteenth-century epidemiological paradigm of infection by microorganisms. One is reminded that the contention about the vital role of infection in the rampant spread of disease in populations had roiled the greater part of the

previous century before the idea was fully accepted. By the beginning of the twentieth century, however, historical conviction about infection as a cause of disease was entrenched and not easily rejected.

In the face of persisting disbelief, as mentioned above Goldberger devised new strategies to back his results. Thus, he followed up his human experiments with a survey of residents in mill villages in South Carolina. Two research teams visited the homes, one to assess diets, and the other to assess sanitary conditions. In order to measure the pollution levels of water drawn from streams as a possible source of infection, Goldberger used a scaled assessment developed by Wade Hampton Frost, his contemporary and equally adroit colleague in the Public Health Service. The thoroughly convincing results from these two sets of data were compiled and analyzed by Edgar Sydenstricker, an officer of the public health service trained in both sociology and economics. This study sealed the famed association of Sydenstricker and Goldberger that was to last until Goldberger's early death in 1929 at the age of 55.

Edgar Sydenstricker (1881–1936), son of Presbyterian missionaries in China and brother of the novelist Pearl Buck, was born in Shanghai. There he was schooled at home until sent to the United States in 1896 at age 15. After studies in sociology and economics, he first taught school in Virginia, and then edited a local newspaper. In 1907 at the age of 26, however, he turned to postgraduate study of political economy at the University of Chicago. Thereafter, he began his systematic surveys on behalf of both the US Immigration Commission and the US Commission on Industrial Relations. These data provided a basis for his reports on industry, standards of living, wages, and also on working conditions with a special focus on the foreign born. With such work behind him, in 1915 he was appointed the first public health statistician for the Marine Hospital Corps (as noted, later to be renamed and converted into the US Public Health Service). Soon after, Sydenstricker together with Goldberger began to study pellagra in the population, a plague more especially in the cotton mill villages of South Carolina.

From the perspective of the epidemiologist, the hunt for the cause of pellagra was well served by these surveys. From the individual perspective of Sydenstricker himself, they were pivotal in his own further development. Pellagra was then a widespread and hugely debilitating if soldom fatal disease of unknown cause. Sydenstricker was of the view that given such a persisting chronic condition, mortality statistics alone could never provide the completeness required for comprehending fully its effects on the health of the people. He was already convinced by his earlier studies of tuberculosis (discussed below), that for such a purpose, one would need data also on the prevalence of the disease and its associated ills (which he defined). For most existing surveys of the almost exclusively chronic disorders under study,

comparisons among such data would often be limited either by the available resources, or by variations across place and season, or by the feasibility and quality of the preferred research designs.

In Sydenstricker's epidemiological work with sick benefit associations, he had soon become aware that for his purposes, from one industry to another data collection and recording were often haphazard and needed to be standardized and better specified. To take a simple example, industrial sick benefit records would show absences from work but not their duration. By sustained effort, Sydenstricker achieved needed improvement. Nevertheless, however satisfactory the improvements in such industrial data might be, they could tell nothing about adults who were either out of work or were not workers, nor about the very young, the very old, and the disabled. Moreover, among the poor, often illiterate, and sometimes uncomprehending villagers in South Carolina, in the course of his interviews he was disconcerted to find that among his respondents, the recall of even quite serious conditions was often poor. Thus, upon repetition, the responses of those interviewed about sickness recorded in the benefit data seemed to retain their validity and accuracy for no more than a few weeks. In any case, he came to realize that such prevalence data could supply neither measures of incidence nor much useful information about a past history of acute conditions. Step-by-step, Sydenstricker focused his thinking about population-based health statistics in terms of what was needed and useful, how such data could be acquired, and to which uses they might be put.

At Sydenstricker's instigation, in 1921 the US Public Health Service initiated the Hagerstown Morbidity Studies. Located in Maryland, not far from Washington, DC, Hagerstown seemed typical enough of much of the country and, in an earlier investigation, the people had proved receptive. Moreover, population studies had some few antecedents to draw on, as in the seldom mentioned survey of tuberculosis morbidity by Armstrong in Framingham in 1917 (Susser and Stein 2005). Sydenstricker wrote as follows about his plans for Hagerstown: "The chief aim of the study was a record of illnesses, as ordinarily understood, that were experienced by a population group composed of persons of all ages and both sexes, and in no remarkable way unusual" (Sydenstricker 1974).

Given the limitation on resources, Hagerstown itself comprised too large a population for a survey to cover completely. The researchers resorted to a selected population sample, validated against the most recent census, and representative of the economic strata of the area. The surveys were conducted at various times to reflect seasonal changes in health and sickness. Further, given the duration of the study over a 25 month period, the researchers devised the technique of "person years of observation" to cope with potential biases incurred by inevitable losses in the study population.

With the experience of the strengths and weaknesses of the completed Hagerstown survey behind him, Sydenstricker set about creating the foundation for the execution of national studies. He began with National Health Interviews, followed subsequently by National Health Examination Surveys. These bold and imaginative undertakings, creatively developed and modified by experience over time, have provided modern epidemiologists in the United States with remarkable databases that are now available online for anyone with an idea or a question to pursue. Long-established surveys include the National Health and Nutrition Examination Survey (NHANES), the National Health Interview Survey (NHIS), the National Immunization Survey (NIS), and the National Survey of Children's Health, to name a few. More recent initiatives include the National Household Study on Drug Abuse, the Medical Expenditure Panel Survey, and the Behavioral Risk Factor Surveillance System.

Behind Sydenstricker's pursuit of valid descriptors of health status in populations there lay a concern with poverty and social disadvantage, as the same concerns also had stimulated the work of Goldberger, Herman Biggs, Seebohm Rowntree, and others. Concern for the disadvantaged had in fact preceded Sydenstricker's turn from economics to the new commitment to the public health. When earlier he had come to work with Goldberger in South Carolina, Sydenstricker applied himself to developing improved descriptors of poverty and of family standards of living. Thus he would check reported wages at the source of payment, and in weighing the adequacy of the amounts he appropriately allowed for family size and especially for the number of dependent children. In his studies of pellagra, the care he took with detail and the specificity of his data had proved of inestimable value in establishing the relation of the disease to standards of living (Figure 14.1 and Table 14.1).

In 1976 I.S. Falk, revered as a socially conscious public official in the Department of Health and Social Services, submitted an insightful foreword to the collected works of Sydenstricker. Falk describes his own early efforts between the two world wars as these two pioneers and collaborators—both members of the US Public Health Service—together struggled to make an impact on health policy in the United States (Sydenstricker had died in 1936). Falk's foreword and analysis seems no less relevant than it did 60 years ago, as in the following brief quotation: "The health services have long been in a troubled state in the United States, because it has not been feasible for the nation to have a clear, unambiguous or durable social policy for health" (Falk 1974). Falk attributed the failure in the United States to achieve a coordinated medical and public health service to the persistent dichotomy through the twentieth century as to what is appropriately public and what private, with a predominant leaning of both government and the public toward the private.

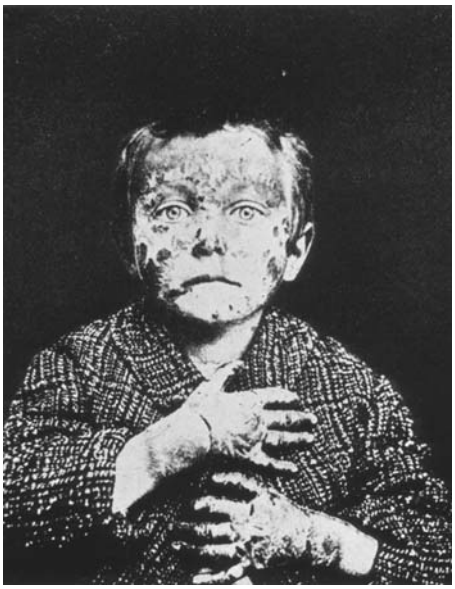


Figure 14.1 Child with Pellagra: Half-length view of a young boy with severe dermatitis on face and hands. Credit: Courtesy of the National Library of Medicine.

Table 14.1 Pellagra incidence during 1916 among households of cotton-mill workers in seven villages of South Carolina, whose supply of fresh meats was less than one pound, and of fresh milk less that four quarts, per adult male unit per 15-day period, classified according to the household supply of wheat flour per adult male unit for a 15-day period between April 16 and June 15, 1916.

Household Supply of Wheat Flour in Pounds per Adult Male Unit, for a 15-day Period	Total Number of Households	Number of Households Affected with Pellagra	Percent of Households Affected with Pellagra
All amounts	175	29	16.6
0–3.9	15	5	33.3
4.0–7.9	8	0	0
8.0–11.9	44	8	18.2
12.0–15.9	49	6	12.2
16.0 and over	59	10	17.0

Adapted from Terris M, ed. *Goldberger on Pellagra* (1964).

In his surveys of villages, towns, and industries, Sydenstricker was made sharply aware of the burdens of chronic and acute illness and disability that beset the families of the poor. In 1923, however, Sydenstricker was invited to work in matters of health and health administration with the League of

Nations in Geneva. There he made himself familiar with European approaches toward social insurance to cover episodes of illness. Strongly critical of such approaches, however, in their stead Sydenstricker favored health insurance that would include prevention. His own preferred approach was to strengthen the capacity of services to prevent disease. For diseases that could not be prevented, he argued for health and medical services that worked toward advancing both diagnostic methods and the timeliness of therapeutic interventions. During the last two decades of his life, he offered these ideas repeatedly both in the counsel he gave and the papers he published.

Health advocates have put forward such recommendations intermittently both before that time, and from that time to this. In the United States, acceptance by lawmakers has been limited, excepting perhaps the era of President Franklin Roosevelt's New Deal initiated in response to the Great Depression of the 1930s. After World War II, in glaring contrast, many such ideas were embodied in the British National Health Service and across much of Europe.

In 1914, the American Association for Labor Legislation had sought to launch health insurance, and in 1916 and 1917 Warren and Sydenstricker, working in the US Public Health Service, set out proposals to build health insurance into a public health program. In 1920 the American Medical Association, together with much of the insurance industry, firmly rejected such proposals and did so repeatedly, foiling the major efforts after World War II and, later, those of the Clinton administration in 1994. Here we perhaps observe a recurrent political pattern in the United States: health professionals generate national proposals for socially beneficial plans, whereupon special interest groups bring all the power they can muster to reject proposals perceived as against their narrow interests.

Epilogue

A closing comment on the status of social epidemiology may be fitting. By the third decade of the twentieth century, there seemed to be a relative hiatus in international communication across the field: relative, that is, to the apparently instant cross-Atlantic transfer of knowledge and technology apparent, for instance, in narrower sciences such as in the microbiology. Chadwick and Simon certainly had connections with many others in England and Europe, as did Herman Biggs and his team in New York with Simon in England, and with others in Prussia and France. Engels was obviously an avid reader of the work of others on poverty, commenting for instance on the critique of Carlyle, not to mention his careful study of Lewis Henry Morgan's *Ancient Society* to develop his own treatise, *The Origin of the Family, Private Property and the State*. Rowntree was naturally deeply influenced by Booth, nor does he

neglect to quote Engels' work. Despite the influence of Rowntree's book *Poverty in England*, Sydenstricker, who references such demographers as Pearl, Dublin, and the other Metropolitan Life writers like Frankel, seemed not to see either Rowntree or his work as worth mentioning.

Nor are all of these gaps in communication limited to cross-Atlantic connections. Thus in many papers, Sydenstricker, always deeply interested in the epidemiology of tuberculosis, reported mortality figures by time and place. However, he makes no mention of the work in the field of either Donald Budd Armstrong in Framingham, Massachusetts, or of Arthur J. Myers in Minnesota. Both Armstrong and Myers were prominent contemporary physicians deeply involved in research and prevention of tuberculosis. The omissions seem the more glaring in that Sydenstricker resorts to the data on tuberculosis in Massachusetts published by the Metropolitan Insurance Company, the organization that sponsored Armstrong's pioneering studies in Framingham. Yet Armstrong's obituary (Wheatley 1968) noted that his published results from Framingham effected a turning point in community action on tuberculosis. Nor in the extensive writings of Myers, as we noted elsewhere, can we find any acknowledgment of the work of Armstrong, who was by no means an obscure figure. In 1916 Armstrong was elected the first Chair of the Sociological Section of the American Public Health Association. He was also elected Assistant Secretary, and later the Executive Officer, of the National Tuberculosis Association. Although Armstrong does not refer to Myers, much of his writing on tuberculosis precedes the main impact of Myers's work in Minnesota. In turn, Armstrong himself—who devised his survey of tuberculosis as well as the subsequent institution of a community-based program in Framingham—makes no reference either to the several community studies of health in England, most obviously Rowntree's *Poverty*, nor does he refer to the mid-nineteenth-century publications of William Budd, Edwin Chadwick, John Simon, and others in England, or to Villermé, Parent-Duchatelet, and others in France (Susser and Stein 2005).

Returning to Sydenstricker, again a publication on the history of surveys in the United States scarcely mentions his name (Duncan 1984). A possible reason for Sydenstricker's lack of comment on Armstrong's Framingham work might reside in his own report on tuberculosis in the New York county of Cattaraugas. There his interpretation of the contribution of the effects of community policies in preventing tuberculosis is cautious indeed. He was not ready to commit himself to the inference that the observed decline in the rates of the disease was owed to the change in policies. He comments several times on the need for caution in evaluating health changes, with undeniably justified reservation. Perhaps the absence of any comment from Sydenstricker on Armstrong or Myers was a reflection of a personality far from combative.

Harry Marks, a strong recent American critic of Sydenstricker's, argues in a comment on his work on pellagra and other disorders that he neglects both African Americans and women (Marks 2003). Marks goes further in seeing these omissions as proof at the time of the lack of interest in the conditions of these groups. In England, parallel criticism was recently leveled at Seebohm Rowntree. Of course, we are all products of our time. The demand for ethical standards in contemporary medical research is couched in terms of universal human rights. Many such rights took time to take form while others took time to be revised or discarded.

References

Booth C (1889). *Labour and Life of the People*. London: Macmillan and Co.

Booth C (1892). *Life and Labour of the People on London*. London and New York: Macmillan and Co.

Booth C (1902). *Life and Labour of the People in London*. London: Macmillan and Co.

Duncan OD (1984). *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation.

Engels F (1844/1993). *The Condition of the Working Class in England*. Ed. with an introduction by David McLellan. Oxford England: Oxford University Press. First published in 1844.

Falk IS (1974). Commentary. In Kasius, ed. *The Challenge of Facts: Selected Public Health Papers of Edgar Sydenstricker*. New York: Prodist, for the Milbank Memorial Fund, pp: 95–107.

Kraut AM (2003). *Goldberger's War: The Life and Work of a Public Health Crusader*. New York: Hill and Wang.

Marcus S (1974). *Engels, Manchester, and the Working Class*. New York: Random House.

Marks HM (2003). Epidemiologists explain pellagra: Gender, race, and political economy in the work of Edgar Sydenstricker. *J Hist Med Allied Sci* 58:34–55.

Rowntree BS, Bradshaw J (2000). *Poverty: A Study of Town Life*. Special centenary ed. Bristol: Policy.

Royal Commission on Poor Laws and Relief of Distress, Webb S, Webb BP (1974). *The Minority Report of the Poor Law Commission*. Clifton NJ: A.M. Kelley; Reprint of the 1909 ed.

Susser M, Stein Z (2005). Commentary: Donald Budd Armstrong (1886–1968)—Pioneering tuberculosis prevention in general practice. *Int J Epidemiol* 34:1191–1193.

Sydenstricker E (1974). *The Challenge of Facts: Selected Public Health Papers of Edgar Sydenstricker*. New York: Prodist, for the Milbank Memorial Fund.

Terris M**,** ed. (1964). *Goldberger on Pellagra*. Baton Rouge: Louisiana State University Press.

Wheatley GM (1968). Donald Budd Armstrong, M.D.—1886–1968. *Am J Public Health Nations Health* 58:2007**–**2009.

15

Epidemiology After World War II: New Times, New Problems, New Players

Prologue

Epidemiology as a discipline is made unique by the combination of its way of looking at things—the theory and methods of population studies—and the subject of its studies, states of health, and their determinants in the populations under study. In this view, what distinguishes the late twentieth-century era from those preceding are the concepts that guided epidemiologists. It is not so much the particulars of their discoveries about states of population health, as it is changes in theories of disease and the consequent creation of a body of methods to test those theories. These theoretic and technical ideas evolved from changes in society and from concurrent changes in the configuration of health and disease.

The focus in this section is on the theories and techniques developed to cope with the changed spectrum of diseases in the latter decades of the twentieth century. A primary source is precisely the literature of epidemiology on theory and method. In the period under review, theory and method have

Chapters 15–19 are an updated and expanded version of Susser M (1985). Epidemiology in the United States after World War II: the evolution of technique. *Epidemiol Rev* 7:147–177.

become predominant in the education of epidemiologists. It is this literature, and the growing numbers of women and men who teach and who learn it, which have given epidemiology standing as a mature discipline.

Epidemiology today is fully institutionalized through academic departments in major universities, an assembly of texts in common, doctoral programs capable of producing faculty with primary training in the discipline, and a research output that has a significant and continual impact on research and practice in the health field. This is a phenomenon of the post World War II era. One may ask why so late? And why then?

To recognize the existence of maturational delay, one might compare the lapse of time, on the one hand between the work of William Harvey on circulation of the blood in the third decade of the seventeenth century, and the founding of physiology as an academic discipline in France only in the mid-nineteenth century and, on the other hand, with that between the work on enumeration of John Graunt and William Petty in the mid-seventeenth century and the like founding of epidemiology as a discipline in the early twentieth century. In this comparison of the basic sciences of medicine and of public health, epidemiology lagged by almost a century. Admittedly, there were stirrings long before that, if not in the form of an established discipline. In fact, the study of associations between environmental factors and disease had gained apace early in the nineteenth century. In France, the work of Louis René Villermé and Parent-du Châtelet, in England, that of Edwin Chadwick, Florence Nightingale, William Farr, John Snow, John Simon, and William Guy, and in Germany, that of Henle and the young Rudolf Virchow, not only enumerated and quantified their data but also studied communities and populations selected to illuminate such issues as the effects of drainage, housing, occupation, and nutrition.

In England throughout the nineteenth century, rapid industrialization, urbanization, and population growth together created the need to confront a changing pattern of disease that flourished in the rapidly growing urban slums. The challenge excited the medical scientists of the time. John Simon, as Chief Medical Officer of the General Board of Health from 1855 to 1858, could draw the most brilliant young doctors in the country into his epidemiologic team (Brockington 1965). Florence Nightingale and Edwin Chadwick, too, although misguidedly committed to miasma theory until each died in the last decade of the nineteenth century, had a substantial and beneficial impact on maintaining hygiene.

By the last quarter of the century, however, the pace had slowed. An unfavorable political climate and reactionary politics frustrated the efforts of public health workers (Simon 1890; Lambert 1963). At the same time, however, another brilliant but differently focused era opened for public health with the discoveries of microbiology.

Some of the maturational lag of epidemiology can be attributed to the enormous successes of the germ theory ascribed to Louis Pasteur, Jacob Henle, Robert Koch and others (Ackerknecht 1948; Dubos 1959; Cassel 1964; Susser 1973). In terms of the fresh insights of Ludwik Fleck and Thomas Kuhn about scientific revolution, this was the acceptance by the scientific community of a new paradigm (Kuhn 1970; Fleck 1935), one that eliminated its chief rival, the persistent miasma theory (Susser 1973). As a result, the Henle-Koch postulates—laboratory-based criteria for judging the causal relationship of organism to lesion (often attributed to Koch)—displaced the population-based inferences typical of the work of the earliest modern epidemiologists. In the late 1880s, Koch began to apply the discoveries of Pasteur to the elucidation of specific human infections. Quite soon, the focus of epidemiology was reduced to the pursuit of specific microbial agents, singular causes, and the means of preventing their consequences.

In the face of germ theory, in the first half of the twentieth century the population concepts and quantitative skills developed in the first half of the nineteenth century were scarcely maintained. In England, some rare souls like J. Brownlee, M. Greenwood, W.0. Kermack, and A.G. McKendrick continued the quantitative tradition as epidemiology (Susser 1973). In the United States, too, these notions of epidemiology were sustained by a few like Joseph Goldberger, Edgar Sydenstricker (as discussed in Chapter 14), Charles V. Chapin and Wade Hampton Frost: in 1921, Frost was appointed to the founding chair of the discipline at the Johns Hopkins School of Hygiene and Public Health. In distilling general principles of the epidemiology of specific infections, Frost expressed concisely the interactions of microorganisms, populations, and environmental conditions (Frost and Maxcy 1941). In studying particularly chronic infectious diseases like tuberculosis, and infections by unknown organisms like poliomyelitis, he was responsible for theoretic formulations that were adapted and developed in the subsequent phase of the epidemiological study of chronic noninfectious disease (Frost and Maxcy 1941).

Parasitologists like Ronald Ross (1857–1932) in Britain and Theobald Smith (1859–1934) in the United States also contributed much to broadening the focus of attention in epidemiology. Ross' Theory of Happenings of 1910, aimed at the prevention of malaria, has a niche as the beginning of multivariate modeling in epidemiology (Susser 1973). It became apparent that the life cycles of parasites that caused human disease involved reciprocal relationships between agent, host, and environment that could not be ignored. Theobald Smith, in his book *Parasitism and Disease* (Smith 1934) elaborated the concept of host and parasite. Others used these ideas of reciprocal interactions between multiple factors to launch a new phase in epidemiology that challenged the dominance of the germ theory.

Transitions

For the opening of the new phase, World War II is a convenient watershed. Leading epidemiologists perceived and prepared themselves to meet the *epidemic transition* (Omran 1971) from the predominance of infectious disease to that of chronic disease.

This transition was the outcome of objective and continuing change in society at large. The Great Depression of the 1930s did not for long interrupt a steady increment in economic productivity that led to higher living standards, better nutrition, and large demographic changes in the Western world. The population distribution shifted substantially toward older ages. The decline in infant and childhood mortality increased the numbers surviving to middle and old age; the decline in births reduced the numbers of the young. With these age changes came the now familiar increment in age-related disease such as heart disease, stroke, cancer, and senile dementia. Alongside the economic and demographic transformation, the scientific application of sanitation and vaccines, and of chemotherapy and antibiotics, undoubtedly contributed to the conquest of the more severe effects of infectious diseases.

While the infectious diseases declined among the forces of mortality, new and mysterious diseases, so-called "diseases of civilization," began to emerge. Middle-aged males in particular began to suffer and die from them at an unprecedented rate—after the turn of the twentieth century, from peptic ulcer, and in the 1920s and 1930s, from coronary heart disease and lung cancer (Morris 1957). Hypertension and mental disorders were also thought to be on the rise, although no good population data discounting Emil Durkheim's classic text on suicide (Durkheim 1897) were available to bear this out.

At the time of this epidemic transition, in medicine the prevailing explanatory paradigm for chronic disorder was that of *degenerative* disease. These diseases were seen in the main as intrinsic failures of the aging organism, brought to attention by increased survival and the rise in the aged population. Distinct from the orthodox medical view, however, a novel perception among epidemiologists entrained a line of research that was to establish a different paradigm for chronic disease. The view that chronic diseases had environmental and behavioral causes that were potentially preventable reflected the optimism and enthusiasm for reform that prevailed at the end of World War II.

After a long and uphill struggle, that paradigm is widely accepted, in medicine as well as in public health. At the same time and, we believe as a result, epidemiology has long transcended its narrowed and singular

role in elucidating the characteristics and effects of infection. It is clearly a needed discipline in the biomedical armamentarium. Epidemiology now serves as necessary to the solution of an array of pressing biomedical problems besides epidemic infections. These problems, once mainly those of the causes of chronic disease, have extended into such fields as the establishment of population norms of all kinds, the control of fertility, and population genetics. The problems have in common only their potential origin in the environment, and the need for population studies to determine their character and origin.

Contemporary discoveries—the role of rubella in congenital anomalies; of cigarette smoking in lung cancer, emphysema, and heart disease; of oxygen in retrolental fibroplasia in the premature baby; of sex hormones in cancer of the breast and of the endometrium; and of viruses in Burkitt's lymphoma, hepatitis B, and of prions in kuru—arise from and are encompassed by a theoretic viewpoint sharply different from that of the early years of the late nineteenth and early twentieth centuries. The exemplary research of the earlier period dealt almost entirely with infectious disease. Many infectious agents and modes of transmission were discovered. They were typically bacteria in acute disorders such as the major epidemic diseases, parasites in the tropical diseases and later, viruses.

Yet most contemporary teaching of epidemiology, in searching for relevance in the past, turned to quite another set of exemplars. What now seemed most meaningful were the discoveries that repeated ingestion of lead in small quantities caused the "Devonshire colic," that soot was the carcinogen responsible for scrotal cancer in chimney sweeps, that citrus fruit could prevent or cure scurvy (a not infrequently fatal disorder), that measles was a transmissible if unidentified infection with a regular incubation period, that cholera, too, was caused by an unidentified infectious agent carried by sewage-contaminated water, that silent infection by an unknown organism underlay the mysterious pattern of the distribution of poliomyelitis, or that pellagra signaled a nutritional deficiency.

This reordering of relevance is an unequivocal indication that epidemiologic thinking was guided by a new paradigm. The most salient methods were perceived to be those that can aid in the discovery of multiple environmental agents of unknown character or of obscure operation. Historical examples that conform with the new criteria of relevance are selected, both in the following text and in others, to indicate the origins of various methods and the priority of those who invented them. An illusion of continuous linear development is created by this selective process. In truth the resurrection of old priorities masks a leap from one paradigm to another (Kuhn 1970).

Epidemiology in Academia

The structure and dimensions of epidemiology in Europe and in the United States, first developed along somewhat different lines. During the interim between the World Wars I and II in Europe, the first academic base for what was described as Social Medicine was established in Zagreb, Croatia, in 1931. Professor Andrije Stampar was designated chairman. Quite soon thereafter, recruited by the Health Organization of the League of Nations, he was sent off for a peripatetic spell of study and lecturing in several countries. In North America as a guest of the Rockefeller Foundation, Stampar consulted with the US Committee on the Costs of Medical Care. Dispatched to China, in the aftermath of devastating floods, he advised on the control of the ensuing epidemics of infectious disease and on the reconstruction of destroyed villages.

When early in World War II Hitler's forces occupied the former Yugoslavia, the Gestapo (the brutal Nazi political police) arrested Stampar. He was held captive as an internee until in April 1945, the War in Europe ended. Thereupon he resumed his professional position, and in 1946 the newly founded World Health Organization (WHO), at its first International Health Conference in New York in 1946, elected Stampar head of the organization.

In many medical schools and universities, new departments of social and preventive medicine began to replace the academic Departments of Public Health first established by Sir John Simon in Britain in the late nineteenth century. In these new departments, epidemiology as a discipline developed. In Britain, immediately after the end of World War II in 1945, there was a surge of social concern to repair the wartime devastation, both physical and social. At Oxford in 1948, the first Chair of Social Medicine in England was filled by Sir John Ryle, a leading physician and gastroenterologist who then occupied the chair of Regius Professor of Medicine at Oxford, with Dr. Alice Stewart his deputy and Reader.

Several other universities across England followed and founded such new departments. Their brief was to promote an academic base that would give new emphasis to broad social concerns in public health. Such notable figures as Francis Crew in Edinburgh, C.F. Brockington in Manchester, A.C. Stephenson in Belfast, W.J.E. Jessop in Dublin and, later, J.N. Morris in London, T. McKeown in Birmingham and John Pemberton in Sheffield, were appointed to chair newly formed university departments of Social and/or Preventive Medicine. The occupants of these chairs assumed a central role in the development of epidemiology as an essential discipline in post-war public health. Especially in Britain during this period, with the seeming recession of acute epidemic infection and the rising rates of chronic disease, the study of infectious disease receded as a central concern.

This new academic base of social medicine in Britain, expanded by the several new chairs, led to the establishment in 1956 of the academic and scientifically oriented Society of Social and Preventive Medicine. This was an initiative of John Pemberton, the newly appointed Chair of Social Medicine at Sheffield University (Porter 1997; Pemberton 2002). Parallel and related developments in social psychiatry were greatly influenced by Aubrey Lewis (Susser 1968), an Australian and major intellect later knighted for his distinguished leadership of the Institute of Psychiatry at the venerable Maudsley Hospital for Psychiatric Disorders. In the course of revolutionizing psychiatric services in England, under the new parliamentary Mental Health Act of 1957, Lewis trained and guided a young team of such gifted social psychiatric researchers as Michael Shepherd, Peter Sainsbury, John Wing, Lorna Wing, Morris Carstairs, George Brown, and others; their brief was the epidemiological study of mental disorders and social factors in their causes, course, treatment, and care.

During the postwar period in Britain, perhaps more than anyone else J.N. Morris ("Jerry" to all) forged a central place in epidemiology for the role of social determinants in health and disease in his brief but seminal, idiosyncratic text (Morris 1957). Early in his career, in collaboration with the distinguished demographer R.M. Titmuss, he foresaw many now contemporary themes. Thus he wrote: "Society largely determines health; ill health is not a personal misfortune due often to personal inadequacy, but a social misfortune due more commonly to social mismanagement and social failure." (Morris 1957). In 1948 the Medical Research Council appointed him as founding Director of the Social Medicine Research Unit. In his many studies he promoted a broad vision of social medicine with its core in epidemiology and social science.

Morris generated ingenious and economical approaches to studies of the pressing health problems of the time. One such project, a comprehensive study of mortality in children in England and Wales, showed that death rates closely mirrored the gradient of social class among their mothers. Later, Morris focused especially on the role of chronic diseases underlying mortality in postwar Britain. Most notable was the seemingly new and growing phenomenon of epidemic coronary heart disease, in which he first established the protective effect of physical exercise. In this ingenious study, he contrasted the sedentary bus drivers with their much more active conductors who dashed up and down the double-decker buses collecting fares. In their work, Morris and his colleagues gave attention both to social factors and individual behavior as interrelated causes. For that purpose, he recruited a small research team that was appropriately and even necessarily multidisciplinary. It included not only such physician epidemiologists as Stuart Morrison and John Last, but also statisticians such as Austin Heady, and a

psychiatric social worker, namely Matilda Goldberg (her book explored the effects of psychological factors and family relations in the then major epidemic of peptic ulcer). Successive editions of *Uses of Epidemiology* (Morris 1957, 1964, 1975), a lively text written in his characteristic staccato style, describe much of his work and thinking.

Contemporaneously, John Brotherston at the London School of Hygiene and Tropical Medicine (later Sir John) recruited to his department no less than three sociologists: Margot Jefferies, Ann Cartwright, and Fred Martin. (In passing, one should not forget that in any discipline, British departments did not compare in scale with those in the United States.) In social medicine, several academics concerned themselves mainly with the study of health systems and the provision of medical care in such particular settings as general practice and mental health services. C. Fraser Brockington provided valuable background on the "history and development of public health as a discipline in Britain" (Brockington 1965). Then leaders of that first post-war generation like "Jerry" Morris, Richard Doll, John Brotherston, Thomas McKeown, John Pemberton, Archibald Cochrane, and others, sought understanding of societal, social, and biological factors integrated in a comprehensive search for the determinants of chronic disease and other health outcomes. Rather than viewing disease etiology solely in microbiological, or behavioral, or environmental, or social terms, these postwar investigators extended epidemiological approaches to examine the full spectrum of health determinants. During this period, Susser and Watson's text, *Sociology in Medicine* (Susser and Watson 1962), accomplished a melding of epidemiological and sociological thought in addressing the significant supporting role that the social sciences can play in the study of health and disease.

During this period, also, Thomas McKeown and the talented team he recruited at Birmingham University made some seminal contributions to contemporary social epidemiology. Among them were Brian MacMahon (later Chair of Epidemiology at Harvard), George Knox, and Alwyn Smith (later Chair of the Department of Social and Preventive Medicine at Manchester University). The pioneering work of McKeown and his faculty included the innovative device designed by George Knox to identify localized clusters that might point to a specific cause.

Thomas McKeown and Reginald Record essayed descriptions of shifting and contrasting national patterns of morbidity and mortality over the past few centuries. These analyses made some bold assumptions about declines in mortality in earlier centuries and the underlying causes, and some critics view their basic assumptions as open to question. As with many epidemiologists in Britain and Ireland during this period, this thinking regarding the determinants of population health was expansive enough to incorporate individual behavior as well as socioeconomic conditions. In his book *The*

Modern Rise of Population (McKeown 1976), McKeown built on a series of his prior analyses to reach some sweeping generalizations about population growth and likely mortality trends in England and Wales from the eighteenth century onward. These sought to explain the increases of population during and after the Industrial Revolution. McKeown interpreted his provocative analysis as indicating that the dramatic reductions in mortality during this period occurred independently of specific public health interventions. Instead, he hazarded the alternative assumption that the rise in population was owed to a decline in adverse socioeconomic conditions and better nutrition. Although this analysis has been strongly criticized, this is not to lessen the lasting impact of his book on contemporary social epidemiology (Susser 1993; Colgrove 2002; Szreter 2002; McKeown 2005) (also see Chapter 6).

The Systematization of the Literature Developments in the United States

From the mid-1950s in the United States, there was a shift in emphasis at leading universities (among others at Yale, C.E.A. Winslow; at Columbia, Gurney Clark; at Harvard, Brian MacMahon; at Johns Hopkins, Abraham Lilienfeld, and at Berkeley, Reuel Stallones). This shift in emphasis led to the search for individual risk factors as "exposures," which predisposed to chronic conditions. The ideas initiating this phase of epidemiology are well reflected in a 1952 review of the discipline by John E. Gordon, Professor of Preventive Medicine and Epidemiology at the Harvard School of Public Health (Gordon 1952). In pointing to past triumphs, Gordon had to rely largely on research in infectious diseases such as yellow fever, mumps, influenza, infectious hepatitis, and pneumonia. In looking to the future, by contrast, he pointed to studies of chronic disease. He espoused a holistic view of disease, framed by Pettenkofer's classic parameters of agent, host, and environment.

Aware of the necessary analytic distinction between levels of organization, Gordon, like others before him, specified the population group as the appropriate unit of study for epidemiological analysis. The viewpoint he adopted in favor of a concept of multiple causality also marks a break from the pursuit of the specific agent of disease initiated by the germ theory. The logic of this concept of multiple causes, one may infer, led Gordon to require the better definition of variables, especially the hypothetical causal variable. For this purpose, he turned to the social sciences and genetics in the hope that they would help in advancing techniques for measuring the social components of causality.

The need for quantification at once follows from a population model of complex causes, and Gordon called for building on quantitative foundations

in the manner of his predecessors Wade Hampton Frost, Joseph Goldberger, and Edgar Sydenstricker (Gordon 1952; Terris 1979). It was already clear to Gordon that to undertake such an enterprise, epidemiologic research in the future would have to be a team affair that would draw on several disciplines (social, biologic, and statistical), and on the skills of different specialists (especially clinicians and laboratory workers).

In the little space Gordon devoted to research design, he differentiated the broad categories of descriptive and analytic epidemiology, and indeed he traced the historical evolution of epidemiology from the descriptive studies of John Graunt in 1662 through the comparative historical geographic and ecologic studies of epidemics of the nineteenth century (Finke 1792 [see Barrett 2000]; Hirsch 1883) to the analytic studies of Peter Ludwig Panum on measles in the Faroe Islands in 1847 (see Panum 1939), and of John Snow on cholera in London in 1855 (Snow 1855). He confined experimental epidemiology to the conduct of large-scale animal experiments, in a usage that persisted through much of the 1950s, and that had taken origin with the work on cycles of epidemic transmission by Amoss and Webster in the United States and that of Greenwood and Wilson in Britain (Webster 1922; Amoss 1922a, 1922b; Greenwood et al. 1927).

More revealing to the modern eye are Gordon's omissions in discussing methods. In matters of design, with the exception of one new development, he referred only to the "field survey." He implies that this design is cross-sectional and elicits prevalence. The one new development noted by Gordon he termed, for want of a name, the "field review." He used this term, under which he described the prospective cohort study of incident disease, to describe the freshly launched Framingham Study of coronary heart disease.

In fact, one searches in vain in the standard texts of the decade after World War II for a description of the basic features of either cohort or case-control designs (Leavell and Clark 1953; Maxcy 1956). Not until the work of MacMahon et al. in 1960 (MacMahon et al. 1960) does a systematic treatment of study design appear. As a result, this pathbreaking book became a founding text for modern epidemiology.

Among antecedents, in Britain the sixth edition of Austin Bradford Hill's limpid work (Hill 1955) incorporated clinical trials but not other designs. In 1957, Taylor and Knowelden's book (Taylor and Knowelden 1958) touched only briefly on the structure of designs. In 1959, however, Richard Doll's chapter in a multiauthor text gave an exposition (Doll 1959) on usable designs. "Uses of Epidemiology," the 1957 text by Jerry Morris (Morris 1957) was seminal, but it taught from application, not method.

A measure that the discipline had reached its majority is that since the early 1970s in the United States, epidemiologists have been able to pursue doctoral studies without a criterion that required a primary degree beyond

the baccalaureate in some other field. At Columbia University School of Public Health in 1972, we were able to establish a rigorous PhD program in epidemiology (the first in the U.S.), although not without academics in the Graduate Faculties raising questions about the academic status of the discipline. The chief criterion for the adequacy of such an academic program is that it produces trained professionals capable of pursuing and publishing research autonomously. Such an epidemiologist is competent in statistics but not a statistician; has a grasp of concrete biomedical reality without being a clinician responsible for the medical care of individuals, and is able to comprehend the basic elements of society and social structure without being a sociologist or anthropologist. Ideally, such a newly minted epidemiologist should be capable of doing better than any one of these in pursuing the defined public health task of addressing the health of populations (Hollingshead and Redlich 1958; Lilienfeld 1958; Reid D.D. 1960; Susser and Watson 1962; Rose and Blackburn 1968; Susser 1971; Morabia 2004; Susser et al. 2006).

This new breed had to await on the synthesis of a solid disciplinary foundation. It was approaching two decades after World War II before epidemiologic writings exemplified by the text of MacMahon, Pugh, and Ipsen (1960), assumed the definable structure of a coherent whole and provided the needed foundation of the discipline. Thus, the coming of age of epidemiology as a discipline signifies a body of research and literature that is sufficient for academic studies in depth. Methods adequate to the various tasks of epidemiology have been devised and clarified. The choice among them depends not only on intuition, natural ability, and a basic knowledge of disease, but also on specialized training in a particular field. In consequence, the autodidact (as represented by the authors themselves) is a dying species. Great epidemiologists of the past, the creators of the discipline, had little special training. The mark of their greatness is precisely that they were the innovators, and of necessity, self-taught.

References

Ackerknecht EH (1948). Hygiene in France, 1815–1848. *Bull Hist Med* 22:117–155.

Amoss HL (1922a). Experimental epidemiology. I. An artificially induced epidemic of mouse typhoid. *J Exp Med* 36:25–43.

Amoss HL (1922b). Experimental epidemiology. II. Effect of the Addition of Healthy Mice to a Population Suffering From Mouse Typhoid. *J Exp Med* 36:45–69.

Barrett FA (2000). Finke's 1792 map of human diseases: The first world disease map? *Soc Sci Med* 50:915–921.

Brockington F (1965). *Public Health in the Nineteenth Century.* Edinburgh: E. & S. Livingstone.

Cassel J (1964). Social science theory as a source of hypotheses in epidemiological research. *Am J Public Health* 54:1482–1488.

Colgrove J (2002). The McKeown thesis: A historical controversy and its enduring influence. *Am J Public Health* 92:725–729.

Doll, R (1959). Retrospective and prospective studies. In Witts, ed. *Medical Surveys and Clinical Trials*. London: Oxford University Press, pp. 64–90.

Dubos RJ (1959). *Mirage of Health: Utopias, Progress, and Biological Change*. New York: Harper.

Durkheim E (1897). *Le Suicide*. New York: Free Press, trans. by J. Spalding and G. Simpson, 1987.

Fleck L (1935). *Genesis and Development of a Scientific Fact*. Chicago: University of Chicago Press (see 1979 edition with comment by Kuhn).

Frost WH, Maxcy KF (1941). *Papers of Wade Hampton Frost, M.D. a Contribution to Epidemiological Method*. New York: The Commonwealth fund. Reprinted in 1977, New York: Arno Press.

Gordon, JE (1952). The twentieth century—yesterday, today and tomorrow. In Winslow, Smillie, Doull, et al., eds. *The History of American Epidemiology*. St. Louis, MO: The CV Mosby Co., pp. 114–167.

Greenwood M, Newbold EM, Topley WWC, Wilson G. (1927). On the mechanisms by which protection against infectious disease is acquired in "Natural" epidemics. *J Hyg* 25:336–350.

Hill AB (1955). *Principles of Medical Statistics*. 6th ed. London: The Lancet.

Hirsch A (1883). *Handbook of Geographical and Historical Pathology*. London: New Sydenham Society.

Hollingshead A, Redlich F (1958). *Social Class and Mental Illness*. New York: John Wiley and Sons, Inc.

Kuhn TS (1970). *The Structure of Scientific Revolutions*. 2nd ed. Chicago: University of Chicago Press.

Lambert R (1963). *Sir John Simon, 1816–1904 and English Social Administration*. London: MacGibbon & Kee.

Leavell HR, Clark EG (1953). *Textbook of Preventive Medicine*. New York: McGraw-Hill.

Lilienfeld AM (1958). The epidemiologic method in cancer research. *J Chronic Dis* 8:649–654.

MacMahon B, Pugh T, Ipsen J (1960). *Epidemiologic Methods*. Boston: Little Brown.

Maxcy KF (1956). *Preventive Medicine and Public Health*. 8th ed. New York: Appleton-Century-Crofts.

McKeown T (2005). Medical issues in historical demography. *Int J Epidemiol* 34:515–520.

McKeown T (1976). *The Modern Rise of Population*. London: Edward Arnold.

Morabia A, ed. (2004). *History of Epidemiologic Methods and Concepts*. Boston: Birkhauser.

Morris JN (1957). *Uses of Epidemiology*. Edinburgh: Livingstone. 2nd ed (1964); 3rd ed. New York: Churchill Livingstone (1975).

Omran AR (1971). The epidemiologic transition. A theory of the epidemiology of population change. *Milbank Mem Fund Q* 49:509–538.

Panum PL (1939). Observations made during the epidemic of measles on the Faroe Islands in the year 1846. *Medical Classics* 3:802–866 [Hatcher AS, transl.].

Pemberton J (2002). Origins and early history of the society for social medicine in the UK and Ireland. *J Epidemiol Community Health* 56:342–346.

Porter D (1997). *Social Medicine and Medical Sociology in the Twentieth Century*. 43. Amsterdam: Rodopi.

Reid DD (1960). *Epidemiological Methods in the Study of Mental Disorders.* 2nd ed. Geneva: World Health Organization.

Rose GA, Blackburn H (1968). Cardiovascular survey methods. *Monogr Ser World Health Organ* 56:1–188.

Simon J (1890). *English Sanitary Institutions, Reviewed in Their Course of Development and in Some of Their Political and Social Relations.* 2nd ed. London: Cassell.

Smith T (1934). *Parasitism and Disease.* Princeton: Princeton University Press. Reprinted in 1963: New York: Hafner Pub. Co.

Snow J (1855). *On the Mode of Communication of Cholera.* 2d ed. London: J. Churchill.

Susser E, Schwartz S, Morabia A, Bromet EJ (2006). *Psychiatric Epidemiology: Searching for the Causes of Mental Disorders.* New York: Oxford University Press.

Susser M (1968). *Community Psychiatry: Epidemiology and Social Themes.* New York: Random House.

Susser M (1971). The public health and social change: Implications for professional education in public health in the United States. *Int J Health Serv* 1:60–70.

Susser M (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* New York: Oxford University Press.

Susser M (1993). Health as a human right: An epidemiologist's perspective on the public health. *Am J Public Health* 83:418–426.

Susser M. (1985) Epidemiology in the United States after World War II: the evolution of technique. *Epidemiol Rev* 7:147–177.

Susser MW, Watson W (1962). *Sociology in Medicine.* London: Oxford University Press.

Szreter S (2002). Rethinking McKeown: the relationship between public health and social change. *Am J Public Health* 92:722–725.

Taylor I, Knowelden J (1958). *Principles of Epidemiology.* Boston: Little, Brown.

Terris M (1979). The epidemiologic tradition. The Wade Hampton Frost Lecture. *Public Health Rep* 94:203–209.

Webster LT (1922). Experiments on normal and immune mice with a bacillus of mouse typhoid. *J Exp Med* 36:71–96.

16

The Expanded Epidemiology Team: Social Scientists and Statisticians Join Epidemiologists in Social Surveys

At the time of the epidemic transition of the twentieth century described above, the cross-sectional field survey was the readily available tool for epidemiologists seeking to describe the rising frequency and significance of chronic disease. As early as 1916 (as discussed in Chapter 14), Edgar Sydenstricker had used this approach. He joined Joseph Goldberger (Goldberger and Sydenstricker 1918) in the landmark study he initiated of living conditions of seven cotton mill communities at high risk of pellagra, in which they identified a nutritional deficiency of Vitamin B as the cause. In December 1921, Sydenstricker, by then Chief of the Office of Statistical Investigations of the US Public Health Service, began a continuous series of morbidity surveys in Hagerstown, Maryland (Sydenstricker 1926). A variety of morbidity surveys had continued in Hagerstown through the 1930s. From 1938 through 1943 a series on chronic disease, planned jointly by Sydenstricker and Wade Hampton Frost, was in the end executed (by Jean Downes and Selwyn Collins) in Baltimore and completed only after Sydenstricker and Frost, the two originators of such studies in the United States, both had died in their fifties (Sydenstricker 1974). In the early 1960s, field research at Hagerstown was effectively resumed by the late George Comstock at the Johns Hopkins Bloomberg School of Public Health and

is currently maintained as the base for an "epidemiologic laboratory" for community studies (Comstock 1970).

In present day epidemiology the search for causes is the nub of the discipline, and in themselves field surveys do not have a prominent place in the pursuit of causation. In this light, one may seem to give surveys undeserved attention in what follows. One reason is paradoxical. It is precisely the manifest weaknesses of the cross-sectional population survey approach to causal research that engendered much of the longitudinal perspective in the thinking and the methods found today in the kit of the epidemiologist. The field survey can be seen as a primitive progenitor of other current research design. First, in the hands of Austen Bradford Hill and of R.A. Fisher in Britain, one may see surveys retooled in the form of the case-control design. Second, in the form of the longitudinal or panel study elaborated by repeated cross-sectional surveys of a given population over time, the field survey might reasonably be seen as among the ancestors of the cohort design.

In the post–World War II phase, developments of the field survey in the United States were intensified in two directions. The first led to the institutionalization and maintenance of morbidity surveys on a national level. The second led to a focus on community surveys of mental disorder. A national morbidity survey of 750,000 families conducted by the Public Health Service with aid from the Works Progress Administration in 1935–1936 was in the direct lineage of descent from Sydenstricker's earlier surveys. In the liberal Truman era after World War II, beginning in January 1950 the Commission on Chronic Illness set out to determine the frequencies to be expected for the various diagnostic categories of chronic disease in urban and rural areas. Medical epidemiologist Morton Levin at the Johns Hopkins School of Public Health was the first director of the Commission, which in four volumes (Commission on Chronic Illness 1956) reported what was learned and provided a foundation for the continuing National Health Survey, conducted at intervals since 1956.

The National Health Survey comprises three elements: repeated sample surveys of self-reported illness; medical examinations among smaller samples of defined age groups; and the collection of data from medical care facilities. The morbidity surveys, which rely on self-reports of diagnosed illness and of disability (measured respectively by bed-days and dysfunction in daily living), were later extended to include data on nutrition, fertility, and other topics. Although these morbidity surveys are descriptive and do not in themselves constitute etiologic research, they provide a valuable resource for the generation of hypotheses, for the estimation of sample sizes, and for testing hypotheses by secondary analysis.

A second line of development of the morbidity survey—in community studies of mental disorder—is probably its main residual use in research. After World War II a signal innovation in the epidemiology of mental disorders was the attempt to escape the confines of medical diagnosis. Previous surveys of mental disorder, which in the United States had begun with Edward Jarvis in nineteenth-century New England, had virtually all relied on counting medically diagnosed cases, most usually those found in medical care facilities.

Two basic problems had emerged in the conduct of such studies. A minority of persons with mental disorders entered medical care and, even for that minority in care, diagnosis was as yet unreliable between different psychiatrists and different institutions. Psychiatric epidemiologists therefore set about the task of forging instruments for sample surveys that would yield reliable measures of the prevalence of the slippery entities of mental dysfunction. The chosen instrument for breaking out of the constraints of treated disorder was the structured interview questionnaire—mainly in the form of symptom inventories—applied in community prevalence studies (Susser 1968; Dohrenwend and Dohrenwend 1981).

The task was a long one and indeed is not yet completed. Two major surveys were conducted by independent investigators in the 1950s, namely, the Midtown Manhattan Study devised by Leo Srole and his co-workers (Srole et al. 1962) and the Stirling County studies in rural and small town Nova Scotia devised by Alexander and Dorothea Leighton (Leighton et al. 1963). The results of these studies were loaded with provocation and uncertainty. The reported frequencies of mental impairment or psychiatric dysfunction were met with open disbelief because of their magnitude. Validation was not attainable because the scores obtained on the symptom inventories could not be readily extrapolated to the distinctly defined mental disorders diagnosed by psychiatrists. Psychiatric researchers, aware of the problem, tried at least to specify "caseness," a condition a psychiatrist would judge to be in need of treatment (Leighton et al. 1963).

To bridge the gap and define the difference between symptom inventories and psychiatric diagnosis, the attempt was made to create, and to test in the field, psychiatric screening and diagnostic instruments that relate to the standard diagnostic manuals used by psychiatrists. In other areas of research, also, many scales of function and disability have been invented, which are specialized, standardized, and to some extent validated. Some such scales relate to nonpsychiatric disorders, as for example respiratory disease, and others relate to life stress and to life situations.

The labor devoted to enhancing the strength of cross-sectional surveys has not expunged the weaknesses in the reliability, validity, and significance of their measures. While the weaknesses remain, however, the efforts to correct

them have encouraged a rigor in method that has been to the advantage of epidemiology in general. Two areas in particular have benefited. One is the design and analysis of surveys; the other is logical inference. These benefits can best be understood in context.

To take first survey design and analysis, most epidemiologists are intensely concerned with the gathering of data and with ensuring that it meets minimum standards of quality. The methodologists who have treated systematically such topics as the construction of scales, interviewing and response bias, reliability, validity, and measurement error have been drawn, oddly, not from epidemiology but from psychology and sociology. These advances, acquired from the so called "soft" social and psychologic sciences (Susser 1968; Susser 1973), were the fruit of a period of the dominance in those fields of functionalism and positivism, a period during which they strove to emulate, in the study of society, the rigor and methods of the physical sciences.

At the same time, medicine was compelled to recognize that it was not self-sufficient. As in many medical research areas, research became ever more specialized and reliance on nonmedical scientists was growing. In the major surveys of the Commission on Chronic Illness (Commission on Chronic Illness 1956), or in Midtown Manhattan (Srole et al. 1962), or in Nova Scotia (Leighton et al. 1963), sociologists and psychologists were recruited to augment physicians and psychiatrists as the technicians and the architects of design. In the process, a new breed of social and psychosocial epidemiologists emerged in the United States; only a scattering of such socially oriented medical epidemiologists as Sydney Cobb and John Cassel had preceded them. In short, these various studies opened the way for the substantial development of psychosocial epidemiology in the United States. They attracted a host of new researchers and generated a stock of new and better methods.

This accession of social scientists was important also for future recruitment to graduate studies in epidemiology. In the postwar ambience of World War II, after the first flush of enthusiasm for social issues had subsided, beside the quick returns and the goldfields of medicine, the long labors and meager material rewards of epidemiology and public health seemed the less enticing and talented recruits were difficult to find. The new discipline of social epidemiologists helped fill the gap. Their presence and productivity demonstrated the potential of epigones untrained in medicine for carrying the work of epidemiology forward.

In any event, the texts on survey method of this period are written by sociologists, psychologists, and statisticians. These texts include such special topics as sampling theory (Deming 1950; Cochran 1953) and interviewing (Hyman 1954) as well as general texts on survey research (Goode

and Hatt 1952; Hyman 1955). Much stemmed from the statistical writings of Ronald Fisher and Austin Bradford Hill in England (Fisher 1925; Fisher 1935; Hill 1937). It was 1959 before a text—a British one—was devoted to health and medical surveys (Witts 1959). A single-author text on the subject, from Israel, did not appear until 1974 (Abramson 1974). Nor are we aware of even a simple epidemiologic text on interviewing before 1975 (Bennett and Richie 1975). Again, this text was British. Perhaps traditional epidemiologists dwelt little on these matters because, on the face of it, the data they sought treated objective manifestations with serious implications for affected persons. Did such data seem more grounded in material reality, and therefore less in need of being challenged for truth and tested for error, than such tenebrous things as subjective opinions and the recesses of states of mind?

Again, perhaps it was because epidemiologists dealt with categoric data produced by the diagnostic process that they made perhaps their chief methodological contribution to cross-sectional studies. This was to recognize and measure the unreliability of diagnostic observations. Thus, their concerns were with intra- and interobserver error; with misclassification and its consequences for analysis and inference; and with instrument reliability, as in the comparison of self-reported illness with medically assigned diagnoses.

One field of study, the early detection of disease by diagnostic screening, epidemiology made its own. The field arose in response to the demands of the public health strategy of secondary prevention through early detection and treatment of disease or of nascent risk. Its theory was built around the fourfold table, which remains at the center of modern epidemiologic thinking. Screening theory, too, has since been taken up in clinical pathology and other diagnostic fields, as well as to help improve the efficiency of surveys. Thus, the technical operations of screening include measures of sensitivity, specificity, predictive value, conditional probabilities, and lead time bias. (Thorner and Remein 1961; Vecchio 1966; Hutchison and Shapiro 1968; Galen and Gambino 1975). The synthesis of screening theory—in a comprehensive model, which reconciles these notions with those of reliability and validity on the one hand, and with those of decision theory on the other—is still to be achieved.

Now let us consider what benefits field surveys have conferred on logical inference and interpretation. These procedures and their problems, first taken up systematically by philosophers of science and statisticians, were an early preoccupation of survey researchers, a preoccupation that is unsurprising in face of the ambiguity and instability of social research data. In epidemiology, the concern with logical inference perhaps reflects most clearly the shift of focus from the discovery of specific infectious agents toward broad etiologic studies of chronic disease. Early attempts to apply the then dominant means

of the cross-sectional field survey to the ends of chronic disease studies made the problems of inference at once obvious.

Yet again, the relevant literature on logical inference from observational data was not indigenous to epidemiology, but was as much also the product of the social and psychological sciences. A substantial body of literature flowed from the practical demands for surveys made on American social scientists during World War II. Information was wanted about people at home, in the armed forces, and in occupied territories (Merton and Lazarsfeld 1950). These social scientists had to contend with two problems of inference exaggerated by cross-sectional data. Such data, that is, lack benchmarks for time order among variables, and they run a high risk of confounding causal relationships among a proliferation of variables.

Statisticians helped cut paths through the multivariate undergrowth, as did those of other disciplines wrestling with similar problems. This concern with confounding and inference led to the early adoption by the social sciences and psychology of the methods of multiple variable analysis invented by Fisher (Fisher 1925; Fisher 1935) and others. Social surveys often used continuous variables that lent themselves easily to multiple regression analysis. Once computers made such analyses accessible to anyone who could command a packaged program, they also lent themselves, alas, to mindless abuse. Multivariate analyses that neglect the known and unknown relationships and causal pathways among variables are as likely to obscure as to reveal reality.

The advance of chaos was stemmed somewhat with the reintroduction of path analysis to the biologic world. The method was invented around the time of World War I by Sewall Wright (Wright 1978), a founding father of population genetics. Path analysis, however, was little used until taken up by the new discipline of econometrics, and then by social scientists, finally to return in the 1970s to population genetics and at last to the biomedical world. Path analytic models impose a degree of order because the analyst must commit himself or herself to explicit causal models. Hence, assumptions are more readily brought to the surface and made visible. Further developments in this direction take account of reciprocal pathways (structural equations) and longitudinal relationships through time ("time lagging" and "causal analysis").

In epidemiology, the first treatment of the problems of logical inference within a multiple cause framework appears in the text of MacMahon et al. on epidemiological methods in 1960 (MacMahon, Pugh and Ipsen 1960). These problems have since been extended and elaborated in other texts, beginning with a concept of variables located in a structure of causal sequences, and with a set of strategies to establish what the sequences among variables might be and to screen out extraneous variables (Susser 1973).

A substantial literature now exists on confounding, interaction, and various estimates of risk. Epidemiologists began to make general use of multiple variable analysis only after the invention of the binary regression method (the use of dummy variables by which discontinuous variables could be entered into multiple regression equations) (Cohen and Cohen 1983) and logistic regression (Suits 1957; Feldstein 1966; Walker and Duncan 1967). We have since seen increasing use and new borrowings of multivariate analytic approaches.

In summary, the attention given by social scientists to the inherent weaknesses of field surveys led to the strengthening of design techniques and the creation of methods for constructing and testing instruments and scales and for dealing with measurement problems. It also led to awareness of the demands of logical inference and the dangers of confounding and, hence, to the early adoption of multiple variable analysis. Because of their training and skills in these matters, statisticians and social scientists have become a growing segment of the epidemiologic profession.

References

Abramson JH (1974). *Survey Methods in Community Medicine.* Edinburgh: Churchill Livingstone.

Bennett AE, Richie PA (1975). *Medical Interviewing.* London: Oxford University Press.

Cochran WG (1953). *Sampling Techniques.* New York: Wiley.

Cohen J, Cohen P (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: L. Erlbaum Associates.

Commission on Chronic Illness (1956). *Chronic Illness in the United States.* Cambridge: Published for the Commonwealth Fund by Harvard University Press.

Comstock GW (1970). The unofficial census as a basic tool for epidemiologic observations in Washington County, Maryland. In Kessler II, Levin ML, eds. *The Community as an Epidemiologic Laboratory: A Casebook of Community Studies.* Baltimore: Johns Hopkins Press.

Deming WE (1950). *Some Theory of Sampling.* New York: Wiley.

Dohrenwend BP, Dohrenwend BS (1981). Socioenvironmental factors, stress, and psychopathology. *Am J Community Psychol.* 9:128–164.

Feldstein MS (1966). Binary variable multiple regression method of analysing factors affecting peri-natal mortality and other outcomes of pregnancy. *Journal of the Royal Statistical Society Series A-General* 129:61–73.

Fisher RA (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fisher RA (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Galen RS, Gambino SR (1975). *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses.* New York: Wiley.

Goldberger J, Sydenstricker E (1918). A Study of the diet of nonpellagrous and of pellagrous households - in textile mill communities in South Carolina in 1916. *J Am Med Assoc* 71:944–949.

Goode WJ, Hatt PK (1952). *Methods in Social Research.* New York: McGraw-Hill.
Hill AB (1937). *Principles of Medical Statistics*. London: The Lancet.
Hutchison GB, Shapiro S (1968). Lead time gained by diagnostic screening for breast cancer. *J Natl Cancer Inst* 41:665–681.
Hyman HH (1955). *Survey Design and Analysis: Principles, Cases, and Procedures.* Glencoe, IL: Free Press.
Hyman H (1954). *Interviewing in Social Research*. Chicago: University of Chicago Press.
Leighton DC, Harding JS, Macklin DB, MacMillan AM, Leighton AH (1963). *The Character of Danger: Psychiatric Symptoms in Selected Communities. Vol III: Stirling County Study of Psychiatric Disorder & Sociocultural Environment.* New York: Basic Books.
MacMahon B, Pugh T, Ipsen J (1960). *Epidemiologic Methods.* Boston: Little Brown.
Merton RK, Lazarsfeld PF (1950). *Continuities in Social Research: Studies in the Scope and Method of "The American Soldier."* Glencoe, IL: Free Press.
Srole L, Langer TS, Michael ST, Kirkpatrick P, Opler M, Rennie TA (1962). *Mental Health in the Metropolis.* New York: Harper & Row.
Suits DB (1957). Use of dummy variables in regression equations. *J Am Stat Assoc* 52:548–551.
Susser M (1968). *Community Psychiatry: Epidemiologic and Social Themes.* New York: Random House.
Susser M (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* New York: Oxford University Press.
Sydenstricker E (1926). A Study of Illness in a General Population Group. Hagerstown Morbidity Studies, No. 1: the method of study and general results. *Public Health Rep* 41:2069–2088.
Sydenstricker E (1974). *The Challenge of Facts: Selected Public Health Papers of Edgar Sydenstricker.* New York: Prodist, for the Milbank Memorial Fund.
Thorner RM, Remein QR (1961). *Principles and Procedures in the Evaluation of Screening for Diseases.* Washington: U.S. Govt. Print. Off.
Vecchio TJ (1966). Predictive value of a single diagnostic test in unselected populations. *N Engl J Med* 274:1171–1173.
Walker SH, Duncan DB (1967). Estimation of probability of an event as a function of several independent variables. *Biometrika* 54:167–&.
Witts LJ (1959). *Medical Surveys and Clinical Trials: Some Methods and Applications of Group Research in Medicine.* London: Oxford University Press.
Wright S (1978). The application of path analysis to etiology. In Morton and Chung, eds. *Genetic Epidemiology.* New York: Academic Press.

17

The Arsenal of Observational Methods in Epidemiology: Classical Designs, the Fourfold Table, Cohort and Case-Control Studies

Introducing the Fourfold Table

The essential comparisons made in epidemiological studies, whether case-control, cohort, or experimental, can all be represented in the fourfold table below (Table 17.1). This table also serves in field studies to test the precision of measures, either of exposure or of outcome (Chapter 20), and in public health applications in screening for affected individuals (as in Chapter 21).

The cells of the table represent, in a given population, the frequencies with which both a manifestation or effect (the dependent variable) and an experience or hypothetical cause (the independent variable) occur either separately or together. In other words, the table represents the degree to which the two variables are associated with each other. Whatever the study design, the cells of the table show all four possible combinations of any two characteristics in a study population.

In a case-control study, individuals are first identified, and classified into case and control groups according to the presence or absence of the manifestation under study. In the table, the group of cases is represented by a+c, the group of controls by b+d. The frequencies of the independent variable in each of the groups are then compared: in other words, a/a+c is compared with b/b+d. Higher frequency of exposure in the case group;

Table 17.1

Independent Variable	Dependent Variable		
	Present	Absent	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

in other words, a/a+c > b/b+d points an association between the two study variables.

In cohort and experimental studies, individuals are first identified and classified according to whether they were subject to the suspected determinant under study. In the table, the group exposed is represented by a+b, the unexposed control group by c+d. The frequencies of the dependent variable among both groups are then compared. That is, a/a+b is compared with c/c+d. Here a higher frequency of cases in the exposed group points to association of the two variables. Within this seemingly simple framework, we shall find ample room for complication, obfuscation, and misinterpretation to lead us into error. Laborious evolution distilled this elegant generalization about the structure of the variables in studies seeking association. A review of that evolution may help to steer one clear of the bogs of false inference.

Cohort Studies

Cohort field studies share with cohort analyses a longitudinal and forward-looking analytic perspective. They ordinarily entail the recruiting of a population at some given entry point. This entry might depend on some particular characteristics common to all recruits to the study, as in cohorts defined at entry by age, for instance either at birth, or at school, or at military recruitment, or at entry to homes for the aged and the like. Actual or recorded observation begins at recruitment, and appropriate means are devised to specify and measure the particular manifestations or outcomes selected as the objectives for study.

Thus the researcher begins by selecting a population identified at a common point of entry with known characteristics appropriate to the aims of the study, and excluding those already affected by the conditions to be studied as outcomes. This population is followed over a time period conveniently but not necessarily determined at the outset. Antecedent estimates are helpful in ensuring the feasibility of the study: for instance, estimates of the anticipated frequency of the events of interest, of the numbers of subjects needed and

hence the ultimate size of the cohort, and of the time that must pass to yield the number of events sufficient to provide the analytic power to yield statistically viable results. All outcomes need not be predetermined at the outset, however, as would be wise whenever an analysis with a defined hypothesis is undertaken. As the history of the cohort unfolds over time, it has the great advantage that, as new opportunities for study are recognized, outcomes additional to those initially selected for study may be added.

Cohorts can thus be assembled to great advantage for the study of the causes and outcomes—hypothesized or known—of preselected diseases or even the subsequent advent of unanticipated outcomes. Such populations are closed in that, once defined, there are no additions to the population and all losses should be accounted for. In the course of follow-up, researchers then observe and measure change in those attributes and emergent events of interest. The researcher is better armed if, in addition to specifying the postulated causes of a given outcome at the outset, measures are included of factors extraneous to those causes that might raise or lower risk from the suspect factors. These factors thereby help to control possibly false estimates that either confound or mask the true risk of the disorder under study.

Among notable early cohort studies, one can comfortably accord priority on the grounds of originality and importance to the national study of British Births. Launched in 1946 and directed by J.W.B. Douglas, the study followed outcomes of interest among all live-born, legitimate single births in England, Scotland, and Wales in the single week of March 3 to 9 (Douglas and Blomfield 1958; Douglas 1964). The stimuli for undertaking so ambitious a study, initiated in a war-ravaged country with depleted resources, stemmed from two quite separable pressures: first, the new postwar Labour Government led by Clement Attlee, having passed legislation to establish the National Health Service still now extant, needed to know the extent of the commitments that had been undertaken; second, demographers raised alarms about a sharp decline in births observed over the previous decade and earlier, and hence by the prospect of a steady and worrying decline in the productive population that was bound to follow. One assumption was that the expense associated with maternity was discouraging would-be parents from adding to their families (Titmuss and Titmuss 1942).

A total of 13,687 births were identified eight weeks after delivery, although follow-up excluded the illegitimate and multiple births among them. Also, to render the study more economical, the final study population was arranged in weighted social strata to yield a balanced sample of 5,362 births drawn randomly from each stratum. All births to the smaller social classes of agricultural laborers and nonmanual workers were included, but only one in four of all the remaining larger classes. The existence of the cohorts also enabled and encouraged the subsequent development of the life course

studies mentioned in Chapter 19 (Wadsworth 1991). Health visitors, trained public health nurses maintained by law by all Local Health Authorities, were assigned to report on the home visits they made routinely in the course of their duties. These reports especially emphasized the quality, adequacy, and costs to families of the services provided prenatally at delivery, and subsequently through childhood.

Birth weight was the main biological variable recorded at birth, and low birth weight especially served as a key variable in many subsequent analyses of immediate and later outcomes. The follow-up continued periodically over a half-century and, at successive intervals through three decades, the director James Douglas himself reported what one might describe as snapshots of cross-sectional observations of the physical and intellectual development and social circumstance of the study population.

Data on early childhood were later supplemented by the reports of the school nurses serving each health district. In addition, at 7 and 15 years of age, reports were obtained from the public health doctors assigned by each local health authority to the schools, and also from teachers. When the cohort population reached its twenties, further reports were solicited by mail.

In the early years of life, the questions emphasized breast feeding, sphincter control, illnesses and hospitalizations, and separations. At later ages, reports described school and measured mental performance, height and weight. The cohort has been followed into adulthood and middle age, with reports of employment and work experience, menopause, and retirement.

As the years passed, the opportunities for posing scientific and social questions have yielded a rich reward. Indeed, the assembled accounts would serve well as a kind of social history of the second half of the twentieth century in Britain. The psychoanalytic observations on child separation of John Bowlby, so influential in childrearing in the 1950s and 1960s, are examined; the effects on the children of steeply rising divorce rates among their parents observed; the disappointing failure to equalize or even to reduce substantially the inequalities of educational opportunities and performance across the social classes acknowledged and reported; the impact of social class and parental interest in school performance measured; and the strengths and weaknesses of the National Health Service frankly reported. The existence of this and subsequent national birth cohorts provided a major resource for future life course studies.

All cohort studies are prospective in the sense that they follow the experience of a defined population forward through time. The position in time of the investigator need not be antecedent to that experience, however; the reconstruction of past experience of defined populations has proved to be a productive research procedure, as described in Chapter 18 under the topic of "natural experiments."

Among physicians in the United States, the Framingham Study stands as the exemplar of epidemiology (Oppenheimer 2005). It is the epitome of successful epidemiologic research, productive in the first instance of insights into cardiovascular disorders and their causes, and of such research applications as cardiovascular risk factor scores. Among epidemiologists mature or in training in the United States, the Framingham Study thus became the prototype and model of the cohort study. For these reasons, it deserves close examination. As best one can determine from accounts of the history of this study—like all histories, these are not in perfect concord (Gordon and Kannel 1970; Dawber 1980; Vandenbroucke 1985)—the Framingham Study set out to develop case-finding procedures for heart disease in healthy volunteers. The seeming intention was to establish *incidence rates* in a general population.

Joseph Mountin, assistant to the Surgeon General of the United States Public Health Service, recognized the public health interest in the rising epidemic of heart disease. With Mountin's support, David Rutstein, Chairman of the Department of Preventive Medicine at Harvard Medical School, joined with the Massachusetts Commissioner of Health (Vlado Getting) to institute a follow-up study among the 28,000 people of Framingham, a town outside Boston. Mountin initially appointed Gilcin Meadors as the epidemiologist who would execute the study. Initiated in October 1947, medical examinations of volunteers began a year later (October 1948). Well into its second year (July 1, 1949), the study was taken over by the newly founded National Heart Institute. Surprisingly, only then was the specific aim of the study formulated, namely to determine "factors influencing the development of heart disease" (Gordon and Kannel 1970).

A new, more representative sampling of 30- to 59-year-old persons was then planned, and an estimate made of required sample size. One account hints that in actuality the estimate then made, rather than determining *de novo* a required sample size, weighed feasibility of the study within a fixed limit of 6,000 subjects. In any event, about two thirds of the selected sample agreed to participate. To meet the shortfall thus incurred, the numbers were supplemented by drawing on unsolicited volunteers available because to turn away anyone desiring an examination in a local community was judged unwise. Some 15 to 18 months after the first cycle of examinations had begun Thomas Dawber, a new principal investigator assigned to succeed Meadors, put into operation a revised protocol for biennial examinations of the recruited cohort that would continue over a 20-year period. Only about nine months later than that, at the end of the third year of initiation of the study (October 1950), did a laboratory become available to carry out cholesterol and other determinations. Unsurprisingly, smoking histories were introduced at a still later stage, and only after Doll and Hill in England

had published their paper on cigarette smoking as a cause of lung cancer in 1950.

As the study progressed over the years, interview schedules were added and refined to improve the quality and content of measurement and to test additional risk factors such as physical activity, diet, and life stress. The proportion of refusals was not negligible, however, and the study could not yield the sound representative incidence rates intended. One suspects also that the distinction between, on the one hand, *cumulative incidence* in a cohort of constant composition and, on the other hand, *period incidence rates* in a general population with continuous accessions and losses had not generally been appreciated, despite Farr's clarification in 1859 (Vandenbroucke 1985; Farr 1859). Perhaps it was with this realization that the goal evolved of establishing among populations the individual risks attached to specific factors. This implicit rationale for longitudinal studies, now routinely assumed, seems not to have been clear and certainly was not explicit at the outset.

When it came to analysis, tactics had to be devised to deal with the potential variability of repeated measurements over time. Where there were unreliability and fluctuation in observed values, how much was the result of regression to the mean? and what were the appropriate base lines for starting measures against which to assess change (Truett and Sorlie 1971). These were major difficulties inherent in longitudinal observations, and the continuity of the project gave some opportunity for solving them. Two of the investigators, in describing the history of this study, wrote that it "might be reasonable to consider measurement problems as one of the justifications for longitudinal studies" (Gordon and Kannel 1970).

Finally, as the study continued, it became apparent that there was no good way of stopping it. (At a meeting of the Royal Society of Medicine in London held to celebrate his 65th anniversary, Sir Austin Bradford Hill read a paper on the 10th anniversary of the national cohort study of the effects of smoking among British doctors. He opened with the remark that there was nothing to be done with a longitudinal study but to pass it on to someone else in one's will.) Even now, no agreed stopping rules for a longitudinal study exist. After 20 years, the National Heart Institute decided that the Framingham Study had realized its original goals, and funding was withdrawn. Determined supporters of the study obtained funding elsewhere. However, after an interval of a few years the study was once again funded by the National Institutes of Health. The project thereafter was devoted to stroke, cerebrovascular disease, the various aspects of aging and, also, to the health status of offspring of the original participants.

No true and completed precursors of the Framingham longitudinal study seem to exist in epidemiology. Conceivably, a number of longitudinal growth and development studies started in the 1920s and 1930s, and also

J.M. Tanner's "auxological epidemiology," as he described his work on growth in the late 1960s could be considered in this light (Boas and Wissler 1906; Shuttleworth 1937; Tanner 1979). Well before that, Sir James MacKenzie, founder of modern cardiology, had sowed an idea that can be followed to its application in Framingham. Around 1920, Mackenzie had left laurels and fame in London to set up, in his native Scotland, the St. Andrew's Institute of Clinical Research. The institute was to be devoted, through the agency of general practitioners in Dundee, to the longitudinal study of the natural history of disease in the community. His aim was to discover the physiologic expressions of disease, from the "beginnings of illness" and then throughout its course. Although Mackenzie died in 1925 before his plan could be fully realized, the American cardiologist Paul Dudley White, who had studied with McKenzie in London, carried the germ of his idea back to Boston. In Boston, White had in turn become a leading cardiologist of his time in the United States and was an early and influential protagonist of the Framingham idea. The modern techniques of the longitudinal study, however, were still to be invented and its challenges discovered.

In 1956, Richard Doll and his mentor A. Bradford Hill reported the first result of a cohort study of smoking and its effects on health in British doctors (all on the national register of general practitioners were invited to volunteer). The method for this remarkably economical study—a hallmark of their work—was precisely laid out from the outset. Follow-up continued over several decades. Beginning with intense attacks on the case-control studies of 1950, tobacco companies had stirred bitter contention by fair means or foul. This 1956 study yielded definitive scientific evidence of the ill-effects of tobacco.

One might argue that in principle the cohort design is an analogue of two preexisting types of design that are both forward-looking and involve the lapse of time. These are the population studies of incidence and, as mentioned above, the repeated cross-sectional survey. A pointer to the connection of the cohort design with studies of incidence can be found in the original intention of Framingham prospectively to determine, not the individual risks of heart disease, but its incidence in the population at large. The connection with the repeated cross-sectional survey rests on the fact that such surveys afford a means of determining prospectively changes in state. In sociology, when applied to a fixed population sample, such surveys constitute the so-called "panel design." Except for the objectives of such study, the design does not differ from cohort studies that yield both cumulative incidence and probabilities according to individual characteristics.

Yet, the crystallization of the cohort method must be attributed primarily to the requirements of studies of chronic disease. This impetus is already seen in Wade Hampton Frost's 1933 paper on the risks of familial contact

in tuberculosis (Frost 1933). Tuberculosis was a bridging condition for epidemiologists making the transition from the dominant study of acute infections to chronic disease. With respect to the bridging role of tuberculosis, we shall have occasion to cite later the development of follow-up studies of reconstructed cohorts. The first application of cohort analysis to disease, to our knowledge, was Andvord's analysis of tuberculosis mortality in England and Wales, Denmark, Norway, and Sweden (Andvord 1930; 1932). Later, the innovative multicenter clinical trial evaluating streptomycin in the treatment of tuberculosis, and the early prophylactic trials of Bacillus Calmette-Guerin (BCG), also reinforce this notion of a bridging function. The Framingham Study, in turn, has been a stimulus of three lines of development, namely, (1) other cohort studies of cardiovascular disease; (2) other types of cohort studies; and (3) long-term community studies.

In the first group, an array of cohort studies of cardiovascular disease in communities and in large organizations tested, validated, and extended the Framingham results. The earlier community cohorts include Tecumseh, Michigan (Epstein et al. 1965); Evans County, Georgia (Cassel 1986); and Honolulu (Worth and Kagan 1970). Cohorts of employees were followed in Minneapolis (Keys et al. 1963); Albany, New York (Doyle et al. 1957); Los Angeles (Chapman et al. 1957); and Chicago (Stamler et al. 1960; Paul et al. 1962).

In the second group, a number of other and less famous longitudinal studies, most often studies of particular exposures and multiple diseases, evolved in parallel with the Framingham Study. This strategy, that is, to follow multiple outcomes in a cohort exposed to a specified risk, is in fact better fitted to the logical uses of longitudinal observation than the initial Framingham approach. Although the Framingham research team considered multiple risk factors, they strictly limited the number and kinds of outcomes.

It is to the testing and cross-fertilization among all these various studies that we owe our understanding of the longitudinal method. Early on in the United States, among the better known, such studies are the continuing follow-up of the Japanese survivors of atomic bombs (Beebe et al. 1962); several follow-up studies of United States veterans (Beebe 1960) and notably Harold Dorn's study of smoking and mortality (Kahn 1966); the American Cancer Society study of the effects of smoking (Hammond and Horn 1954); and the Collaborative Perinatal Project on antenatal factors in postnatal child development (Niswander and Gordon 1972).

The third group—long-term community studies—took up the thread spun on the one side from Sir James MacKenzie in Britain (Frowde et al. 1922) and on the other from Sydenstricker's repeated cross-sectional sample surveys of Hagerstown, Maryland (Sydenstricker 1926). From these evolved the idea of the community as epidemiologic laboratory (Crew 1948; Kessler

and Levin 1970). Such parochial studies have in common the continuing observation through time of a defined community. Their object may be accomplished by follow-up as in Tecumseh, Michigan (Francis and Epstein 1965); The Human Population Laboratory in Oakland, California (Berkman and Breslow 1983); or by repeated cross-sectional survey or census as in Washington Heights, New York City, (Gell and Elinson 1969); Washington County, Maryland (Comstock et al. 1970); or by registers of a given disorder as with mental disorder in Monroe County, New York (Gardner et al. 1963); or by linked record systems as in Rochester, Minnesota (Kurland and Molgaard 1981).

Judging by published results, most community laboratories have produced less than was hoped for, and perhaps less than their potential. Aside from the stopping rule problem, smaller communities are not likely to yield adequate numbers either for hypothesis testing or for the study of secular trends unless the disorders under study are very common. In general, small communities serve best for monitoring trends of incidence and prevalence and as a frame for drawing cohort or case-control study samples.

The use of a special community census for sampling in multiple studies is well illustrated by a continuous stream of research reports stemming from the Washington County Epidemiological Laboratory in Maryland. These address such diverse topics as the effects of the urban factor in lung function, of widowhood on mortality in general, and of soft water on cardiovascular disease (Comstock et al. 1970; Stebbings 1971; Helsing et al. 1981; Comstock 1971). In this way, community laboratories can also serve to test a well-grounded hypothesis. For instance, a seminal test of the effect of social support on mortality was conducted by the Oakland Human Population Laboratory (Berkman and Breslow 1983). The long-standing record-linked system of the Mayo Clinic has proved highly productive in each of these areas (Kurland and Molgaaard 1981). This system has provided the best available frequency and prognostic data on epilepsy, stroke, and many other disorders. It has also been quickly deployed to retest many hypotheses, such as the effects of reserpine on breast cancer, and of diethylstilbestrol on clear-cell adenocarcinoma of the female genital tract.

By the 10-year mark of the Framingham Study, enough had been learned about cohort studies for a full exposition to appear in MacMahon et al.'s 1960 text (MacMahon et al. 1960). In 1959, however, Richard Doll had recommended limiting the use of cohort studies primarily to supporting or refuting a well-formulated and precise hypothesis (Doll 1959). Noted in favor of the method was the lesser opportunity for bias—since knowledge of the outcome could not affect the estimates of exposure—and the simplicity of the relative risk estimates. One may add the advantages of the design for the study of rare and extreme or high-dose exposures in definable groups, as

demonstrated among World War I veterans exposed to mustard gas in World War 2 (Beebe 1960), with the atomic bombs in Hiroshima and Nagasaki (Beebe 1962), or with asbestos dust in exposed New York construction workers (Selikoff et al. 1968).

If the exposure or risk is common, advantage may lie in the detailed description of the factor that can be gathered concurrently with the field observations, as in the Framingham cohort study. In that study, moreover, repeated follow-up allowed both risk factor and outcome data to be extended and refined. This flexibility overcomes the handicap of long-term cohort studies with a once-only starting point, which are tied to initial hypotheses. By the time the results of such studies are known, new knowledge may have made their hypotheses anachronistic.

Whatever the nature of the exposure, multiple and unknown outcomes of a given exposure are not easily discovered except by means of the cohort study. One advantage unique to the design is that only longitudinal observation can take account of selective loss and survival in the time interval between exposure and outcome. Where such losses can occur as a result of the exposure, they may cause serious confounding.

Thus, in the populations exposed to the atomic bombs of 1945, effects of prenatal exposure on measured intelligence were found only for exposures after the sixth week of gestation. Early loss of damaged fetuses probably suppressed the evidence of effects of earlier exposure. This attrition is indicated by a deficiency in the number of births among the cohorts exposed early in gestation (Miller and Blot 1972). Likewise, in a study of the effects of the Dutch famine of 1944–1945 on human development, a reduction in the fertility of the poorer social classes as a result of famine so distorted the social composition of birth cohorts as to give the false appearance that severe famine exposure early in pregnancy improved the mental performance of young adults (Stein et al. 1975).

Various economies have been devised to enhance the advantages of cohort studies. An historical cohort study makes use of data previously collected to define cohorts and outcomes, as Case et al. (1954) did in a study in Britain of bladder cancer in chemical workers. An American example was a study of the effect on mortality of occupational exposure of medical specialists to ionizing radiation (Seltser and Startwell 1963). A later example was the study referred to above of the Dutch famine of 1944 to 1945: for birth cohorts defined by exposure, outcomes were studied at birth, in terms of mortality into adulthood, as well as in young adult males at military induction (Stein et al. 1975).

Economies have also been achieved where the exposed cohorts could be constructed from already collected data, but where outcomes were determined at concurrent follow-up. Such an approach was pioneered more than

a century ago in a study of the mortality of patients after discharge from the Adirondack Cottage Sanitarium, reported in 1904 by Brown, a tuberculosis specialist, and Pope, an actuary then a patient in the sanitarium. The attempt ran foul of a methodological difficulty. Although the observation period began with the onset of symptoms, the cohort comprised only individuals who had survived to admission or discharge and could not account for selective bias by attrition up to that time (Brown and Pope 1904).

In 1933 Frost noted, with his usual unnerving concision, his application of a variant of this method to the familial spread of tuberculosis. Subsequent studies of secondarily constructed cohorts, such as those of the Atomic Bomb Casualty Commission and of the US veterans, were eminently successful. With data gathered at follow-up, at some extra cost this approach obtains the advantage, of meeting the need for precise measurements of outcome data instead of having to rely on what opportunity brings.

Almost in passing, Frost carried the method a step further in order to study the aggregation of tuberculosis in families. To do this, he reconstructed the mortality experience of family cohorts separately around cases of tuberculosis and around unaffected controls; he took account of age and duration of exposure by adjusting the denominators from a life table of person years experienced at each age. In genetic epidemiology this is precisely the method now commonly applied to compare risks of recurrence among relatives of affected probands and those of unaffected controls. That is to say, a historical cohort design is constructed around subjects selected in the manner of a case-control study.

The Framingham Study itself has left its mark. This is so much the case that among many physicians and referees for journals, the conviction persists that only prospective studies can yield sound epidemiologic knowledge. Yet today, were one to judge by the initial protocols referred to in various accounts of its history, the Framingham Study could not hope to be funded. Peer reviewers would demand a sound rationale for the study; structured hypotheses; an estimate of sample size likely to yield significant answers (derived from the levels of significance and power, anticipated relative risk, and frequency of the outcome in the unexposed population); a description of the mode of recruitment of subjects, including criteria for eligibility; details of how the data are to be collected; the type and quality of the data and measures of their reliability and validity; and an account of how the data are to be analyzed. In addition, Institutional Review Boards are required to review every questionnaire to be used in the study.

A matter for historical and philosophic reflection is whether the standards since created and righteously applied are not in some way lacking. For the Framingham Study was indeed one of two major intellectual levers in bringing about a shift in the chronic disease paradigm, the other being the British

Doctors study. These are undisputedly the foundation stones for current ideas about risk factors in general and the prevention of ischemic heart disease and lung cancer in particular.

Case-Control Studies

Case-control designs rest on observations made of each individual at a single point of time in the life course of the population under study (although case collection overall may continue over protracted periods). The study population comprises both cases with the given condition under study, and controls unaffected by that condition. In all other respects, the two groups are as similar as possible. Their histories of exposure to suspected potential causes provide the crucial comparative data that differentiate them from each other. A number of variants on this design have been developed, as with the *case-crossover* (Taylor and Knowelden 1957) and the *nested case-control* designs. This economical design, however, depends on histories elicited of such potentially causal past experiences as are hypothesized by the analyst. Thus the design circumvents the need for long-term prospective observation. Analysis collapses time past into the present moment. The same economy can also apply to other suspect factors that might masquerade as causes or otherwise mask them, provided only that they are specified beforehand.

Cohort studies and community studies are expensive. Case control studies are less so and have come to dominate American epidemiology. Not only can they examine specific hypotheses but they serve also as exploratory studies. In 1979 Philip Cole reported that in the prior 20 years (since the mid-1950s), case-control studies published in four leading journals had increased four- to sevenfold (Cole 1979).

We owe the term case-control to Philip Sartwell (Sartwell 1981). The term *case history* coined by MacMahon, Pugh and Ibsen (1960) in their landmark text, *Epidemiology: Principles and Methods* aimed to escape the ambiguities of *retrospective* as another term then in use to describe the design. Other descriptors of the same design, such as *case-referent* (Miettinen 1970) and *case-compeer* (Hulka et al. 1978) have fallen by the wayside.

In conformity with our observation on the effects of the epidemiologic transition on epidemiologic methods (Cole 1979), one may reasonably attribute the proliferation of case-control studies to an emergent predominance of chronic disease. The utility of this method for studying disease with a long latent interval between exposure and onset soon became evident. The backward-looking study design collapses the time interval between exposure and outcome. Thereby, it avoids the labor and time involved in awaiting the slow evolution characteristic of chronic disease. Equally to the point, and unlike

the cohort design, the case-control method, is well suited to the investigation of those small clusters of chronic disease that sporadically appear.

The design has proved efficacious, too, if within limits, for the study of epidemic outbreaks of acute disorders of unknown origin, as with such newly identified infectious disease syndromes as Legionnaire's disease (Fraser 1977) and toxic shock syndrome (Langmuir 1982). It has served no less well for chronic disease caused by infection with a long latent period. Such infections have assumed fresh significance since the discovery of the slow viruses of kuru and Creutzfeld-Jacob disease, the identification of hepatitis B virus as an agent in both serum jaundice and hepatic cancer, and the emergence of syndromes with an immunologic basis such as subacute sclerosing panencephalitis and acquired immunodeficiency syndrome. This turn of the wheel, toward the study of infections by means of research designs typically used for noninfectious chronic disease, points to an essential if seldom manifested unity of the epidemiology of infectious and noninfectious disease.

In principle, as mentioned above, the perspective of the case-control study can be seen as an analogue of the prevalence survey of morbidity at a given point in time, a development from it that vastly increased efficiency. Both approaches begin with the manifestation of morbidity and both proceed to ascertain the attributes of subjects of the sample surveyed so that those who exhibit morbidity can be compared with those who do not. Like the field survey, in execution the case-control design is cross-sectional but only to the extent that it elicits for comparison, both the current state and past history of the study and control subjects at the same point in time. The survey ordinarily elicits the current attributes and exposures of a given population from a cross-sectional perspective. The attention now routinely given to retrospect—so that past history is documented with maximum precision and the risk of confusing the time order of cause and effect is reduced—is a subsequent if natural evolution of the case-control approach. The link between the two designs is emphasized by difficulties in separating them. Studies that compare the current attributes of cases with those of a simultaneously observed population at large from which the cases are drawn are clearly cross-sectional field surveys. When the comparison of cases is with a sample of the population, however, ambiguity enters into classification. Somewhat arbitrarily, if the survey aims to establish the relationships of cases to antecedent factors in a retrospective manner, one might best assign it to the case-control category (Paneth et al. 2002).

The use of a cross-sectional approach that resembles the case-control design has been traced, in papers of William Guy, as far back as in mid-nineteenth century England (Lilienfeld and Lilienfeld 1979). A well-developed form of the method, used by J.E. Lane-Claypon in 1926, is judged by Philip

Cole as the first "true case-control study of modern type" (Cole 1979). That work was passed by with relatively little notice, however. The prototypes that initiated and led to the present-day upsurge in case-control studies might best be credited to those in 1950 on cigarette smoking and lung cancer.

Conflicting results about such an association of cancer with smoking prevailed in the literature up to 1950, although a pre–World War II case-control study by Muller (Muller 1929) in Germany strongly supported such an association. Some 20 years later, in 1950, Muller's work had scarcely been noted in the English literature when four successive reports of case-control studies reported an association (Schreck et al. 1950; Wynder and Graham 1950; Levin et al. 1950; Doll and Hill 1950). Among these (Schreck et al. 1950) acknowledged the stimulus of an earlier verbal presentation of results by Wynder and Graham, although none referred to Muller.[1] From the account of Morton Levin (personal communication, 1984), the fact that the papers on smoking and lung cancer of Wynder and Graham (Wynder and Graham 1950) and of Levin et al. (Levin et al. 1950) both appeared in the same issue of the *Journal of the American Medical Association* was due to a happy accident. In pre-1950 English literature the hypothesis of the association between smoking and lung cancer was not widely bruited in the literature. Indeed, the postulate was hitherto so little documented that the reluctant editor hesitated before publishing these papers. Levin subsequently convinced the editor of the importance of the two papers, however, which were then published in the same issue. The persuasiveness of the last of these, Doll and Hill's study (Doll and Hill 1950), was enhanced by the consistency of its results with those of its predecessors. In its own right, however, this British paper set a new standard for the case-control study. This work—together with Doll and Hill's national cohort study of British doctors that followed (Doll and Hill 1956)—provided a second major intellectual lever in shifting the causal paradigm of chronic disease from an intrinsic to an environmental model. The paper stands as a classic exemplar for the investigation of a given outcome and an array of exposures. To our knowledge, no previous research paper lays out the essentials of the case-control method with such understanding and meticulous care.

The authors consider a wide range of design problems. Their analysis deals with the comparability of cases and controls and the precision of matching (for age, sex, place of interview); selective recruitment and attrition; confounding (by social class); the validity of the outcome variable and ways of refining it; the quality of the hypothetic causal variable, including

[1] One may perhaps infer that the prewar literature of Nazi Germany was tainted, especially given Hitler's vicious aversion not only to smoking but to Jews, several of whom contributed to the design and execution of the postwar studies on cancer and smoking in the United States especially.

the reliability of the history of exposure, the persistence of recall, and interviewer and response bias; measures of the hypothetic causal variable, including refinements of dose (amount and duration) and of type (variations in kinds and methods of smoking).

In their interpretation, the authors raised alternative hypotheses and disposed of alternative explanations that could reside in faulty study design or execution. The hypothesis that smoking is causal is reviewed in the light of its consistency with other results and its coherence both with population data on the distribution of lung cancer and smoking and with animal experiments.

Finally, Doll and Hill (Doll and Hill 1950) set out statistics for the significance levels and strength of the observed associations. Noteworthy is their extreme caution when they make estimates of relative risks (not odds ratios). Thus, the authors lay out their assumptions in detail. Their caution extends to not actually stating the risks, although these are on display in a figure. This tentativeness in estimating risks underlines the importance of the contributions to analytic procedure of Cornfield (Cornfield 1951) and the developments that followed from his work. In this 1951 paper, Cornfield established today's everyday knowledge of the conditions under which a simple odds or cross-products ratio fairly represents relative risk.

Epidemiologists had to wait until 1959, however, before Mantel and Haenszel (Mantel and Haenszel 1958) described a procedure that lifted the case-control study into the multivariate age. The procedure, the basis for which had been laid by Cochran (Cochran 1954), permitted the weighted combination of chi-square differences in order to arrive at the odds ratios of stratified variables. Joseph Fleiss at Columbia University School of Public Health had the gift of clarifying and simplifying. In 1973, Fleiss drew from his lectures to construct his new statistics text around the analysis of rates and proportions; this was the first text to meet the particular needs of epidemiologists in analyzing the fourfold table.

One subsequent current of thought has sifted the assumptions for attaining comparability between cases and controls, especially matching. It is fair to attribute a good deal of development to the poignard of Joseph Berkson's skepticism in regard to the case-control method (Berkson 1946). In the United States also, Olle Miettinen was a leading theoretical contributor to our understanding of the utilities and inutilities of matching, as well as to problems of confounding and risk estimation (Miettinen 1970, 1976). We have learned that in most instances, excepting either known relationships or extreme confounding, the matching procedure is best reduced to a minimum. When comparison groups are matched on one or more variables, the matching removes the effects of those variables from further consideration. A text wholly devoted to the exegesis of case-control studies is now available, and

other texts are devoted largely to the multivariate analysis of such studies by logistic regression and log-linear methods (Schlesselman 1982; Breslow and Day 1980).

Furthermore, the versatility of case-control studies has been greatly enhanced by the advent of computer programs that both facilitate access to complicated methods and permit their ready use. Loglinear analysis and its offshoots address directly the use of multiple categoric variables, without at once running the analyst into the problem of diminishing numbers in each cell as the data are stretched thin across stratified contingency tables (Goodman 1964; Grizzle et al. 1969; Bishop et al. 1975). By means of mathematical modeling, statistical virtuosi can reduce complex associations adjusted for multiple variables to simple and elegant curves. As always, in interpreting such masterly simplifications, their assumptions must be closely watched.

The utility, the robustness, and also the hazards of the case-control design have gradually come to be generally appreciated. Thus, several tests on cohort data—including almost 50 years of the Framingham study—have shown that the case-control design yields essentially the same results. Although through the decades since the 1950s and 1960s, there has been a steady increment in the number of case-control studies, many fields remain yet unexplored by this technique. To take one example, at Johns Hopkins Lilienfeld and Pasamanick (Lilienfeld and Pasamnick 1954), tutored and advised by the statistician Jacob Yerushalmy reported a case-control study of childhood epilepsy in the city of Baltimore. Aside from a related replication by Henderson et al of such a study among Blacks in the same city, (Henderson et al. 1964), the findings remained untested by any case-control study of epilepsy over the next 25 years. (In that instance, the findings were in fact not replicated in cohort studies (Susser et al. 1985).) Essentially the same applies to the case-control study of cerebral palsy in New York State by Lilienfeld and Parkhurst in 1951 (Lilienfeld and Parkhurst 1951).

The use of the case-control design has been extended in various ways. One extension is to economize in the use of large scale longitudinal cohort data. Thus, in the "nested case control study," an outcome of interest selected from the cohort data comprises the cases. The recorded antecedent experience of the cases is compared with that of a sample of unaffected individuals drawn from the cohort to serve as controls. Another extension is designed for the monitoring of population hazards. Thus, the Boston Collaborative Drug Surveillance Program (Boston Collaborative Drug Surveillance Program 1973) set out to monitor any adverse effects of drug treatments by continuously sampling patients admitted to hospital wards and taking their histories. Initially, the monitoring related to current hospital experience, but it was soon expanded to relate to drug use over the months previous to the admission diagnosis. The matrix of data on drug exposure and diagnoses

was systematically scanned and screened for associations by computer. Associations were examined in detail for robustness and then followed up and tested in case-control studies. In an early example, the study demonstrated a relationship of thrombophlebitis to oral contraceptive use that was conditional on the blood group of those affected.

This approach has not fully conquered the problem, common in epidemiology, of assessing the statistical significance of associations of multiple factors with multiple end points. In general, in a particular instance, the best resort is to the logic of causal inference as the guide to judging statistical inference. Ultimately, only consistency on replication under varying conditions might fully reassure. A different approach to monitoring was used in a continuing case-control study of spontaneous abortions (Kline et al. 1977), a method also applied in Britain to monitor time clusters of congenital anomalies (Weatherall and Haskey 1976). The cumulative summation method (CUSUM) was adapted specifically to monitor chromosomal anomalies. The exploration of such clusters through special case-control studies, however, has not yet proved itself productive.

The case-control method, we may confidently assert, is now in full bloom. Refinements of analytic techniques to control confounding and also to achieve comparability cascade from the journals. The method is applied in evaluation research, such as the efficacy of screening and secondary prevention (Clarke and Anderson 1979), or of prophylaxis for primary prevention (Sargent and Merrell 1940; Rhoads and Mills 1985), and of medical care. Further extensions and uses are bound to appear.

References

Andvord KF (1930). Hvad kan vi lære ved å følge tuberkulosens gang fra generasjon til generasjon? *Norsk Magazin for Lægevidenskapen* 91:642–660. For recent translation, see: Andvord KF, Wijsmuller G, Blomberg B (2002). What can we learn by following the development of tuberculosis from one generation to another? 1930. *Int J Tuberc Lung Dis* 6:562–568.

Andvord KF (1932). continued studies of tuberculosis considered as a generation illness. *Int J Epidemiol* [reprint and translation of 1932 lecture; epub Aug 2008].

Beebe GW (1960). Lung cancer in World War I veterans: possible relation to mustard gas injury and the 1918 influenza epidemic. *NNCI* 25:1231–1232.

Beebe GW, Ishida M, Jablon S (1962). Studies of the mortality of the A-bomb survivors: panel study of the mortality in the medical sample. *Radiat Res* 16:253–280.

Bennett AE, Richie PA (1975). *Medical Interviewing.* London: Oxford University Press.

Berkman L, Breslow L (1983). *Health and Ways of Living the Alameda County Study.* New York: Oxford University Press.

Berkson J (1946). Limitations of the application of fourfold table analysis of hospital data. *Biometrics* 2:47–53.

Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Boas F, Wissler C (1979). Statistics of growth. Report of US government on education for 1904. Washington, DC, 1906:25–132. Cited by Tanner JM. A concise history of growth studies from Boston to Boas. In: Falkner F, Tanner JM, eds. *Human Growth*. Vol 3. New York: Plenum Press, 515–593.

Boston Collaborative Drug Surveillance Program (1973). Oral contraceptive and venous thromboembolic disease, surgically confirmed gall bladder disease, and breast tumours. *Lancet* 1:1399–1404.

Breslow NE, Day NE (1980). *Statistical Methods in Cancer Research. Vol 1. The Analysis of Case Control Studies*. Geneva: International Agency for Research on Cancer, WHO.

Brockington CF (1965). *Public Health in the Nineteenth Century*. Edinburgh: Livingstone.

Brown L, Pope EG (1904). The post discharge mortality among patients of the Adirondack Cottage Sanitarium. *Am Med*. 8:879–882. Cited by Comstock GW (1985). Early studies of tuberculosis. In: *Selection, follow-up and analysis in prospective studies: a workshop*. Washington, DC. National Cancer Institute Monograph no. 67, 5:23–27.

Campbell DT, Stanley JC (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Case RAM, Hosker M, McDonald DG, Pearson JT. (1954). Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dye stuffs in the British chemical industry. *Br J Ind Med* 11:75–104.

Cassel J (1964). Social science theory as a source of hypotheses in epidemiological research. *Am J Public Health* 54:1482–1488.

Cassel JC, ed. (1971). Evans county cardiovascular and cerebrovascular epidemiologic study. *Arch Intern Med* 128:883–986.

Chapman JM, Goerke LS, Dixon W, Loveland DB, Phillips E. (1957). The clinical status of a population group in Los Angeles under observation for two to three years. *Am J Public Health* 47(suppl):33–42.

Clarke EA, Anderson TW (1979). Does screening by "pap" smears help prevent cervical cancer? A case-control study. *Lancet* 2:1–4.

Cochran WG (1953). *Sampling Techniques*. New York: John Wiley & Sons.

Cochran WG (1954). Some methods for strengthening the common x^2 test. *Biometrics* 10:417–451.

Cole P (1979). The evolving case-control study. *J Chronic Dis* 32:15–27.

Comstock GW (1971). Fatal arteriosclerotic heart disease, water hardness at home, and socioeconomic characteristics. *Am J Epidemiol* 94:1–10.

Comstock GW, Abbey H, Lundin FE, Jr (1970). The unofficial census as a basic tool for epidemiologic observations in Washington County, Maryland." In: Kessler II, Levin ML, eds. *The Community as an Epidemiologic Laboratory: A Casebook of Community Studies*. Baltimore: The Johns Hopkins University Press.

Cook TD, Campbell DT (1979). *Quasi-experimentation: Design and Analysis; Issues for Field Settings*. Chicago: Rand McNally.

Cornfield J (1951). A method for estimating comparative rates from clinical data: application to cancer of the lung, breast and cervix. *JNCL* 22:719–748.

Crew FAE (1948). *Measurements of the Public Health Essays on Social Medicine*. Edinburgh: Oliver and Boyd.

Davis DL, Bridboard K, Schneiderman M (1983). Cancer prevention: assessing causes, exposure, and recent trends in mortality for US males 1968–78. *Int J Health Serv* 13:337–372.

Dawber TR (1980). *The Framingham Study: The Epidemiology of Coronary Heart Disease*. Cambridge, MA: Harvard University Press.

Deming WE (1950). *Some Theories of Sampling*. New York: John Wiley & Sons.

Dohrenwend B, Dohrenwend BP (1981). Socio environmental factors, stress, and psychopathology. *Am J Community Psychol* 2:123–164.

Doll R (1959). Retrospective and prospective studies. In: Witts LJ, ed. *Medical Surveys and Clinical Trials*. London: Oxford University Press, pp. 64–90.

Doll R, Hill AB (1950). Smoking and carcinoma of the lung. *Br Med J* 2:740–748.

Doll R, Hill AB (1956). Long cancer and other causes of death in relation to smoking. *Br Med J* 2:1269–1275.

Douglas JWB, Blomfield JM (1958). *Children under Five*. London: Allen and Unwin.

Douglas JWB (1964). *The Home and the School a Study of Ability and Attainment in the Primary School*. London: MacGibbon & Kee.

Doyle JT, Heslin AS, Hilleboe HE, Formel PF, Korns RF. (1957). measuring the risk of coronary heart disease in adult population groups. III. Prospective study of degenerative cardiovascular disease in Albany—Report of three years experience. I. Ischemic heart disease. *Am J Public Health* 47(suppl):25–32.

Epstein FH, Ostrander LD, Johnson BC, Payne MW, Hayner NS, Keller JB, Francis T Jr. (1965). Epidemiological studies of cardiovascular diseases in a total community—Tecumseh, Michigan. *Ann Intern Med* 62:1170–1187.

Farr W (1859). On the construction of life tables, illustrated by a new life table of the "Healthy Districts of England" in the *Transactions of the Royal Society* 838–841. Cited in Vital Statistics: a memorial volume of selections from the reports and writings of William Farr. Humphreys NA, ed. (1885). London: Offices of the Sanitary Institute.

Fleiss J (1973). *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.

Francis T (1957). *Poliomyelitis Vaccine Center. Evaluation of 1954 Field Trial of Poliomyelitis Vaccine; Final Report*. Ann Arbor: University of Michigan.

Francis T Jr, Epstein FH (1965). Survey methods in general populations: studies of a total community. Tecumseh, Michigan. *Milbank Mem Fund* Q 43:333–342.

Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, et al (1977). Legionnaires' disease description of an epidemic of pneumonia. *New Engl J Med* 297:1189–1197.

Frost WH (1933). Risk of persons in familial contact with pulmonary tuberculosis. *Am J Public Health* 23:426–432. Republished in Maxcy KF, ed. (1941). *Papers of Wade Hampton Frost*. New York: The Commonwealth Fund.

Gardner EA, Miles HC, Bahn AK, Romano J (1963). All psychiatric experience in a community. A cumulative survey: report of the first year's experience. *Arch Gen Psychiatry* 9:369–378.

Gell C, Elinson J, eds. (1969). Washington Heights master sample survey. Milbank Mem Fund Q 47:Part II, no. 1.

Goodman, LA (1964). Simple methods for analyzing three-factor interaction in contingency tables. *J Am Stat Assoc* 59:319–352.

Gordon T, Kannel W (1970). The Framingham, Massachusetts study—20 years later. In: Kessler II, Levin ML, eds. *The Community as an Epidemiclogic Laboratory: A Casebook of Community Studies*. Baltimore: The Johns Hopkins University Press, pp. 123–148.

Grizzle JE, Starmer CF, Hutch GG (1969). Analysis of categorical data by linear models. *Biometrics* 25:489–504.

Hammond EC, Horn D (1954). The Relationship Between Human Smoking Habits and Death Rates. *JAMA* 155:1316–1328.

Helsing KJ, Szklo M, Comstock GW (1981). Factors associated with mortality after widowhood. *Am J Public Health* 71:802–809.

Henderson M, Goldstein H, Rogot E, Goldberg ID, Entwisle G (1964). Perinatal factors associated with epilepsy in Negro children. *Public Health Rep* 79:501–509.

Holland WW, Bennett AE, Cameron IR, Florey CDV, Leeder SR, Schilling RSF, Swan AV, Waller RE. (1979). Health effects of particulate pollution: reappraising the evidence. *Am J Epidemiol* 110:533–659.

Hulka BS, Hogue CJR, Greenberg BG (1978). Methodological issues in epidemiologic studies of endometrial cancer and exogenous estrogen. *Am J Epidemiol* 107:267–278.

Kahn HA (1966). The dorn study of smoking and mortality among US veterans: report on eight and one-half years of observation. In: Haenszel W, ed. *Epidemiological Study of Cancer and Other Chronic Diseases*. Washington, DC: US Department of Health, Education and Welfare. National Cancer Institute Monograph 19, 1–125.

Kessler II, Levin ML, eds. (1970). *The Community as an Epidemiologic Laboratory: A Casebook of Community Studies*. Baltimore: The Johns Hopkins University Press.

Keys A, Taylor HL, Blackburn H, Brozek J, Anderson JT, Simonson E. (1963). Coronary heart disease among Minnesota business and professional men followed fifteen years. *Circulation* 18:381–395.

Kline J, Stein Z, Strobino B, Susser M, Warburton D. (1977). Surveillance of spontaneous abortions power in environmental monitoring. *Am J Epidemiol* 106:345–350.

Kurland LT, Molgaard CA (1981). The patient record in epidemiology. *Sci Am* 245:54–63.

Langmuir AD (1982). Toxic shock syndrome: an epidemiologist's viewpoint. *J Infec Dis* 145:588–591.

Levin ML. Goldstein H, Gerhardt BR (1950). Cancer and tobacco smoking. *JAMA* 143:336–338.

Lilienfeld AM, Lilienfeld DE (1979). A century of case-control studies: progress? *J Chronic Dis* 32:5–13.

Lilienfeld AM, Parkhurst E (1951). A study of the association of factors of pregnancy and parturition with the development of cerebral palsy: a preliminary report. *Am J Hyg* 53:262–282.

Lilienfeld AM, Pasamanick B (1954). Association of maternal and fetal factors with the development of epilepsy. 1. Abnormalities in the prental and perinatal periods. *JAMA* 155:719–724.

Lilienfeld AM, Pederson E, Dowd JE (1967). *Cancer Epidemiology: Methods of Study*. Baltimore: The Johns Hopkins University Press.

MacMahon B, Pugh TF, Ipsen J (1960). *Epidemiological methods*. Boston: Little, Brown & Co.

Mantel N, Haenszel W (1958). Statistical aspects of data from retrospective studies of disease. *JNCL* 22:719–748.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 2:769–783.

Miettinen OS (1970). Matching and design efficiency inretrospective studies. *Am J Epidemiol* 91:111–118.

Miettinen OS (1976). Estimatibility and estimation in case-referent studies. *Am J Epidemiol* 103:226–235.

Miller RW, Blot WJ (1972). Small head size after in utero exposure to atomic radiation. *Lancet* 2:784–787.

Muller FH (1929). Tabakmissbrauch und lungencarcinom. *Z Krebsforch* 49:57–85.

Niswander K, Gordon M (1972). *The woman and their Pregnancies: The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. Philadelphia: WB Saunders.

Oppenheimer GM (2005). Becoming the Framingham study 1947–1950. *Amer J Pub Hlth* 95:602–610.

Paneth N, Susser E, Susser MW. The early history and development of the case-control study. Part I. Early evolution. *Social & Preventive Medicine* 2002; 47:5: 282–288.

Paneth N, Susser E, Susser MW. The early history and development of the case-control study. Part II. The case-control study since Lane-Claypon. *Social & Preventive Medicine* 2002; 47:6:359–365.

Paul O, Lepper MH, Phelan WH, Dupertuis GW, MacMillan A, McKean H, Park H. (1962). A longitudinal study of coronary heart disease. *Circulation* 28:20–32.

Reports of the St. Andrews Institute for Clinical Research, St. Andrews, Fife I (1922). London: Oxford Medical Publications, Henry Frowde and Hodder and Stoughton, 1922. II. London: Humphrey Milford, Oxford University Press, 1924.

Rhoads GG, Mills JL (1985). The role of the case-control study in evaluating health interventions: vitamin supplementation and neural tube defects. *Am J Epidemiol* 120:803–808.

Sargent CA, Merrell M (1940). Method of measuring effectiveness of preventive treatment in reducing morbidity. *Am J Public Health* 30:1431–1435.

Sartwell P (1981). Retrospective studies: a view for the clinician. *Ann Internal Med* 94:381–386.

Schlesselman JJ (1982). *Case-Control Studies: Design, Conduct, Analysis.* New York: Oxford University Press.

Schreck R, Baker LA, Ballard G, Dolgoff S (1950). Tobacco smoking as an etiological factor of cancer. *Cancer Res* 10:49–58.

Selikoff IJ. Hammond EC, Churg J (1968). Asbestos exposure, smoking and neoplasia. *JAMA* 204:106–112.

Seltser R, Sartwell PE (1963). The influence of occupational exposure to radiation on the mortality of American radiologists and other medical specialists. *Am J Epidemiol* 81:2–22.

Shuttleworth FK (1937). Sexual maturation and the physical growth of girls aged six to nineteen. *Monographs of the Society for Research in Child Development* 2(5):1–253.

Stamler J, Lindberg HA, Berkson DM,Shaffer A, Miller W, Poindexter A (1960). Prevalence and incidence of coronary heart disease in strata of the labor force of a Chicago industrial corporation. *J Chronic Dis* 11:405–420.

Stebbings JH (1971). Chronic disease among nonsmokers in Hagerstown, Maryland. IV. Effect of urban residence on pulmonary function values. *Environ Res* 4:283–304.

Stein Z, Susser M, Saenger G, Marolla F (1975). *Famine and Human Development: the Dutch Hunger Winter of 1944/45.* New York: Oxford University Press.

Susser M (1964). The uses of social science in medicine. *Lancet* 2:425–429.

Susser M (1968). *Community Psychiatry: Epidemiologic and Social Themes.* New York: Random House.

Susser M and Adelstein A (1975). Introduction to reprint of *Vital Statistics: a memorial volume of selections of the reports and writings of William Farr.* [Humphreys NA,

ed. London: Offices of the Sanitary Institute, 1885.] Reprinted by the Library of the New York Academy of Medicine. Metuchen, NJ: Scarecrow Press.

Susser M, Hauser WA, Kiely J, Paneth N, Stein Z (1985). Quantitative risks in perinatal brain disorders. In: Freeman J. ed. *Risk Factors in Perinatal Brain Disorders.* Washington, DC: NIH.

Susser MW, Watson W (1962). *Sociology in Medicine.* London: Oxford University Press.

Sydenstricker E (1926). A study of illness in a general population group. Hagerstown morbidity studies no. I: the method of study and general results. *Public Health Rep* 41:2069–2088.

Tanner JM (1979). A concise history of growth studies from Boston to Boas. In: Falkner F, Tanner JM, eds. *Human Growth*. Vol 3. New York: Plenum Press. pp. 515–593.

Taylor I, Knowelden J (1957). *Principles of Epidemiology.* Boston: Little, Brown & Co.

Titmuss, RM, Titmuss K (1942). *Parents Revolt: A Study of Declining Birth Rates in Acquisitive Societies.* London: Secker and Warburg.

Truett SH, Sorlie P (1971). Changes in successive measurements and the development of disease: the Framingham study. *J Chronic Dis* 24:349–361.

Vandenbroucke JP (1985). On the rediscovery of a distinction. *Am J Epidemiol* 121:627–628.

Wadsworth MEJ (1991). *The Imprint of Time: Childhood, History, and Adult Life.* Clarendon Press. Oxford.

Walker SH, Duncan DB (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54:167–179.

Weatherall JA, Haskey IC (1976). Surveillance of malformation. *Br Med Bull* 32:39–44.

Worth RM, Kagan A (1970). Ascertainment of men of Japanese ancestry in Hawaii through World War II Selective Service registration. *J Chronic Dis* 23:389–397.

Wright S (1978). The application of path analysis to etiology. In: Morton NE, Chung CS, eds. *Genetic Epidemiology.* New York: Academic Press.

Wynder EL, Graham EA (1950). Tobacco smoking as a possible etiological factor in bronchogenic carcinoma. *JAMA* 143:329–336.

18

Epidemiologic Experiments: Natural and Contrived

Although in the present day the experimental design is certainly one of the less common types of study undertaken by epidemiologists, it is far from the least important. The experiment is the design of choice where the investigator seeks internal validity in results. That requirement arises especially where a highly specific hypothesis can be tested, and where precision is wanted in the estimate of effects. These conditions are characteristically sought with medical treatments, especially medicaments. It is consistent that the earliest and most famous of early experimental trials—indeed, the first known such study, published in 1753 by the naval surgeon James Lind, demonstrated the superiority of lemons over several other treatments for scurvy (Lind 1753; Carpenter 1986).

The randomized trial, although descended from Ronald Fisher[1] by way of Austin Bradford Hill, is neither solely nor strictly epidemiologic ground. This ground, best defined by statisticians and widely exploited by epidemiologists, has overlapped with the interests of clinicians, pharmacologists, and others concerned with therapy. Such other disciplines have perhaps most often called upon epidemiologists and statisticians in multicenter trials, as with

[1] Early experiments in the field of psychology (Thorndike and Woodworth) and social science (especially Chapin) are described by Ann Oakley (1998).

the University Group Diabetes Program trial of oral hypoglycemic agents, controversial for its demonstration of none but adverse effects (Gilbert et al. 1975). Many elements contributing to the initiation of modern clinical trials derive from the multicenter trial of streptomycin in the treatment of tuberculosis, designed in the late 1940s by Austin Bradford Hill for the Medical Research Council in Britain (Medical Research Council 1948).

Epidemiologists are on their own firm ground, however, with prophylactic trials of vaccines, as for example in testing with the Bacillus Calmette-Guérin (BCG) vaccine for childhood tuberculosis (Palmer et al. 1958). In the United States, the most dramatic of such trials was the landmark poliomyelitis vaccine trial of 1954 (Francis 1957). The drama resulted from the epidemic threat to healthy children of crippling poliomyelitis—especially those from the higher social classes. Given their generally hygienic living conditions, they were the least likely to have acquired immunity and hence at highest risk of a devastating epidemic disease. The drama and fear aroused by the threat was heightened by the high level of publicity generated by the voluntary agency that funded the trial (the March of Dimes), and by concerns about the unknown risks of administering either one of the two vaccines developed in response to repeated epidemics, namely the killed Salk vaccine, or the live Sabin vaccine. The drama was heightened soon after by poliomyelitis cases caused by contaminated lots of the Cutter vaccine administered in the first large-scale trial of the vaccine.

This prophylactic trial involved nearly 2 million children. As initially designed, the trial was not randomized. It took much pressure from participating scientists to bring about a second randomized design, which was then superimposed on the initial plan and run in parallel with the first. After the trial, the severity of the criticism of the nonrandom segment was such that a large-scale nonrandomized trial of a vaccine against such severe infection is unlikely again to be undertaken without great caution. This caution does not hold for other preventive interventions as many examples show (one such instance in England being the prophylaxis of neural tube defects by preconception multivitamin supplements administered before conception, in which hospital boards blocked the randomized approach on allegedly ethical grounds (Smithells et al. 1980)).

A prophylactic trial can reduce the necessity for very large numbers if a population can be found with a high incidence of the condition to be prevented. A trial of hepatitis B vaccine used this high-risk strategy to bring one cycle of the saga of hepatitis virus—at the center of which is Baruch Blumberg's discovery of the Australia antigen—to a satisfying conclusion. Szmuness and colleagues (Szmuness et al. 1980) identified several groups at high risk of hepatitis B infection: people living under poor sanitary conditions in the Third World; renal dialysis patients; and multipartnered male

homosexuals. In a masterly study, Wolf Szmuness recruited and randomized some 1,083 active male homosexuals in New York City within one year. Eight months later, he had given definitive proof of the vaccine's efficacy.

The case for randomized trials to test drugs and vaccines seems unassailable, although there is room for trying out modifications to make things easier, for example sequential trials, factorial designs, lesser numbers of controls, and others. With regard to preventive experimental intervention Joseph Goldberger, as noted in Chapter 14, is a revered role model for American epidemiologists (Terris 1964). In the first decades of the twentieth century his conclusive demonstration of a dietary deficiency that was the cause of pellagra depended on dietary experiments that confirmed his remarkable observational studies. He carried out these studies among orphanage children prone to pellagra and its then sometimes fatal consequences. (An irony is that today, this work of a highly principled investigator might not meet the standards demanded by an Institutional Review Board [Susser et al. 1978].)

Some epidemiologists have remained cautious about experimental intervention. Perhaps they have been deterred on debatable ethical grounds, or perhaps they have been reluctant to sacrifice generalizability and representativeness for specificity and internal validity. After all, the differences in the strength of the inferences about causality drawn from different types of design are less qualitative than they are a matter of degree. Aside from the ethical questions always present, a number of other considerations weigh. In time, effort, and expense, the scale of a preventive trial must be thought of in much the same terms as a longitudinal study. As with cohort studies, these are best justified when there is a closely specified hypothesis, and when the same answer cannot readily be obtained by other means. It goes without saying that both the intervention and the outcome must be well defined and measurable.

Natural Experiments

The term "natural experiment" can be credited to John Snow in reporting his study of the epidemic of cholera, described in Chapter 8. John Last's *Dictionary of Epidemiology* provides this definition: "Epidemiologists have made good use of circumstances in which subsets of the population have different levels of exposure to a supposed causal factor, in a situation resembling an actual experiment where human subjects would be randomly allocated to groups" (Last and Abramson 1995). We noted in Chapter 8 that William Budd, like Snow, saw the usefulness of the "natural experiment." Hence, his observation during the great stench of London in 1858, when a murky haze clouded the view of the River Thames, mortality did not increase

concomitantly, thus countering the expectations of the adherents of miasma theory. Ignaz Semmelweis also, as discussed in Chapter 8, derived his ideas for a "contrived experiment" from his previous observations in the natural experimental situation of the leading clinic in Vienna. Thus he persuaded the obstetricians and nurses of his unit to ensure that ministrations to the obstetric patients were only undertaken after hands were carefully cleaned. Infection rates at once declined.

An experience from the work of our own research team well illustrates the differences in both parsimony and ethical issues between observational and experimental studies. In the 1960s, a major subject of research interest to ourselves and a number of others was to gain an understanding of the effect of maternal diet during pregnancy on both the viability and mental competence of the offspring. As we explored the question, we formed the opinion that observational studies, however carefully constructed, could not alone provide decisive answers, and that an experimental approach was essential. We designed two studies to test this critical question. One was a record-based study that made use of a natural experiment (the severe Netherlands famine in the last months of World War II in 1944–1945) and thus avoided the need to manipulate human subjects intentionally. This study, outlined below, was accomplished with relative speed and economy. The ethical issues were limited to those involving confidentiality, and simple solutions could be sought.

In World War II during the winter of 1944–1945, the cities of the Netherlands endured a six month blockade by Hitler's enemy forces, precipitating a famine that became increasingly severe especially during the last two months of the blockade. Even under severe stress, in the orderly Dutch society food was carefully measured and rationed. Since the caloric content of the diets were recorded, consumption could be estimated. Information about the exposure of infants born in these cities before, during, and after the period of famine could be assembled from hospital records and assigned for virtually all births in the maternity units of major cities. In particular the records included date of birth, birth weight and even, in a number of centers, postpartum maternal weight. Hence, in a first step, one could measure exposure to a diet greatly reduced in amount and even content and, in a second step, assess the effect of diet on pregnancies. The main prior hypotheses, however, concerned the effects on cognitive function in survivors. The routine records for military enlistment included a battery of cognitive tests. Thus the results were on record for some 400,000 18-year-old men born before during and after the period of famine and medically examined for military service in the centers of the regions affected by the famine experience. With remarkable generosity, the National Health Department made available to us for study (anonymously, of course) the relevant data in areas both affected and unaffected by famine before, during, and after the famine period (Stein

et al. 1975). To our considerable surprise, we found that among the specified cohorts exposed to prenatal starvation, and of lower-than-expected weight at birth, there was absolutely no effect on cognition (Stein et al 1975).

Another natural experiment, like the Netherlands famine fortunately unlikely to recur, involved earlier studies among survivors of the residents in Hiroshima exposed to the atomic bomb in 1945. At first, effects on births and in childhood, and later also of adults, provided unique scientific data on the hazards of irradiation on the intrauterine embryo, and then on effects at various ages through childhood, adolescence, and in adulthood (Neel et al. 1974; Schull and Neel 1959). (With regard to specifying research designs, we note that both these studies might equally well be described as retrospective cohort studies.)

Contrived Experiments

By contrast with the relative ease of accomplishing the Netherlands study given the infant data, a second study we designed to test the relation of maternal diet to the viability and health of the offspring was a contrived experiment undertaken in New York City. From the start, this study faced a conflict between scientific and human considerations. This trial of prenatal nutritional supplementation to prevent low birth weight (Rush et al. 1980) had both a precise hypothesis and a well-specified intervention and outcome. It might have been argued—and some reviewers of the initial proposal did so argue—that the answer could be obtained through observational studies (Susser et al. 1978). In truth, reductions in birth weight following prenatal exposure to famine had been demonstrated in at least three studies of World War II, although outcome in those offspring at later ages was studied only in the Netherlands (Stein et al. 1975).

Nonetheless, when it came to a controlled trial in New York City of nutritional supplements to raise birth weight among pregnant black women at high risk of low birth weight, who were randomly assigned to either of two forms of prenatal supplementation, one high in protein and the other high in carbohydrates, neither diet succeeded in raising birth weight significantly. In fact, to our consternation, a high-protein supplement actually produced the adverse effects of an excess of prematurity, of low birth weight, and of newborn deaths, while the carbohydrate supplement produced a modest increment in birth weight without evident ill effects.

Among controlled trials, a critical issue that this New York study could not resolve with entire certainty, however, was to ensure that the experimental intervention actually took place as specified among all the participants. Epidemiologists have learned that large proportions of patients may

not adhere to prescribed treatments. With food dispensed among a largely poverty-stricken population, there was the further risk that the treatment might either simply replace, rather than supplement, the regular diet, or else be distributed also among other family members of a household.

To avoid this ethically and practically doubtful expedient, we developed an elaborate system for taking diet histories, counted the unused cans of dietary supplement left on the shelf at each replenishment, measured the output of a biochemical marker in the urine, and observed weight changes in the women, all measures aimed at monitoring the intake of the nutritional supplement. No single measure provided entirely secure information about the amount of supplement consumed by individual women, although the available data did enable us to compare outcome in the groups assigned to each diet with some confidence.

The lack of firm knowledge in the field of human nutrition is owed essentially to just this problem: nutrient intake can seldom be measured with precision in large, free-living populations. Once again this weakness resides in potential difficulties for epidemiological investigations, whether natural or contrived, to reconcile scientific needs with appropriate demands far close ethical surveillance. Many closely controlled environments, as with Goldberger's orphans and prisoners, or as with those in mental hospitals, may not readily lend themselves to the development and acceptance of experimental studies.

A lesson here is that among human beings, researchers can seldom if ever be assured of anticipated success of an intervention aimed at counteracting the observed effects of a deficiency or some other adverse factor. For one thing, many variables interact to cause an outcome. Interventions select but one or a few of these variables, thereby ignoring or bypassing unknown interaction effects. Even if the selected experimental interventions are exactly the obverse or the counterpart of adverse factors—and a degree of speculation must reside in the judgment that they are—the effect on the given outcome is of necessity uncertain. Results that invariably confirm existing ideas are suspect in science, which to a degree is sustained by its subversive challenge to existing belief. As seen in this study of prenatal nutrition (Rush et al. 1980), as also in the University Group Diabetes Program (Gilbert et al. 1975), this uncertainty always includes the possibility of rendering either no effect or, more disconcertingly, an unanticipated harmful effect. Wisdom is to expect the unexpected and reconsider one's understanding accordingly.

When a prophylactic trial is aimed at preventing chronic disease, sample size tends to be a limiting factor. Very large numbers are sometimes needed to confer statistical power sufficient to demonstrate significant effects. In face of such constraint, sponsors of recent preventive trials have tended to be government agencies, since few other agencies command the resources for

assembling such risky and expensive multicenter projects. (Admittedly, in the 21st century, such wealthy and significant donors as the Gates Foundation have been bold and willing enough to do so.) Thus, toward the end of the twentieth century, the National Institutes of Health launched several experiments in the prevention of hypertension and its effects (1979; 1976b; 1982a; 1982b and also the Multiple Risk Factor Intervention Trial (Multiple Risk Factor Intervention Trial Group 1982) and, most recently, the National Children's Study (Landrigan et al. 2006).

In order to provide the subjects of a trial, the Hypertension and Detection Follow-up Program screened defined populations at 14 centers (N = 178,009) to detect 10,940 persons with diastolic blood pressure greater than 90 mmHg. Participants were assigned randomly either to special care, so-called "stepped care," or to the care of their community physician, so-called "referred care." (About one fourth in both groups were already receiving antihypertensive drugs.) The stepped care group received especially comprehensive care with an emphasis on a hierarchy of drug treatments.

Five-year mortality was first reported as an end point (Hypertension Detection and Follow-up Program Cooperative Group 1979a). A significant reduction in mortality obtained in the experimental stepped care group. Most notable, both for its magnitude and for the size of the group, was the effect in participants with mild to moderate hypertension (diastolic blood pressure 90–104 mmHg).

It is perhaps premature to make a final judgment from the findings for this single end point, but one may raise a number of questions. First, one might ask if the result was not owed solely to the effect of the higher frequency of drug treatment in the experimental groups (the explanation preferred by investigators) but also to effects of the comprehensive care that the group received? In favor of a drug effect is the congruence of the treatment effect with the proportions among the experimental and comparison groups who reported using antihypertensive drugs as directed.

Second, one might ask if it was appropriate for the investigators to abjure a placebo treatment group? This decision rested on the ethical ground that antihypertensive drug treatment had previously been shown to reduce the morbidity among males from severe hypertension. Were these grounds legitimate for mild hypertension, or even for women? For these classes, no effects of antihypertensive treatment had been previously demonstrated. Whether women could have been considered a potentially separate biologic class was perhaps almost as much a political as a moral or scientific question. Third, one may ask what the implications of the study are for preventive intervention. Given the favorable effect on mortality among those with mild hypertension, the sponsors concluded that everyone with any degree of hypertension should be vigorously treated.

Studies of large questions with large consequences plunge epidemiology into the arena of major policy. On some estimates, treatment of mild hypertension could extend to 15% of the total population and 30% of adults or, in the United States, more than 30 million people. On more stringent criteria, not less than 6% or 7%, say 9 million people, would be treated. Indeed, the National High Blood Pressure Education Program, under the auspices of the National Institutes of Health, claimed that as many as half the adult population (60 million Americans), had high blood pressure and are candidates for treatment (The National High Blood Pressure Program 1978).

If this prescription were to be followed, one is perhaps obliged to ask what might be the appropriate recommendation for several millions at risk of treatment among subgroups in whom no distinctive differences in mortality emerged. The investigators themselves brought this question to attention. Their analysis of groups subdivided by race, sex, and age, the subject of a second report, showed an effect on blood pressure level in all groups, but no effect on mortality either in white women, or in members of either sex aged 30 to 49 years (Hypertension detection and Follow-up Program Cooperative group 1979b). The results for these two groups cannot be played down for lack of power, as implied by the investigators, which is better than 90% for detecting a reduction of 20% mortality.

Subgroup analysis provides a shaky platform for statistical inference unless the strata analyzed have been defined *a priori* as conditions for experimental testing. Post hoc manipulation of the data (so called data dredging), while valuable in itself, raises three particular problems: if voluntary behavior such as adherence to treatment is involved, the strengths of nonselective random assignment to comparison groups peculiar to an experimental design are abandoned; if an *a priori* hypothesis was not set up for testing, the analysis departs from the model of a rigorous experimental test and becomes exploratory; if different segments of the data are tested, each "look" at a subgroup carries its own probability of observing a result by chance, and thereby alters the subsequent conditional probabilities that attach to the hypotheses under test. The third problem, namely, that of data-dredging, can be rephrased in terms of the search for statistical interaction (which is present when results vary across subgroups). To test systematically for interaction is greatly to reduce statistical power and to raise the likelihood of both false negative and false positive results. In tests for interaction, even random misclassification can produce spurious associations (Greenland 1980).

It turned out that subjects in the Hypertension Detection and Follow-up Program had been randomized only within three predetermined strata of blood pressure level, but within no other strata. By the strict conventions of hypothesis testing, therefore, the absence of effects in the two special sex and

age subgroups not randomized is without force upon the general result. This is in contrast with the effect in mild hypertension, a subgroup set up *a priori*. Yet, in this instance, the scale and seriousness of the public health problem lends same merit to the plea that subgroups should not be entirely ignored in reaching a conclusion and that statistical inference should not stand in the way of logical inference. In publishing results for subgroups constructed *post hoc*, the investigators were surely moved both by the momentous issues they faced and the potential importance of the variations in the data they had collected.

In order to make a judgment, many epidemiologists will want data to be brought to bear from other sources, as well as more results from the study itself. The effects of "stepped care" on mortality from mild hypertension withstand additional probing for potential confounding of the results either by treatment or by the extent of organ damage prevailing at entry into the trial (Hypertension Detection and Follow-up Program Cooperative Group 1982a). A major Australian trial (Report by the Management Committee 1980) of drugs versus placebo for hypertension, randomly stratified by age and sex, also supports the experimental results of the Hypertension Detection and Follow-up Program on mild hypertension (>94 mmHg). The experimental subgroup analysis was ambiguous, however. No significant effect was obtained either in women or in the age group under 50 years, although in both these subgroups the direction of the result conformed with a treatment effect.

The supportive conclusions of another analysis of the Hypertension Detection and Follow-up Program—that more than half of the excess risk in the referred care group could be attributed to treatment—remain suspect because of self-selection for compliance with treatment (Hardy and Hawkins 1983). More convincingly, the Program finally demonstrated an effect on a different end point—cerebrovascular disease—that was consistent with average blood pressure levels and held across all age and sex groups (Hypertension Detection and Follow-up Program Cooperative Group 1982b).

Growing certitude was undone by the Multiple Risk Factor Intervention Trial (Multiple Risk Factor Intervention Trial Group 1982) since it failed to demonstrate an effect on mortality of antihypertensive treatment combined with other preventive measures**.** The Multiple Risk Factor Intervention Trial aimed to use the high-risk selection strategy to reduce mortality from coronary heart disease by interventions aimed at reducing multiple risk factors. A trial involving diet alone had been given up as infeasible, mainly because of the huge numbers and cost in the face of low rates of compliance and the expense of providing the diet at special food centers (Report of the Diet-Heart Review Panel of the National Heart Institute 1969).

About 350,000 men at 22 centers across the country were screened in order to recruit close to 13,000 participants aged 35 to 57 years.

These men, selected for high "Framingham coronary risk scores," were randomly assigned to intervention. The intervention program consisted of advice about reducing dietary cholesterol, counseling sessions on smoking, and stepped drug treatment for hypertension. Sample size for the study was estimated from realistic anticipation of the reduction that could be achieved in each risk factor and from the supposition that in turn these reductions together might reduce coronary mortality by about one fourth. The sample could be relatively small because sample size did not allow for hypotheses about subgroups and because a one-tailed test of significance was used. After an average follow-up of seven years, the first set of results (Multiple Risk Factor Intervention Trial Group 1982), showed no advantage for the experimental group in total mortality (actually +2%) and a minor nonsignificant advantage in coronary heart disease mortality (actually –7%). Several interpretations of the results have been considered:

1. The regimen for intervention was ineffectual. If so, it was not ineffectual because of any failure to reduce risk factor levels; better than 80% of the goal for risk factor reduction was achieved. Furthermore, unlike the Hypertension Detection and Follow-up Program, this regimen did not have to compete with the active referral of the control group to treatment elsewhere. Since many high-risk individuals at the ages recruited for the trial are likely to have had advanced atherosclerosis, however, it is quite possible that the attempt to intervene came too late for primary prevention.
2. Effects were beneficial in most groups, but in a minority of subgroups counterbalanced by adverse effects of antihypertensive drugs (a view favored by the investigators). Thus, men in the experimental group who had both hypertension and electrocardiographic abnormalities actually experienced excess coronary heart disease mortality (+65%). Also, those who had mild hypertension did no better than controls. Once again, however, the analysis of subgroups not created *a priori* does not by itself controvert the statistical inference of a null result in a given experiment as designed.
3. The regimen was effective, but "contamination" of the control groups reduced statistical power to the point at which the effect could not be detected. The regimen for the experimental group had to compete with sensitization to risk in the control group. Those assigned as controls and their doctors were of course informed about their high risk and followed annually, and their voluntary response was marked. Thus, with regard to risk factors, after six years in the study, blood pressure and cholesterol levels of the control group were reduced by margins

almost as large as those of the experimental group. Only the reduction in cigarette smoking distinctly favored the experimental group.

With regard to outcome, the experiment also had to compete with history. Over the period of observation, mortality from coronary heart disease in the United States had declined almost as much as was hoped for from the special intervention program. At the six-year point, the number of deaths in the control group was a little less than half the expectation (219 versus 442). This reduction in control group risk factors and expected deaths together reduced the chance of detecting an effect on mortality from around 90% to 60%. That is, actual statistical power was low.

Thus, in the result, these major and courageous experiments are beset by ambiguity and conflict. Epidemiologists who undertake large-scale tests of hypotheses dear to many, and who produce negative or adverse results, must be ready to face criticism from both friends and enemies. Such issues are likely to be buffeted about by ideology, by commercial interests, or by the constraints on national budgets that ruling economists no longer find tolerable.

To conduct large-scale epidemiologic experiments is still to pioneer. Avoidable errors as well as unavoidable inconsistencies are costly but inevitable. At this time, no better means offer for answering sharply posed questions about interventions in a definitive way. Other approaches might be possible under special conditions. For instance, were there to be an adequate database covering total health care for a number of populations, of the kind that is available for Olmsted County through the Mayo Clinic, quasi-experimental studies could yield valuable answers. Again, case-control studies of exposure to intervention could be mounted around selected endpoints (Clarke and Anderson 1979; Sargent and Merrell 1940; Rhoads and Mills 1985). Yet, the effort and expense of large-scale experimental studies cannot be entirely avoided if we are to learn how to do better in sharpening the answers to our questions.

References

Carpenter KJ (1986). *The History of Scurvy and Vitamin C*. Cambridge: Cambridge University Press.

Clarke EA, Anderson TW (1979). Does screening by "Pap" smears help prevent cervical cancer? A case-control study. *Lancet* 2:1–4.

Francis T (1957). Poliomyelitis vaccine center. Evaluation of 1954 field trial of poliomyelitis vaccine; final report. Ann Arbor: University of Michigan.

Gilbert JP, Meier P, Rumke CL (1975). Committee report assessing the biometric aspects of controlled trials of hypoglycemic agents. *JAMA* 231:584–608.

Greenland S (1980). The effect of misclassification in the presence of covariates. *Am J Epidemiol* 112:564–569.

Hardy RJ, Hawkins CM (1983). The impact of selected indices of antihypertensive treatment on all-cause mortality. *Am J Epidemiol* 117:566–574.

Hypertension Detection and Follow-up Program Cooperative Group (1979). Five-year findings of the hypertension detection and follow-up program. I. Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 242:2562–2571.

Hypertension Detection and Follow-up Program Cooperative Group (1979). Five-year findings of the hypertension detection and follow-up program. II. Mortality by race, sex and age. *JAMA* 242:2572–2577.

Hypertension Detection and Follow-up Program Cooperative Group (1982). Five-year findings of the hypertension detection and follow-up program: the effect of treatment on mortality in "mild" hypertension. *N Eng J Med* 307:976–980.

Hypertension Detection and Follow-up Program Cooperative Group (1982). Five-year findings of the hypertension detection and follow-up program. III. Reduction in stroke incidence among persons with high blood pressure. *JAMA* 247:633–638.

Landrigan PJ, Trasande L, Thorpe LE, Gwynn C, Lioy PJ, D'Alton ME, Lipkind HS, Swanson J, Wadhwa PD, Clark EB, Rauh VA, Perera FP, Susser E. (2006). The National Children's Study: a 21-year prospective study of 100,000 American children. *Pediatrics* 118:2173–2186.

Last JM, Abramson JH (1995). *A Dictionary of Epidemiology*. New York: Oxford University Press, p. 110.

Lind J (1753). *A Treatise of the Scurvy*. Edinburgh: Sands, Murray and Cochran.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 2:769–783.

Multiple Risk Factor Intervention Trial Group (MRFIT) (1982). Multiple risk factor research trial: risk factor changes and mortality results. *JAMA* 248:1465–1477.

Neel JV, Kato H, Schull WJ. (1974). Mortality in the children of atomic bomb survivors and controls. *Genetics* 76:311–336.

Oakley A (1998). Experimentation and social interventions: a forgotten but important history. *BMJ* 31:1239–1242.

Palmer CE, Shaw LW, Comstock GW (1958). Community trials of BCG vaccination. *Am Rev Tuberculosis* 77:877–907.

Report by the Management Committee (1980). The Australian therapeutic trial in mild hypertension. *Lancet* 1:1261–1267.

Report of the Diet-Heart Review Panel of the National Heart Institute (1969). Mass field trials of the diet-heart question: an assessment of seven proposed experimental designs. American Heart Association Monograph 28. New York: American Heart Association, Inc.

Rhoads GG, Mills JL (1985). The role of the case-control study in evaluating health interventions: vitamin supplementation and neural tube defects. *Am J Epidemiol* 120:803–808.

Rush D, Stein Z, Susser M (1980). Diet in pregnancy: a randomized controlled trial of nutritional supplements. Birth defects: original article series 16, no. 3. March of Dimes Birth Defects Foundation. New York: Alan R Liss.

Sargent CA, Merrell M (1940). Method of measuring effectiveness of preventive treatment in reducing morbidity. *Am J Public Health* 30:1431–1435.

Schull WJ, Neel JV (1959). Atomic bomb exposure and the pregnancies of biologically related parents. a prospective study of the genetic effects of ionizing radiation in man. *Am J Public Health: Nations Health* 49:1621–1629.

Smithells RW, Sheppard S, Schorah CJ, Seller MJ, Nevin NC Harris R et al (1980). Possible prevention of neural tube defects by periconceptional vitamin supplementation. *Lancet* 1:339–340.

Stein Z, Susser M, Saenger G, Marolla F (1975). *Famine and human development: the Dutch hunger winter of 1944/45.* New York: Oxford University Press.

Susser M, Stein Z, Kline J (1978). Ethics in epidemiology. *Ann Am Acad Pol Soc Sci* 437:128–159.

Szmuness W, Stevens CE, Harley EJ, Zang EA, Oleszko WR, William DC et al. (1980). Hepatitis B vaccine: demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States. *N Engl J Med* 303:833–841.

Terris M (1964). *Goldberger on Pellagra*. Baton Rouge: Louisiana State University Press.

The National High Blood Pressure Program. NHLBI Infomemo. New prevalence estimates provide a more complete view. US DHHS, Public Health Service, May 1978, no. 13.

19

New Designs and Models

It has become clear that flexibility and ingenuity is called for in both design and analysis. In causal analysis, the need for tight control is a shibboleth which the discipline of epidemiology must live by. At the same time, no competent scientist will want to be shackled by rules. Thus, at least five circumstances come to mind—most of them rare—each of which can be indicative of causal associations in the absence of conventional controls.

1. Exceptional cases. In 1831, a cholera epidemic first struck Britain from the Continent. A second epidemic of the highly fatal disease broke out in 1848. John Snow, general practitioner and anesthesiologist, undertook his investigations of the then mysterious ravages of a large cholera cluster around the Broad Street pump in central London as described in Chapter 8. He also followed the case of a Mrs. Eley, the widow of a cartridge maker (whose firm's shotgun cartridges author Mervyn Susser still used as a boy in the South African Bushveld in the mid-1930s). She had moved to Hampstead, then a village north of London's bounds (Vinten-Johansen et al. 2003). Mrs. Eley, herself once a resident in Broad Street and partial to the water from the pump, had water from the Broad Street pump trundled daily to her home. She fell ill with cholera and died. The first and only case in Hampstead, her case became truly exceptional when the doctor who called to attend to her

himself fell ill. On that hot summer's day, having availed himself of a glass of her water, he too had a bout of watery diarrhea (Susser 1964).

2. Exceptional clusters. For the sake of terminological clarity (if not numerical exactitude) one might usefully sustain a deliberate distinction between *clusters* of a few cases, *outbreaks* of a substantial number (as occurred around the Broad Street pump, which then coincidentally caused the isolated death of Mrs. Eley in Hampstead), and *epidemics* involving the spread of large numbers across populations. Before the Broad Street outbreak, clusters of a few cases had occurred in a housing development south of London. Snow had at once set off to investigate these, which raised his profound suspicions about a water supply that, on close inspection, was clearly open to fecal contamination from a nearby stream. Clusters of a few cases often raise the suspicion of epidemic spread, but seldom result in a verdict beyond "not proven": in the few cases, the exposure dose too usually remains in doubt. A firmer verdict can be supported where the cluster comprises rare or new disorders, the exposures are of known time order, and the measured strength of the true risk is high. In such circumstances, small point epidemics quite commonly incriminate rare and hazardous exposures even without benefit of controls. Thus, among many other instances, the rare cancers of angiosarcoma following polyvinyl chloride exposure and of mesothelioma following asbestos exposure can be firmly attributed to those specific exposures.

In 1943, Norman McAlister Gregg, an Australian ophthalmologist confronted by an unusual series of infant cataracts discovered, when prompted by his nurse, that each case was preceded by maternal rubella infection acquired during pregnancy in the course of a recent epidemic. (The finding proved to be a spur to fetal epidemiology; Kline et al. 1989.) Likewise in Hamburg, Germany, in 1953, Lenz used a register of births that recorded congenital anomalies to show the association of phocomelia (failure of limb development) with maternal use in early pregnancy of the then new antiemetic thalidomide (now banned) (Lenz 1962). In another instance, Christopher Wagner, a South African pathologist, recognized the conjunction of several cases of the rare malignant mesothelioma of the pleura in persons who had lived around asbestos mines (Holland et al. 1979).

3. Recurrent episodes. The fact that episodic spells characterize a case series, or sometimes even a single case of a disorder, may support or negate a causal inference. Such an inference makes use of the *timing* of *event-related* episodes. For instance, American neurologists long shared a belief, drawn from just such data, that alcohol withdrawal precipitated "alcoholic seizures." On further study, the relationship did not hold. The distribution of the interval between last drink and the onset of seizures, assessed against a random distribution of intervals, bore no relation to withdrawal. However, it did bear

a strong relation to the amount of alcohol ingested before the seizure (Ng et al. 1988). Also, in the course of a single disorder, treatment effect can be inferred from a design based on successive periods of treatment versus treatment withheld (Lilienfeld et al. 1967). A somewhat related approach is the *case-crossover* design (Taylor and Knowelden 1957).

4. Medical histories. In the not very usual instances where an outcome of a disease is known to have been uniform and virtually inevitable, only a few cases in which intervention averts the outcome might suffice for an inference that a given intervention was the effective cause of recovery. Such egregious examples occur when, in a disease previously always fatal, a new treatment is followed by instances of survival. In the later nineteenth century, survival of a child bitten by a rabid dog sufficed to prove the efficacy of Pasteur's vaccine against rabies. Likewise, around the mid-twentieth century, the definitive efficacy of penicillin was made clear when the anticipated fatalities from victims of subacute bacterial endocarditis, septicemias, and severe gram positive pneumonias began to survive. A decade later again, the survival of a few cases of tuberculous meningitis with the advent of the antibiotic streptomycin (even though sometimes accompanied by auditory nerve deafness) gave acceptable proof of its efficacy. The new antibiotics were quickly accepted as the proven treatment for each of these previously fatal conditions.

5. Case-case studies. Under some circumstances, studies within a series made up solely of cases can proceed without the preselected controls normally required (Begg and Zhang 1994). For instance, with a lump in the breast, an initial diagnosis by mammogram when confirmed and refined by biopsy allows the cases to be segregated by pathogenetic origin into benign and malignant types. In the search for determinants of malignancy, the benign types are then well-matched to serve as controls for breast cancer, since the selection for entry to study of those with and without cancer is identical and without bias. Similarly, in a study of the factors underlying anginal pain in patients found on angiography to have coronary artery occlusion, their uniform selection for entry to study ensures that those with angina are likely to be well-matched against those without it (Shea et al. 1984).

Epilogue

How then should one assess this period of the history of epidemiology? The diffusion of new epidemiologic knowledge undoubtedly contributed to major changes in medical consciousness and practice, in health behavior generally and in reductions in morbidity and mortality: smoking, diet, and physical exercise all changed for the better; so too did the regulation of

environmental and occupational hazards, drugs, and dangerous substances. Yet the developmental history we have described has taken place in the shadow of a paradox. While epidemiology has flowered, the unprecedented frightfulness of nuclear destruction has beset us. That ultimate damage and the dread of it is common to every sentient person living in our age. But the wonders and terrors of science are not unconnected. Even in its most benign aspect, epidemiology must endure the irony of its ties with the malignant outgrowths of scientific creation. Today's epidemiologists, like all scientists, are the affines of the Manhattan Project and Los Alamos.

From those undertakings stems the era of "big science," a multifaceted science wrapped in high technology that can only be sustained by government. Epidemiology cannot but share in this evolution. Much of modern epidemiology must be on a large scale because of the demands inherent in its new ambitions. Like other sciences in the United States, and despite occasional buffetings from uncomprehending governing powers, for the most part epidemiology has been cossetted by the federal government. The sins of science arise from the power of its applications. The discoveries of scientists have consequences, unanticipated as well as anticipated, harmful as well as helpful. As the handmaiden of public health and medicine, epidemiology becomes by definition applied. Thus, epidemiologists have sought to go beyond understanding disease, in order to prevent or control it. In the contemporary United States, the discoveries of epidemiology—or more commonly its ambiguous findings—are seen to have consequences as never before.

To demonstrate a carcinogen or a mutagen is at once to bring regulatory agencies and affected industries into conflict or collusion. The failure to demonstrate a favorable effect as, say, in the case of nutritional supplementation, may be the occasion for budget cutting on the part of government, and for hunger on the part of the poor. The failure to demonstrate toxic effects of waste chemicals may form the grounds for relaxed regulation and more environmental contamination. In studies of air pollution, the tough-minded criticism of serious design faults that detract from reported effects (Holland et al. 1979) lends political support for reducing standards of air quality. With that damage done, to demonstrate that the critique neglects many considerations (Shy 1979), such as the doubtful results produced by weak statistical power and weakly measured variables, may be insufficient to counter the damage. A probably inflated estimate that 40% of cancers have an occupational source leads to one kind of government action and industry reaction; a possible underestimate that such cancers comprise no more than 5% leads to another kind of action, or more likely, to none (Davis 1983; Doll and Peto 1981).

In such matters in the United States today, epidemiologists testify before congressional committees, official commissions, and judges. Not

uncommonly and to some detriment of the discipline, they are to be found sometimes as paid experts, on both sides of the question. Contention on all these issues has strengthened the rigor of epidemiologic research and inference. The inconsistency of results across studies characteristic of epidemiology raises the pitch of argument. Inconsistency is a fact of epidemiology that must be lived with. If not by chance, inconsistency may arise from weaknesses and differences in method, from subtle elaborations of analysis, or from true differences and changes in populations, environments, and agents. Thus, the testing of results by replication under the same and under different conditions is a necessary resort, however tedious. To a large extent in the United States, the federal government agencies have become the arbiters of what the important problems are that need research and should be funded. Thus, these agencies mould the direction of the greater part of the national research effort. The incentive for epidemiologists to be responsive is enormous if not irresistible.

Some speculate that the combination of generous funding with guidance in the direction of research dampens imagination. Indeed, some applaud the precarious situation of many researchers today—in which they may have little assurance of continuing funds for themselves or their co-workers—as promoting the survival of the fit. Yet, one may well argue the converse. To scurry for funds certainly requires energy and activity. It is a question whether, in the larger scientific context, that activity is well directed. The once well-funded system for training and research of past years has been the key to creating the gratifying number of competent epidemiologists in the country today. There is no reason to think that comfortably funded researchers who have the gift of scientific imagination do not exercise it well.

Greater dangers may arise as corporate sponsors enter the field in response to the federal regulation of occupational health and safety. These sponsors are helping to create a new specialization in occupational epidemiology (alongside the few other specializations such as social, environmental, and genetic epidemiology that take origin from the independent variable). They are bound, by the nature of their economic, organizational, and bureaucratic needs, to prove less disinterested than the National Institutes of Health. Epidemiologic research takes place within the larger society. Its guiding concepts cannot be disconnected from that context, although they may transcend it. Within those limits, epidemiologists have forged research instruments with ever sharper cutting edges. With an appearance of bland neutrality, however, technique may veil the silent intrusion of values and the shaping of judgments by unrecognized societal forces. In respect of such forces and pressures, epidemiologists need to remain alert.

Modern epidemiology is a mature academic discipline, and epidemiologists properly espouse the values of science and the search for truth or, at the least, explanation and adequate interpretation. Those values are not incorruptible, as many scientists have fondly believed. Our obligation is to cultivate awareness and to see that what is learned is not misused. At the same time, epidemiologists do well to look to the origins of epidemiology as an applied discipline that sustains another set of values. The founders were zealous for the public weal. If we go only as far as the founders, we shall use our discipline to the best advantage in the prevention of disease and the preservation of health.

References

Begg CB, Zhang ZF.(1994). Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev* 3(2):173–175.

Davis DL, Bridboard K, Schneiderman M (1983). Cancer prevention: assessing causes, exposure, and recent trends in mortality for US males 1968–78. *Int J Health Serv* 13:337–372.

Doll R, Peto R (1981). The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *JNCI* 66:1191–1308.

Holland WW, Bennett AE, Cameron IR, Florey CDF, Leeder SR, Schilling RSF, Swan AV et al. (1979). Health effects of particulate pollution: reappraising the evidence. *Am J Epidemiol* 110:533–659.

Kline J, Stein ZA, Susser MW (1989). *Conception to Birth: Epidemiology of Prenatal Development*. New York: Oxford University Press.

Lenz W (1962). Thalidomide and congenital abnormalities. *The Lancet* 1(Jan 6): 45.

Lilienfeld AM, Pedersen E, Down JE (1967). *Cancer Epidemiology: Methods of Study*. Baltimore: The Johns Hopkins University Press.

Ng SK, Hauser WA, Brust JC, Susser M (1988). Alcohol consumption and withdrawal in new-onset seizures. *N Engl J Med* 319(11):666–673.

Shea S, Ottman R, Gabrieli C, Stein Z, Nichols A (1984). Family history as an independent risk factor for coronary artery disease. *J Am Coll Cardiol*.4(4):793–801.

Shy CM (1979). Epidemiologic evidence and the United States air quality standards. *Am J Epidemiol* 110:661–671.

Susser M (1964). The uses of social science in medicine. *Lancet* 2:425–429.

Susser MW, Watson W (1962). *Sociology in Medicine*. London: Oxford University Press.

Taylor I, Knowelden J (1957). *Principles of Epidemiology*. Boston: Little, Brown & Co.

Vinten-Johansen P, Brody H, Paneth N, Rachman, S, Rip M (2003). *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow*. New York: Oxford.

20

Social Science in Epidemiology

Introduction

In the 1950s the contributions of the social sciences to the theory and practice of public health, medical practice, and epidemiology were still uncharted. Indeed, one of the avowed purposes of Sociology in Medicine (Susser and Watson 1962) was to accomplish such a task. The chapter headings convey the scope they envisaged for the field (Table 20.1).

Some of these topics have been developed over the past half century, some neglected, and some now mainly absorbed into other fields. We follow these topics somewhat selectively, as we have other themes in this book: in particular, we have not dealt here with the considerable expansion of research and writing arising from the HIV epidemic, much of which is discussed in Chapter 23.

First, the topics covered in the chapter "Health, Population and the Economy" have developed considerably, and the World Health Organization

This chapter is a revised and expanded version of Susser M and Myer L (2007) Social Epidemiology, in Holland WW, Olsen J, Florey CDV, eds. *The development of modern epidemiology: Personal reports from those who were there.* New York: Oxford University Press.

Table 20.1 Chapter Titles from *Sociology in Medicine* (Susser and Watson 1962)

1. Health, Population, and the Economy
2. Culture and Health
3. Social Class and Epidemiology
4. Social Mobility
5. Medicine and Bureaucracy
6. The Cycle of Family Development
7. Mating and Marriage
8. Infant to Adult
9. Social Support

has taken leadership here: for instance in comparisons across countries, in measures of Disability, of Quality of Life, and of many other areas.

"Culture and Health" is now well represented, far beyond the early but still deservedly much studied text *Health, Culture and Community,* a series of case studies edited by Benjamin Paul in 1955 (Paul 1955). A more recent text, *Medical Anthropology in the World System*, published in 1997 (Baer et al. 1997), and again in 2003 (Baer et al. 2003), brings the field up-to-date, as does *Cultural Diversity in the United States* (Susser I and Patterson T.C. 2001). Ethnographic studies advance epidemiological understanding well beyond cruder and broader-based classifications, such as country of origin or ethnicity. A focus on the social factors relating to particular diseases (as, for example, mental disability, epilepsy, cancer, and many others) has been addressed mainly by social scientists. Thus in Chapter 22 we cite the joint work of the anthropologists and epidemiologists who contributed to unraveling the mystery of Kuru in New Guinea.

Several studies of social class and social mobility are documented in more detail in what follows. Because social position can be defined more or less precisely by censuses and public health records, this field has lent itself to many epidemiological studies. With regard to studies of the family cycle, subsequent studies of the health of women have emphasized the influence and effects of their changing roles as well as on the function of health systems. The subject of social support, with notable exceptions in several fields such as the family, aging and homelessness, has been somewhat neglected by epidemiologists. The topic of "Medicine and Bureaucracy" devolves around studies and evaluation of health care policies and systems, and development has been substantial among economists and sociologists although, with notable exceptions, less so among the health professions.

1950s–1980s

In earlier eras, a prime social concern was the issue of poverty, with its accompaniments of poor sanitation, housing, hunger, high mortality rates,

and the struggle for mere survival. In sharp contrast, however, in the post–World War II period in Western Europe and the United States, the major dislocations of the industrial revolution had subsided. In these countries and no less in others of the northern hemisphere, although the residual social ills had not disappeared, they were much reduced.

In medicine and public health, new capabilities generated a sense of confidence in the potential of prevention and treatment. From the late 1930s sulphonamide chemotherapy, and from the 1940s penicillin and other antibiotics, had succeeded in curbing the predominant scourges of bacterial infection. Further advances in microbiology and immunology had reinforced understanding of physiological defenses. Unsurprisingly, the apparent conquest of epidemic infection turned the attention of epidemiologists to a rising toll from chronic disease. With the realization that the distribution of serious health disorders had shifted, likewise many epidemiologists shifted their central preoccupations to a search for causes and potential risk factors in cardiovascular, pulmonary, renal, and cerebral disorders (cf. Kermack et al. 1934). Over and above the broad societal determinants of population health, in the 1950s through the 1980s, the effects of such individual behavior as smoking, exercise, and diet began to attract close attention (Susser and Stein 1962; reprinted in 2002). Indeed, in Britain even more than in the United States, this focus dominated most centers of academic epidemiology (Morris 1957).

There is little doubt, however, that the pathbreaking event in chronic disease epidemiology in post–World War II—and, one might say, in the twentieth century—was the definitive demonstration of Richard Doll and his mentor Austin Bradford Hill that cigarette smoking was the predominant cause of the mystery of the rising epidemic of lung cancer (Doll and Hill 1952). With that, the central focus of epidemiological research in Britain shifted sharply to searching for causes of chronic disease.

In the United States, in the first three decades of the twentieth century Joseph Goldberger and Edgar Sydenstricker had set standards for epidemiological research that encompassed societal and social factors in states of health (Goldberger et al. 1920; also see Chapter 14). Wade Hampton Frost (1880–1938) was the founding professor of epidemiology at the first school of public health in the United States, established at Johns Hopkins University in Baltimore. He did much significant and rigorous work—first, on the sources of largely waterborne epidemic infectious disease. He also participated in the initial Hagerstown surveys with Edgar Sydenstricker. From the mid-1950s, at Johns Hopkins, however, Abraham Lilienfeld and George Comstock, and then at Columbia University, Gurney Clark, and at Harvard, Brian MacMahon (a student of McKeown in Birmingham), and in California, Warren Winkelstein and Reuel Stallones at Berkeley, led a shift in emphasis toward the search for individual risk factors as "exposures" predisposing

to chronic conditions (MacMahon B et al. 1960). But in the United States, epidemiological study of societal determinants of mental health stemmed as much from the social sciences as from epidemiology itself. Most mainstream epidemiological research gave social and economic determinants relatively modest attention.

After a beginning in the 1930s by L.J. Henderson, however, American sociologists had studied biomedical practice for some time, thus creating a sociology *of* medicine (Henderson 1970). The 1950s also marked new ventures in medical sociology that addressed the influence of social conditions on the causes of disease, and more especially mental disorders. Many built on the work in Chicago of Faris and Dunham. They found the highest rates of schizophrenia in the most deprived circumstances in the city (Faris and Dunham 1939). Their findings posed the question of whether the high rates were a causal effect of deprived circumstances or a result of the downward social drift of those already affected. In 1958 at Yale University, sociologist August Hollingshead and psychiatrist Frederick Redlich published their book, *Social Class and Mental Illness* (Hollingshead and Redlich 1958), a study of the social distribution of mental disorder in the town of New Haven. They too demonstrated a gradient of mental disorders, lowest in the highest social class and rising with descending social class that was consistent with that described by Faris and Dunham for schizophrenia. Hollingshead and Redlich attributed the pattern largely to a parallel decline in the quality of psychiatric care. At the University of Michigan, the physician Sidney Cobb and his student, Stanislav Kasl, centered their work on the psychological effects of stress and cardiovascular disease (Kasl et al. 1968; Kasl and Cobb 1980). Hollingshead's student Leonard Syme was one among several sociologists who applied epidemiologic methods to the study of social determinants of cardiovascular disease. Syme and Leo Reeder (Syme's sociological colleague at the University of California Berkeley) also concentrated on the role of stress as a factor in cardiovascular disease (Syme and Reeder 1967). Later, Syme and his student Michael Marmot (Marmot and Syme 1976), in a study of the incidence of cardiovascular disease among Japanese-Americans, showed that independently of known risk factors, morbidity increased with the degree of Western acculturation.

The work of Hollingshead, Syme, Reeder, and their students gave rise to further research into how social factors might affect the occurrence of disease and death. Perhaps the best known of such work is that of Lisa Berkman and Syme (her mentor) in their study in Alameda County, California in 1965 to 1974 (Berkman and Syme 1979). They devised a measure of social integration and its relation to mortality (taking account of marital status, community group membership, and contacts with family and friends). Persons with low levels of social integration were two to three times more likely to

die than were those better integrated. The association held regardless of various self-reported high-risk behaviors.

Among epidemiologists in the United States during this period, John Cassel also stimulated interest in social epidemiology (Cassel 1964; Cassel 1976). Upon completing his internship at Witwatersrand University Medical School in 1945, Sidney Kark had recruited him as his deputy in his government-supported project to develop a national network of health centers. As the grip of the Apartheid State tightened in South Africa in the mid-1950s, however, Cassel departed to take up a fellowship at the University of North Carolina, an unintended prelude to permanent residence in the United States. Appointed to the newly created Chair of Epidemiology at the University, Cassel became a highly influential teacher and a leading proponent of psychological stress as a causal factor in disease.

Like others among those influenced by Sidney Kark (including ourselves), Cassel and his colleague H.A. Tyroler looked to the social sciences for insight into how social conditions might shape health outcomes. True to Kark's precepts and practice, they turned to studies of migrants from the hills of North Carolina, a population then in transition as they fled rural poverty and sought work in urbanizing environments.

Much of this research posited that psychological stress enhanced both physiological and psychological vulnerability and ultimately manifested in cardiovascular disease. Cassel's influential and widely cited paper, "The Contribution of the Social Environment to Host Resistance" (Cassel 1976), stimulated many to follow him in the study of psychosocial responses induced by social and economic factors that might influence health. Cassel advocated an approach that viewed the health effects of social conditions in broad rather than disease-specific terms, a perspective that has had a lasting impact on social epidemiology.

At this point we mention the contribution of Sidney Kark, Cassel's mentor, and the school that emanated from his work. Kark, as a medical student at Witwatersrand University in Johannesburg and already socially committed, joined the then strong South African National Department of Health immediately on graduating in 1937. Emily Jaspan, Kark's wife and partner, graduated in 1938, and in 1939, with the support of the National Health Department, they established the first health center committed to the practice of "social medicine" (Susser 1993). Pholela, their chosen site deep in rural Natal province, was peopled predominantly by tribal Zulus. The avowed intention of the Health Department and Kark's project was ultimately to develop a network of such centers providing comprehensive sociomedical services that would serve defined communities across the country. In the first instance, these areas were those populated by poverty-stricken Africans, those most in need. In Pholela the Karks, together with the colleagues they had recruited,

documented the social and economic conditions that facilitated the spread of disease among impoverished inhabitants.

In a classic paper, Kark showed how the spread of sexually transmitted infections stemmed from migration patterns created by structural economic conditions (Kark 1949; recently reprinted Kark 2003). A money economy, outside the African tribal traditions, obliged the males of the community perforce to seek work on the gold mines where they labored a mile and more underground, or in the industrial enterprises of Johannesburg, or even as kitchen and garden domestics in the homes of whites. Living in single-sex dormitories, miners had easy access to the sex workers who likewise had migrated to the city in search of work and money. In a routine created by the gold mining industry to sustain its labor force, the men would work for nine months of the year, and for the remaining three months return to their rural homes to plough the fields and plant the crops—corn in the main—and tend the cattle or sheep that would keep their families fed.[1]

Of particular note, during the 1950s, several of Kark's colleagues and followers—among them the physicians Harold Phillips and his wife Eva Salber, Bert Gampel and his wife Julia Chesler, Joseph Abramson, social psychologist Guy Steuart, and sociologist Helen Navid—all left the Apartheid State of South Africa. Some joined John Cassel's Department of Epidemiology at the University of North Carolina, and others joined Sidney and Emily Kark in Israel at the Hadassah Medical School where Sidney Kark was encouraged to found a Department of Social Medicine. There unhindered, the group as a whole focused on the development and teaching of social medicine and epidemiology (e.g., see the book on survey methods of Joseph Abramson [1974]). The authors of this text too were strongly influenced by Kark's teachings and indeed propagated them (Susser and Watson 1962; Susser 1968). Since we ourselves were not then *persona grata* with the Apartheid government, public service however benign was not an option. Obliged to follow our own independent course, together with our close colleagues, the married couple Michael Hathorn and Margaret Cormack, we applied (to the degree possible) Karkian principles in our own health center practice among the 80,000 Blacks living in Alexandra Township on the northern outskirts of Johannesburg (Susser et al. 1955; Susser 1957; Susser 2006).

The epidemiological studies of Cassel in the United States and Kark in South Africa in some ways anticipated the challenging paper of Geoffrey Rose, published in 1985 and expanded further in 1992 (Rose 1985; Rose 1992). This work provided the critical justification, should one be needed,

[1] As one predicted, this pattern, at the root of epidemic syphilis in South Africa, exactly foreshadowed the devastating spread of the HIV/AIDS pandemic beginning in that country in the 1990s four decades later (Myer et al. 2003).

for the development of social epidemiology. Thus Rose articulated a central question: should the exploration of disease in populations with a view to effecting prevention and improvement, focus only on individual biological and social determinants, or should at least equal weight be given to the characteristics of the population? These ideas are further amplified in Chapters 24 and 25 of this book.

Social epidemiology in the latter twentieth century, post 1980

Toward the end of the twentieth century, research attention to the societal determinants of health had increased sharply. With this, social epidemiology grew into a discrete specialist area within the discipline. Since 1990, several notable texts have integrated epidemiological and sociological perspectives on population health. Berkman and Kawachi (2000) published the first volume specific to social epidemiology. In this period of an evolving social perspective in epidemiology, research has addressed thematic areas that go beyond descriptions of the general association between increasing socioeconomic status and improved health, and that seek better understanding both of the disparities in health among ethnic groups, and of the effect of community characteristics on individual health (Diez-Roux et al. 1997; House 2002; Winkleby et al. 2006; Masi et al. 2007).

In Britain throughout this period, also within an evolving discipline of social epidemiology, the strong research emphasis on socioeconomic factors as health determinants continued. This sustained interest gave rise to a series of government commissions investigating national inequalities in health. In Britain, both the 1980 Black Report (Gray 1982; Black et al. 1982) and the 1998 Acheson Report (Acheson et al. 1998) pointed to the persistence, despite steady economic growth, of socioeconomic gradients in morbidity and mortality. These inquiries drew heavily on epidemiological analyses showing how social class is inversely related to morbidity and mortality across a range of causes, and secured the place of social epidemiologists within the domain public health.

The Whitehall studies, developed by Michael Marmot and Geoffrey Rose, exemplified the situation. They selected a cohort of British civil servants for their research enterprise with the aim of analysing the relation of social class distributions to health. The highly organized structure of the Civil Service neatly and formally categorized the work and social positions of civil servants into a hierarchical gradient. The initial Whitehall study (Marmot et al. 1984) reported a risk of death in the lowest grade of civil servants roughly three times that in the highest grade. Cause-specific mortality manifested the same gradients as reflected, for example, in cardiovascular death rates. Allowance for the effects of differences in smoking, blood pressure,

and plasma cholesterol explained only part of these differences. On further follow-up, such psychosocial occupational factors as low job control at the lower civil service grades proved to be strong predictors of substantial risk. Indeed, they were better predictors of the risk of coronary heart disease than many established risk factors. The Whitehall studies are particularly important because they show marked gradations in health even in a hierarchy among office workers rendered relatively homogenous by the site and demands of their work. The results make plain the impact on health throughout a refined social gradient well removed from frank poverty (Marmot and Brunner 2005). Comparable US studies include the work, for instance, of Link and Phelan and others in the United States (see below).

Many epidemiologists have viewed social and economic factors temporally proximate to the onset of disease as thus potentially causal, but also as relatively static phenomena. By contrast social inequalities, as they change or persist during the life course, have been a primary concern in understanding the causes of adult disease and death. British epidemiologists especially, taking a longitudinal view of development, have sought to isolate the potential impact of socioeconomic conditions in early life on disease during adulthood.

Much of this work arose in the early social medicine era, beginning with the 1946 National Birth Cohort (Wadsworth et al. 2006) followed thereafter through life by J.W.B. Douglas, and succeeded by the 1970 British Cohort Study, also followed from birth by Neville Butler (Elliott and Shepherd 2006). The force and the effects of socioeconomic factors are plainly evident across many aspects of child development (Stein and Susser 1969b), but none so clear-cut as in mental performance. In the 1920s retarded mental ability was first shown by E.O. Lewis (Chapter 21) to be of two major types. One type, severe and recognizable in very young children, was distributed equally, regardless of social position of the families and often accompanied by a range of physical disabilities. In time, some of these conditions were established as genetic in origin. In a second type of mild mental retardation, first noted by Lewis and Penrose, and later studied in some detail in several settings and countries by ourselves and others, show a clear connection with the social status of the families (Stein and Susser 1969a). Mild mental retardation is defined primarily in the quantitative terms of intelligence test score rather than in qualitative terms. The accompanying table (see Table 20.2), taken from a study of all 11-year-old children in Warsaw in 1963, shows their performance on written tests, stratified by a global index of parental occupation and education (Firkowska et al. 1978). As with the Whitehall data of Marmot, for social epidemiologists these associations pose questions about the widespread impact of social status. In societies across the world, whether rich or poor in resources, similar data sets have shown similar associations.

Table 20.2 Mean Score on Raven Test by Global Index of Parental Occupation and Education in 13,108 Children of the 1963 Birth Cohort Living in Warsaw (Data Incomplete for 1,130 Children).[1]

Global index	Score	Standard Deviation	*N*
0	85.85	15.56	103
1	90.63	16.88	221
2	91.68	15.66	510
3	93.63	15.45	1278
4	96.02	14.89	1976
5	97.14	14.90	1116
6	100.15	14.23	1547
7	101.86	13.60	895
8	102.35	13.44	1618
9	104.27	13.01	608
10	105.42	12.97	1362
11	105.20	12.82	432
12	108.07	12.08	1442

[1] Reprinted from Firkowska AN, Ostrowska A, Sokolowska M, Stein Z, Susser M, Wald I (1978). Cognitive Development and Social Policy. *Science* 200:1357–1362. Copyright (1978) with permission from the American Association for the Advancement of Science.

These Polish findings, moreover, were later shown to be highly relevant and indeed predictive, of the subsequent careers of these children in terms of their further education, their socioeconomic status, and even their sense of life satisfaction (Firkowska-Mankiewicz 2002; Stein and Susser 2004).

Several decades later, George Davey Smith, Diana Kuh, David Barker, and others have turned to mining cohort data through different stages of the life course. These studies have sought to uncover links between social factors—more especially social class—with causes of adult morbidity and mortality, as well as with potential markers for cardiovascular disease (Barker et al. 1992; Kuh and Ben-Shlomo 1997; Davey Smith et al. 1997; Davey Smith et al. 1998; Ben-Shlomo and Kuh 2002; Kuh and Ben-Shlomo 2004).

These life course studies tend to be considered separately from those longitudinal studies that have followed populations exposed in early life to particular events or life experiences as, for instance, famine, or irradiation, or placement outside the family of origin, or family breakup, or loss of a parent. Both approaches have yielded valuable social and psychological insights. But in life course studies the emphasis has often been more explicitly broadly social.

In the United States since the 1980s, selecting again now for a contrasting situation, social epidemiology has had a different trajectory. For some years before and even later, social epidemiology remained a somewhat marginal and even unspecified subspecialty. During this period social scientists

continued to make contributions to the understanding of forms of practice and of social determinants of health, but the vast majority of senior epidemiologists were medically trained, with a focus on the pursuit of individual behaviors or exposures as risk factors for disease. The concept that social and economic conditions were significant determinants of health reached its nadir when one prominent and generally excellent epidemiology text took the position that socioeconomic position, rather than belonging in any causal chain leading to disease, was confined to that of a confounder of exposure/disease associations (Rothman 1986). Nevertheless in recent years, the study of socioeconomic factors among the causes and distribution of disease has gradually secured its place in American epidemiology. The rise of social epidemiology has been accompanied by developments on a variety of fronts.

Several sociologists, including Link and Phelan, Wilkinson, House, Gottfredson, and others, have argued for a fundamental cause to explain inequalities of health across social categories of wealth and poverty. Thus Link and Phelan drew attention to the universality of this phenomenon across time and across societies (Link and Phelan 1995). Across societies, moreover, they pointed out that the less remarkable the social inequality, the less the inequalities of health status. Likewise, the less preventable are mortality and morbidity from particular diseases, the less is the inequality across social groups. Thus those societies that best spread both knowledge of health risks and disease and access to services across the range of social groups, the less marked are the disparities in health.

Further, it is manifestly evident that both knowledge about risks and the optimal responses to them, as well as about access to resources, will generally be connected in turn to educational attainment. As demonstrated earlier in this chapter, in the Warsaw study the monotonic relationship regularly found between childhood intelligence and both parental education and occupation matches the relationships between occupational status and health. Not surprisingly then, it has been proposed that intelligence (presumably of individuals) would provide a broad set of resources relevant to health maintenance (Gottfredson 2004). On the contrary, Link and Phelan argue that education rather than intelligence alone would account for any relationship with useful health knowledge. In our view, among individuals and certainly among groups, intelligence and educational attainment are closely connected. At the same time, we must acknowledge that we do not understand the causes of the relationship of social factors to intelligence any better than we understand the causes of their relation to health.

In the United States, ethnicity has attracted close attention as a significant factor in the distribution of health. Compelling epidemiological evidence shows that rates of morbidity and mortality both overall, and for several specific conditions, are higher among African-Americans and other minority

groups. Adjustment for such markers of socioeconomic position as employment, income, and education does not dispel the imbalance. Although these ethnic disadvantages are widely acknowledged, the explanations for them generate contention. It is evident that contributory factors include illiteracy, poor education, ignorance, and poverty. Nancy Krieger and others, however, take a narrower contextual view of racial disadvantage in health and social discrimination—or, more directly, crass racism—to be the principal causal factor (Krieger 1999). In this light, a sustained socioeconomic differential in health is not merely a confounder requiring analytic adjustment, but the vehicle for the malign effects of race discrimination. Further study will be needed to establish such an hypothesis as fact.

A second recent contribution of social epidemiologists in the United States has been to advance understanding of the ways in which social and economic conditions influence health outcomes at such different levels of social organization as the individual, household, community, or society. These and other epidemiologic analyses of social determinants of health across different levels of organization echo the work of Durkheim on suicide a century before (Durkheim 1897). Sampson and colleagues, for example, demonstrated in a seminal paper that the social features of given neighborhoods helped explain levels of violent crime, independent of the characteristics of the individuals living there (Sampson et al. 1997). Populations as entities may themselves have features that shape health outcomes beyond the aggregated characteristics of the individuals comprising them (Rose 2001). Group-level variables serve as an important vehicle for social epidemiologists (Susser 1973).

In the design and analysis of research, the differing properties of group-level and individual-level variables demand alertness if the distinctions between them are to be maintained. Multilevel analysis poses new challenges to epidemiological thinking (Diez-Roux 2000). Thus multilevel analysis takes into account the differing patterns of relations among variables that may emerge at successive levels of organization, from the simplest to the most complicated (Duncan et al. 1998; Blakely and Woodward 2000; Diez-Roux and Aiello 2005; Susser E et al. 2006). At the simplest level, individuals may be taken alone, or nested within households. Households are nested within communities, and communities within still larger population conglomerates. The analysis of such data structures has required the application of such complex analytic approaches as hierarchical modeling, generalized estimating equations, and mixed effects modeling. In large part, these derive from statistical methods developed for the analytic needs of the social sciences. Such statistical tools have enabled a body of research that shows how communities or neighborhoods encapsulate a range of structural properties that influence individual health. Approaches to unraveling the associations

between the character of social groupings and health are at the cutting edge of social epidemiology.

Social Gradients and their Effects

Theoretical and methodological challenges face social epidemiology. The most general challenges relate to the associations of wealth and poverty with health and disease. In this chapter, starting with mental disorders as an instance, we have sketched the growing interest in these associations across the whole gamut of morbidity and mortality, both within societies and across societies and across time. These are questions crucial to the public health. Should the quest for causes be for a single unitary factor or, as seems more appropriate , should it seek out groupings of such environmental and personal disadvantages that seem to be linked and overlapping: for example, poverty and discrimination, which undermine education and socioeconomic advantage, and which in turn limit appreciation and access to such facilities for prevention and care as are provided within the particular societies.

Conclusion

During much of the later twentieth century in the United States, the study of effects of societal and economic conditions on health was seldom in the mainstream of epidemiology. In the 1990s, however, such issues regained the attention of the discipline. In Britain, beginning in the post–World War II period under the umbrella of social medicine, the social determinants of health had received sustained attention from epidemiologists. In America at that time, it had chiefly been sociologists and mental health researchers who drove research in an area seen to fall outside the necessary attention to infectious and chronic disease.

In general the practice of research in social epidemiology, both historically and in the present has been a fundamentally interdisciplinary undertaking that balances the inputs of medicine, sociology, anthropology, economics, demography, and other social sciences. Perspectives from social epidemiology have played a considerable role in pressing epidemiologists of all kinds to consider how social contexts influence health outcomes, and to examine in turn the potential for societal interventions to modify or reduce morbidity and mortality. However this renaissance of the study of societal determinants has raised important questions about how the study of health relates to the broader discipline. Is social epidemiology best considered a subdiscipline within epidemiology, alongside genetic epidemiology, cancer epidemiology

and infectious disease epidemiology? Or is understanding the ubiquity of social and economic conditions in shaping morbidity and mortality a concept that should influence epidemiology generally, and with which all epidemiologists should be familiar. While the answer to these questions remain unclear, epidemiologists working in various specialized areas have paid attention to socioeconomic determinants. Such a perspective, rather than simply defining a specialty field, can best be seen as relevant to the discipline as a whole.

References

Abramson JH (1974). *Survey Methods in Community Medicine.* Edinburgh, UK: Churchill Livingstone.

Acheson D, Great B, Her Majesty's Stationery Office (1998). *Independent Inquiry into Inequalities in Health Report.* London, UK: Stationery Office.

Baer HA, Singer M, Susser I (1997). *Medical Anthropology and the World System: A Critical Perspective.* Westport, CT: Bergin & Garvey. (2003) 2nd ed. Westport, CT: Praeger.

Barker DJP, Medical Research Council (Great Britain), Environmental EU (1992). *Fetal and Infant Origins of Adult Disease: Papers.* London, UK: British Medical Journal.

Ben-Shlomo Y, Kuh D (2002). A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *Int. J Epidemiol.* 31:285–293.

Berkman L, Kawachi I, eds. (2000). *Social Epidemiology.* New York, NY: Oxford University Press.

Berkman LF, Syme SL (1979). Social networks, host resistance, and mortality: a nine-year follow-up study of alameda county residents. *Am J Epidemiol* 109:186–204.

Black D, Townsend P, Davidson N (1982). *Inequalities in Health: The Black Report.* Harmondsworth, Middlesex, UK: Penguin Books.

Blakely TA, Woodward AJ (2000). Ecological effects in multi-level studies. *J Epidemiol Community Health* 54:367–374.

Cassel J (1964). Social science theory as a source of hypotheses in epidemiological research. *Am J Public Health Nations Health* 54:1482–1488.

Cassel J (1976). The contribution of the social environment to host resistance: the Fourth Wade Hampton Frost Lecture. *Am J Epidemiol* 104:107–123.

Davey Smith G, Hart C, Blane D, Gillis C, Hawthorne V (1997). lifetime socioeconomic position and mortality: prospective observational study. *BMJ* 314:547–552.

Davey Smith G, Hart C, Blane D, Hole D (1998). Adverse socioeconomic conditions in childhood and cause specific adult mortality: prospective observational study. *BMJ* 316:1631–1635.

Diez-Roux AV (2000). Multilevel analysis in public health research. *Annu Rev Public Health* 21:171–192.

Diez-Roux AV, Aiello AE (2005). Multilevel analysis of infectious diseases. *J Infect Dis* 191 Suppl 1:S25–S33.

Diez-Roux AV, Nieto FJ, Muntaner C, Tyroler HA, Comstock GW, Shahar E, Cooper LS, Watson RL, Szklo M (1997). Neighborhood environments and coronary heart disease: a multilevel analysis. *Am J Epidemiol* 146:48–63.

Doll R, Hill AB (1952). A study of the aetiology of carcinoma of the lung. *BMJ* 1:1271–1286.

Duncan C, Jones K, Moon G (1998). Context, composition and heterogeneity: using multilevel models in health research. *Soc Sci Med* 46:97–117.

Durkheim E (1897). *Le Suicide.* See New York: Free Press, trans. by J. Spalding and G. Simpson, 1987.

Elliott J, Shepherd P (2006). Cohort profile: 1970 British Birth Cohort (BCS70). *Int J Epidemiol* 35:836–843.

Faris R, Dunham H (1939). *Mental Disorders in Urban Areas: an Ecological Study of Schizophrenia and Other Psychoses.* New York: Hafner Publishing Co.

Firkowska AN, Ostrowska A, Sokolowska M, Stein Z, Susser M, Wald I (1978). Cognitive development and social policy. *Science* 200:1357–1362.

Firkowska-Mankiewicz A (2002). *Intelligence and Success in Life.* Warsaw: IFiS Publishers.

Goldberger J, Wheeler GA, Sydenstricker E (1920). A Study of the relation of family income and other economic factors to pellagra incidence in seven cotton-mill villages of South Carolina in 1916. *Public Health Reports* 46:2673–2714.

Gottfredson LS (2004). Intelligence: Is it the epidemiologists' elusive "Fundamental Cause" of social class inequalities in health? *J Pers Soc Psychol* 86:174–199.

Gray AM (1982). Inequalities in health. The Black Report: a summary and comment. *Int J Health Serv* 12:349–380.

Henderson LJ (1970). *On the Social System: Selected Writings.* Chicago, IL: University of Chicago Press.

Hollingshead A, Redlich F (1958). *Social Class and Mental Illness.* New York, NY: John Wiley and Sons, Inc.

House JS (2002). Understanding social factors and inequalities in health: 20th century progress and 21st century prospects. *J Health Soc Behav* 43:125–142.

Kark SL (2003). The social pathology of syphilis in Africans. 1949 (Reprinted). *Int J Epidemiol* 32:181–186.

Kark S (1949). The social pathology of syphilis in Africans. *S Afr Med J* 23:77–84.

Kasl SV, Cobb S (1980). The experience of losing a job: some effects on cardiovascular functioning. *Psychother Psychosom* 34:88–109.

Kasl SV, Cobb S, Brooks GW (1968). Changes in serum uric acid and cholesterol levels in men undergoing job loss. *JAMA* 206:1500–1507.

Kermack WO, McKendrick AG, McKinley PL (1934). Death rates in Great Britain and Sweden: expression of specific mortality rates as products of two factors, and some consequences. *J Hyg* 433–451.

Krieger N (1999). Embodying inequality: a review of concepts, measures, and methods for studying health consequences of discrimination. *Int J Health Serv* 29:295–352.

Kuh D, Ben-Shlomo Y (1997). *A Life Course Approach to Chronic Disease Epidemiology.* Oxford, UK: Oxford University Press.

Kuh D, Ben-Shlomo Y (2004). *A Life Course Approach to Chronic Disease Epidemiology.* New York, NY: Oxford University Press.

Link BG, Phelan J (1995). Social conditions as fundamental causes of disease. *J Health Soc Behav* Spec No: 80–94.

MacMahon B, Pugh T, Ipsen J (1960). *Epidemiologic Methods.* Boston, MA: Little Brown.

Marmot M, Brunner E (2005). Cohort profile: the Whitehall II Study. *Int J Epidemiol* 34:251–256.

Marmot MG, Shipley MJ, Rose G (1984). Inequalities in death—specific explanations of a general pattern? *Lancet* 1:1003–1006.

Marmot MG, Syme SL (1976). Acculturation and coronary heart disease in Japanese-Americans. *Am J Epidemiol* 104:225–247.

Masi CM, Hawkley LC, Piotrowski ZH, Pickett KE (2007). Neighborhood economic disadvantage, violent crime, group density, and pregnancy outcomes in a diverse, urban population. *Soc Sci Med* 65(12):2440–2457.

Morris JN (1957). *Uses of Epidemiology.* Edinburgh, UK: Livingstone.

Myer L, Morroni C, Susser ES (2003). Commentary: the social pathology of the HIV/AIDS pandemic. *Int J Epidemiol* 32:189–192.

Paul BD (1955). *Health, Culture, and Community Case Studies of Public Reactions to Health Programs.* New York, NY: Russell Sage Foundation.

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

Rose G (1992). *The Strategy of Preventive Medicine*. New York, NY: Oxford University Press.

Rose G (2001). Sick individuals and sick populations. 1985. *Bull World Health Organ* 79:990–996.

Rothman KJ (1986). *Modern Epidemiology.* Boston, MA: Little, Brown.

Sampson RJ, Raudenbush SW, Earls F (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science* 277:918–924.

Stein ZA, Susser MW (1969a). Mild mental subnormality: social and epidemiological studies. In Redlich, ed. *Social Psychiatry.* Baltimore: Williams and Wilkins, pp: 62–85.

Stein ZA, Susser MW (1969b). the social dimensions of a symptom: a sociomedical study of enuresis. In Sokolowska et al., eds. *Sociological Studies in Medicine*. Warsaw: Books and Knowledge. Reprint from *Soc Sci Med* 1967;1:183–201.

Stein ZA and Susser M (2004). Essay Review. Measured intelligence in childhood, social class and adult outcomes across Poland's sociopolitical transitions, 1945–1995. *Int J of Epi*, 33:227–230.

Susser E, Schwartz S, Morabia A, Bromet EJ (2006). *Psychiatric Epidemiology: Searching for the Causes of Mental Disorders.* New York, NY: Oxford University Press.

Susser I, Patterson TC (2001). *Cultural Diversity in the United States: A Critical Reader.* Malden, Mass: Blackwell.

Susser M (1957). African Township; a sociomedical study. *Med World* 86:385–397.

Susser MW, Watson W (1962). *Sociology in Medicine*. London: Oxford University Press.

Susser M (1968). *Community Psychiatry: Epidemiology and Social Themes.* New York, NY: Random House.

Susser M (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* New York, NY: Oxford University Press.

Susser M (1993). A South African Odyssey in community health: a memoir of the impact of the teachings of Sidney Kark. *Am J Public Health* 83:1039–1042.

Susser M (2006). A personal history: social medicine in a South African setting, 1952-5. Part 2: Social medicine as a calling: ups, downs, and politics in Alexandra Township. *J Epidemiol. Community Health* 60:662–668.

Susser M and Myer L (2007). Social epidemiology. In Holland WW, Olsen J, Florey CDV, eds. *The Development of Modern Epidemiology: Personal Reports from those Who Were There.* New York: Oxford University Press, pp. 207–218.

Susser M and Stein Z (1962). Civilisation and peptic ulcer. *Lancet* 1:115–119. Reprinted in 2002 *Int J Epidemiol* 31:13–17.

Susser M, Stein Z, Cormack M, Hathorn M (1955). Medical care in a South African Township. *Lancet* 268:912–915.

Susser MW, Watson W (1962). *Sociology in Medicine*. London, UK: Oxford University Press.

Syme SL, Reeder LC (1967). Social stress and cardiovascular disease. *Milbank Memorial Fund Quarterly* 45.

Wadsworth M, Kuh D, Richards M, Hardy R (2006). Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol* 35:49–54.

Winkleby M, Cubbin C, Ahn D (2006). Effect of cross-level interaction between individual and neighborhood socioeconomic status on adult mortality rates. *Am J Public Health* 96:2145–2153.

21

Epidemiologists and Geneticists: A Developing Détente

The new talents drawn into public health and epidemiology immediately following World War II tended to be, like their predecessors, socially conscious individuals. They sought especially the means of preventing those social and physical conditions that might prove amenable to change, as in the developing world with poverty, and in the developed world with industrial and environmental pollutants and irradiation. By contrast, like earlier concepts such as the Hippocratic notion of humoral constitution, or the concept of "host," as it evolved in the late nineteenth and early twentieth centuries, genetics was seen as dealing with fixed and unchangeable entities. Moreover, misconstrued genetics carried with it the recent evil history of its political applications by Nazism as a central criterion for the elimination of such "undesirables" as the mentally disordered or mentally handicapped, as well as such *untermenschen* as Jews and other non-Aryans. All too degrading horrors followed from the race-based Nazi policies that sought to exterminate Jews, Gypsies, and the mentally or physically handicapped. This unregenerate theory stained and burdened social Darwinism and eugenics for decades thereafter. As epidemiologists sought to follow the mantra of preventing disease and testing treatments, they came to accept—if somewhat slowly—that both gene and environment are central contributors to disorders

of health. Separately and together, both disciplines can provide a basis for prevention and public health action.

Using the clinical entity of mental retardation, we trace the paradigmatic changes occurring soon after World War II that introduced genetics into epidemiology and influenced public health practices. Four different classes of disorder, each underlying a distinctive form of mental retardation, will serve to illustrate this theme and this time. For each class the potential for primary and secondary prevention exists. Epidemiologic studies have furthered understanding of these quite different genetic phenomena. One class of disorders arises with chromosomal anomalies of number or structure, as in Down syndrome. A second class arises with genetic disorders of metabolism, as in phenylketonuria. A third class probably arises from the interaction of maternal and/or embryonic genotypes with maternal environment, as in neural tube defects. A fourth class, mild mental retardation, probably arises from the interaction of familial and psychosocial environment with a postulated genetic or polygenic trait. This last-mentioned condition, discussed elsewhere by the present authors and others (Stein and Susser 1970; Stromme and Magnus 2000; Croen et al. 2001; Leonard et al. 2003; Yaqoob et al. 2004; Durkin et al. 2007), has already found mention in the preceding chapter.

Mental Retardation

Epidemiological studies of mental retardation, as distinct from individual or familial studies, were undertaken early on in Britain. In 1929, E.O. Lewis and A.H. Wood (the latter a civil servant at the Board of Education) reported on what was probably the first community-based sample survey of mental retardation (Wood et al. 1929). Lewis executed the study on commission from a joint committee of the Board of Education and Board of Control, in six areas of England and Wales. Here the term "mental retardation," variously defined and applied over the past century, referred specifically to deficient intelligence.

Lewis's report (which became known as the Wood Report, since Wood chaired the committee) stimulated a larger, more detailed study carried out over seven years (1931–1938), among the 1,000 mentally disabled residents confined in an institution in Colchester, England. Lionel Penrose (1898–1972), chosen to execute the study, proved to be an outstanding scientist. His description of the Colchester residents, published by the Ministry of Health in 1938, remains a landmark of thoroughness (Penrose 1938). The survey provides a valuable cross-section of what could be known about mental retardation at that time, and it serves as an historic yardstick against

which to measure the incidence and prevalence of its heterogeneous forms. He attributed several conditions to genetic disorders.[1] In 1960 in London, at the first conference of the newly founded Association for the Scientific Society for the Study of Mental Deficiency (later renamed the International Association and known as the IASSMD), Penrose gave the keynote address (Richards 1962). The meeting coincided with a great ferment in biology and genetics: Crick and Watson's publication on the role of DNA in helical gene structure had appeared in 1953 (see Watson and Crick [2003] for a recent reprint in *Nature*); the number of chromosomes in the human genome had been established by Joe-Hin Tjio and Albert Levan in 1955 and published in 1956; and in 1959 the first description of a clinical condition associated with an extra chromosome appeared, namely, Down syndrome (Lejeune J et al. 1959a; Lejeune J et al. 1959b). In his address, Penrose made the bold claim that the study of mentally retarded populations would yield clues to the understanding of human genetics. Certainly the three quite different conditions described below are exemplary in advancing both genetic understanding and potential avenues for prevention and public health.

Down Syndrome[2]

From the outset, Penrose took a great interest in what was then described as "Mongolian idiocy" (a term based on obsolete mid-nineteenth-century classifications of racial types and characteristics). In his Colchester survey, he had carefully categorized the key clinical phenomena associated with this condition and he confirmed the long-noticed association with maternal age.[3] On the basis of his analysis of the blood groups of Down syndrome subjects, he dismissed the then current idea of any connection with Mongolians. He also ruled out any close associations with late birth order, with the duration of the interval between second last and last birth, and with paternal age.[4]

[1] The term "gene," as mentioned in Chapter 13, referred explicitly to an intergenerational entity, embodying characteristics that were inherited. In time, the term "gene" took on a morphological form, a specific component of the cell nucleus (a sequence of nucleotides). Although the concept of inheritance was maintained, it became understood that environmental exposure, for instance irradiation, could damage somatic cells. Such changes might persist as precursors of cancerous growths, for example, or might be repaired, hence benign; if germ cells were affected, offspring might be affected. At the time of writing, the definition of a gene becomes even further removed from morphology, although it does seem to retain the notion of inheritance.

[2] Renamed for the English physician, John Langdon Haydon Down in 1866. In 1867, Down wrote, "the proclivity resides in the germ cell or sperm, as a result of gradual degeneration." Later writers prefer TRS21 for Trisomy 21 (Roubertoux and Kerdelhue 2006).

[3] G.E. Shuttleworth first described this association in 1886 (Carter 2002).

[4] Although maternal age at the time of conception has been shown to be the most significant risk factor for nondisjunction of chromosome 21, paternal age, health, and environmental factors are also undergoing investigation (Sherman et al. 2005).

He concluded that intrauterine environment was not the cause, in view of the discordance for Down syndrome among nonidentical twins, as well as the fact that among older mothers who had borne a Down syndrome infant, prior births were seldom affected. He surmised instead that some change must have taken place in the aging ovum, although he did not favor the extra chromosome theory of Waardenburg and Bleyer, described below.

Penrose was preceded in his hypothesis of the aging ovum by the Dutch ophthalmologist-geneticist P.J. Waardenburg. In 1932, in his monograph on the human eye, he put forward the specific proposition that nondisjunction of a chromosome in the aging ovum might occur, giving rise either to monosomy or trisomy, or to translocation or deletion (Waardenburg 1932); see also Allen (1974) and Carter (2002). His supposition followed from the knowledge that many different aberrant phenomena (eyes, shape of head, tongue, hand, heart, retardation, and more) were associated with this syndrome, an unlikely result of an error of polymorphism involving a single gene (a phenomenon already understood from Garrod's early work on alkaptonuria). In 1934, Adrien Bleyer, a thoughtful American pediatrician, independently put forward a proposition similar to Waardenburg's, but focused more specifically on trisomy. Bleyer was impressed by the range of "distortions" in those affected and thought that these Down-related anomalies were too diverse and numerous to attribute the condition to intrauterine phenomena (Bleyer 1934; Wolstenholme et al. 1967); (see discussion by Waardenburg (1932); Carter [2002]).

Nondisjunction, a failure in the aging ovum, is precisely the mechanism to which later studies of Trisomy 21 in older women had led Penrose. But Penrose did not stop there. From close study of the mother's age in a large series of Down syndrome cases, he discerned two quite distinct age relationships: first, the highest risk (as others had observed before him) was to older women. Second, and more rarely, isolated or recurrent Down syndrome births sometimes occurred in women younger than average. Thus Penrose's initial supposition of two separable causal processes accounting for the Down's anomaly has proved to be correct (Carter 2002); also see Roubertoux and Kerdelhue (2006). Recurrent Down syndrome births in younger women have turned out often not to be simple trisomies, but unbalanced translocation anomalies of the chromosomes (either inherited from translocations in a parent, in which the parental chromosome set was balanced overall but unbalanced in the germ cell; or arising unbalanced *ab initia* during mitosis in the fertilized egg). Another phenomenon associated with Down syndrome, but not with age of the mother, is mosaicism, a post-fertilization phenomenon of nondisjunction affecting only a proportion of cells.

The new knowledge of these chromosomal anomalies made possible the prevention of trisomic births. A first approach was to publicize the higher

risk of an affected birth for older women. A second technical approach arose once the technique of amniocentesis (and later, chorionic villus sampling) became available to older pregnant women and could be applied selectively for terminating an affected pregnancy. These diagnostic approaches were useful only given the social acceptance of abortion. The marked decrease in the incidence of Down syndrome births in several societies indicates the general acceptance of such procedures.

The risk of a trisomy 21 birth in women in their late thirties is about one in 150. Although the risk in younger women is very much less—perhaps only one in 1,000—the much larger proportion of births overall among them results in larger numbers of trisomy 21 births contributed by maternities at younger ages (Huether et al. 1998). It follows that prenatal judgments of the risk of a Down syndrome birth based solely on chronological age of the mother alone is neither specific enough nor sensitive enough to serve as an indication for prenatal diagnostic procedures. For individual younger women at lower risk, even though prenatal procedures provide virtually foolproof diagnosis of a trisomic conception, they may not choose to undergo them: neither procedure is completely without some risk that a woman might inadvertently terminate a normal conceptus.

Fortunately, since invasive diagnostic procedures were first introduced, prenatal screening techniques have been much refined. In an affected fetus in the prenatal period, ultrasonography can distinguish such markers as unusual thickness at the base of the neck (nuchal translucency). Markers in maternal serum (such as pregnancy-associated plasma protein A and the free beta subunit of human chorionic gonadotrophin) also serve as noninvasive indicators of normality or otherwise. Thus, by 2005, predictions of the normality of the conceptus at 11 weeks after conception had become much more reliable than formerly. Although chorionic villus sampling in the first trimester is an invasive procedure, it is safe for the mother and generally safe for the pregnancy, and thus can be carried out for women positive on any of these three indicators. In practice, by the end of the first trimester, high sensitivity rates can be achieved for diagnostic tests, although not without some few false positive cases. Thus relatively refined and accurate sophisticated screening procedures have become available at a woman's choice, depending on her age and preferences, the cost of the tests and the expertise and experience of the service (Malone et al. 2005; Breathnach and Malone 2007).

In parallel with advances in technical understanding, a succession of novel kinds of intervention have given rise to such new therapeutic possibilities as social and psychosocial management, and public health action. In the event that a Down birth does occur, better control and effective treatment of infection and emphasis on social stimulation in the young child has extended the life expectancy and improved the societal and mental capacities of affected individuals.

In developed societies, distinctly fewer affected infants are born than before, and the lives of those who are born are now generally more fulfilling than previously (Hayes and Batshaw 1993; Smith 2001; Durkin et al. 2007).

As Penrose predicted, Down syndrome studies have led to discoveries relating to quite other conditions. A range of other chromosomal anomalies can now be recognized between conception and birth, and much more is understood about trisomy in general. For example, trisomic forms of almost all the other chromosomes have now been recognized, all of which occur more commonly among the conceptuses of older women. Excepting one third of trisomy 21 conceptions, however, few trisomies come to term. Recurrent anomalies such as chromosomal translocations sometimes account for persistent infertility and repeated miscarriages. The understanding gained about the chromosomes also informs the clinical techniques of assisted reproduction (Kline et al. 1989).

As time passed, we have learned that the diminished mental capacities of individuals with trisomy 21 who survive into middle age deteriorate further, and at postmortem show evidence of Alzheimer's disease (supposedly, this is connected with a double dose of a gene on Chromosome 21 known to be related to Alzheimer's disease). Penrose had earlier noted the connection with clinical Alzheimer's disease (Ohara 1972). In fact, his young co-worker Peter Ohara was the first to describe the fibrillary tangles in the brains of deceased Down syndrome individuals.

We may anticipate a future where first trimester maternal screening renders a safe, accurate, and efficient prediction of the outcome of a pregnancy.[5] This would clearly make diagnosis more widely available. Despite what is feasible and acceptable in well-equipped prenatal centers, however, screening with such precision is unlikely to become widely available anytime soon across the world. In many low-resource countries and among some other communities where unplanned pregnancies are the rule, there will always be children born to older women. As infant mortality declines, with routine immunization and treatment of infections, more affected offspring in these localities will survive. By contrast, in the US and Europe, a proportionate increase in births to older mothers will usually reflect the voluntary deferment of childbearing for economic and professional reasons. Most such women and their partners will doubtless have been deferring pregnancy by contraception for some years, the advent of a pregnancy will have been planned, and they are likely to avail themselves of the optimal range of available screening measures. Hence the incidence, the age-specific prevalence and the quality of life

[5] A recent review suggests that first trimester screening followed by second trimester screening (contingent upon first trimester results) provides optimal results, though the field continues to evolve rapidly (Breathnach and Malone 2007).

of trisomy 21 individuals will vary over time and place, depending as much on demographic and social factors as on advances in genetics.

Phenylketonuria

A somewhat different march of scientific, epidemiologic, and public health action relate to phenylketonuria (PKU). In 1934, a Norwegian physician and biochemist, Asbjørn Følling, confirmed the tell-tale odor of phenylalinemia in the urine of two severely mentally retarded children (reported to him by their parent) and the pallid coloring of their hair and skin. He identified the nature of the abnormality in the urine and confirmed similar findings among eight other mentally retarded persons (Følling 1934a; Følling 1934b; Christ 2003). Subsequently, following the description by Følling, Penrose found two cases in Colchester. In residential facilities for the mentally retarded, the frequency of PKU at that time ranged from 1% to 5%.

By 1934, it was readily understood that the mating of two otherwise normal parents, both carriers of the same recessive gene, could produce such a condition in the offspring. As early as 1905, Archibard E. Garrod had first recognized another such circumstance: alkaptonuria, which follows upon the transmission of a recessive gene to an affected child from each of its unaffected parents (Garrod 1908). Garrod, in correspondence with Bateson, had noticed the additional risk to siblings and to offspring of cousin marriages. Følling was well aware of Garrod's work on the transmission of recessive genes, and confirmed the pattern of inheritance for PKU. Such a pattern sharply limits the means for preventing the conception or birth of children with PKU. The inapparent clinical features of recessive genes on both sides ruled out any option of counseling prior to the postpartum recognition of the birth of an affected child. This circumstance left little if any prospect of reducing incidence in the population. In the future, however, a prospect for prevention might well reside in routine testing of both potential parents before conception for the presence of such inimical genes.

An alternative approach to the problem, first proposed 50 years before it was applied, focused on the affected children (Jervis 1937). This approach relied on the diagnosis of the condition in the neonatal period, and then on modifying the diet of the young affected child, in order to minimize or eliminate the toxic by-product of metabolic malfunction. In the 1950s two controlled trials, one in the US and one in the UK, each demonstrated that prevention of the effects of PKU very early in life by beginning the rigorous dietary restriction of food containing phenylalanine was both feasible and efficacious (Bickle et al. 1953; Guthrie and Whitney 1964). In both trials, presumably by inhibiting the adverse effects of phenylalanine on the brain,

the mental performance of the affected child was significantly improved, if not quite to the level of unaffected siblings.

The results of these trials of dietary restriction faced public health with a considerable challenge in realizing its prime task of prevention. The rare homozygotic infants had first to be identified as soon as possible after birth. The parents were then encouraged and enabled to feed the affected child exclusively on a complicated and unnaturally restricted diet. For most parents, the prospect of warding off mental retardation in an otherwise potentially normal child was sufficient motivation to adhere to the stringent dietary requirement. For public health, the main argument for action in so rare a condition was to reduce the anticipated buildup of the prevalence of surviving affected individuals. Despite the severe degree of mental retardation in untreated children, life expectancy was unaffected. But the logistics of early identification were formidable, and for each case the burden of care and the cost was heavy.

Robert Guthrie's discovery of a simple and inexpensive test to screen the urine of neonates for phenylalanine (Guthrie and Susi 1963) greatly enhanced the prospects of prevention. In some respects the test was imperfect; within the first 48 hours of birth it is neither perfectly specific, nor is it perfectly sensitive. Thus later follow-up testing is needed to ensure detection of both false positives and false negatives. Despite such potential weaknesses, routine screening of all the newborn sharply improved the prospects of preventing adverse manifestations of PKU in affected children. In New York State, the site of Guthrie's invention of the screening test, the number of "true positive" PKU births undetected by the test is very small. Failures are recognized as the adverse effects become manifest. Given the normal lifespan of affected children, most of those undetected and untreated survive, and the diagnostic signs of PKU soon become apparent. Since screening was initiated, a major reduction of this form of mental retardation has certainly been evident.

Several lessons can be drawn from the PKU experience. First, newborn screening can be readily implemented, given adequate resources and commitment of public health departments, and is most feasible where the majority of births take place in supervised settings, and are retained for testing in hospitals for at least 48 hours after birth. Second, the questions of the specificity and sensitivity of testing has import not only with regard to the personnel and costs entailed, but also because an unavoidable if small number of "false positive" tests cannot but cause parents deep anxiety. Experience with PKU in particular has taught public health practitioners to undertake screening for any condition with both care and caution, and to do so only where benefit clearly outweighs potential error.[6]

[6] With the advent of the HIV/AIDS epidemic, another consideration related to testing emerged. Infants could be tested, but not without prejudice to the confidential HIV status of the mother:

Genetic conditions often have many facets.[7] Not all the complexities arising with PKU are resolved by early identification followed by appropriate dietary restrictions. In time it emerged that when affected young adults reached normal or near-normal maturity, they tended to broaden their diet with little or no apparent ill effects on their mental condition. The inherited metabolic defect of hyperphenylalaninemia (HPA) persisted nevertheless, and, among women who became pregnant, HPA adversely affected embryonic growth. These offspring, although heterozygotic for PKU, suffered a range of physical and mental handicaps untreatable by postnatal dietary intervention. Thus women with PKU face the alternatives either of foregoing pregnancy, or before conception, returning strictly to the special diet on which they were reared and maintaining it throughout pregnancy.

Other Neonatal Screening Differences

Screening neonates for inherited metabolic disorders began in the 1960s with the Guthrie test for phenylketonuria. This led to a search for additional tests that might yield timely knowledge of the defects and benefit the child and/or the family. Technological advances have made it possible, using a single specimen, to screen simultaneously for 40 or more known disorders, a process known as tandem mass spectrometry or MS/MS. Some of these disorders are very rare, and in many cases diagnosis of the disorder will not lead to any benefit for the child. Not surprisingly, there is little agreement on which disorders should be deemed appropriate and/or cost-effective for public health intervention. Although it is generally accepted that the frequency and natural history of some of these rare anomalies should be studied, there is not general agreement at this time about the extent to which routine screening programs should be designed to address them all (Wilcken 2007; Arn 2007).

Improved understanding of the natural history and postbirth prognosis of the many conditions that can potentially affect births will better equip public health decision makers to determine preventive and prophylactic policies. In one investigation of 29 conditions detectable on screening, only five appeared amenable to improved prognosis with treatment. It has nevertheless

---

she might or might not know her own HIV status, nor want it to be known by others. In much of Africa, India, and the rest of the developing world, the stigma of a sexually transmitted disease can carry with it major adverse consequences for those affected. Once a therapeutic intervention for HIV that could be helpful to the infant was found and could be applied, however, the terms of this debate changed radically.

[7] In practice, even PKU, the best understood of this group of anomalies, is far more complex in terms of its genetic characteristics than was originally understood.

been argued that knowledge of test outcome would be helpful to parents in undertaking further pregnancies. Alternatively, screening the blood of parents who prefer to have the prognostic knowledge yielded by blood tests before undertaking pregnancy and childbearing might be reassessed. Then, should testing indicate a risk from homozygosity in their offspring, prebirth knowledge obtainable either from chorionic villi sampling or from amniocentesis could guide preventive effort. As with trisomy 21, the future holds out both promise and peril.

Neural Tube Defects

Penrose identified neural tube defects as a third cause of mental retardation. Whether he considered it to be a genetic disorder is not clear. Over the intervening half-century, and indeed still today, epidemiologists have explored the potential contributions and interplay between genetic, environmental, and nutritional factors.

Neural tube defects comprise several clinical entities. They arise when the neural tube, formed from a groove in the embryo that gives rise to the brain and spinal column, fails either partially or completely to close as it should normally do. To avoid anomaly, closure must occur by the 28th day after conception, a critical period for human development. Some defects, like *anencephaly*, are incompatible with life; others like *spina bifida* are compatible with life but may be attended by a range of mental and physical disabilities. The interconnections between these various seemingly related neurological anomalies remain somewhat opaque. For example, a child born with either anencephaly or spina bifida raises the risk for the mother of a subsequent affected birth; in a subsequent birth the risk for the recurrence of the same anomaly is greater than for the other.

A variety of associations with this class of anomalies have been found in epidemiologic studies: in the United Kingdom with lower social class and with season of birth; in Holland during the famine conditions late in World War II, with maternal starvation in the first trimester (MacMahon et al. 1953; Carter and Evans 1973; Stein et al. 1975; Nevin et al. 1981; Susser et al. 1998). Particular definable populations give rise to different risks for these neurological anomalies (Berry et al. 1999; Velie et al. 2006). Although upon relocation, as from Europe to Australia, rates of neural tube defects have been found to decline with improved nutritional conditions, the population of origin still remains predictive of the risk.

Many questions remain. Repeated epidemiological observations and controlled trials in several countries—some limited to women who have experienced an affected birth, others population-wide—have firmly established that supplementation of the mother's diet with folic acid before or very soon

after conception reduces the risk of an affected birth (Mills and Signore 2004). The need for supplementation of the maternal diet during the periconceptional period follows from the fact mentioned above, that closure of the embryonic neural tube (the presumed immediate cause of the defect) takes place 25–29 days after conception. This early closure of the neural tube thus typically occurs before a woman becomes aware of a pregnancy. Hence, to reduce incidence, only the provision of folic acid, which is both general and antecedent as, for instance by means of the regular and population-wide addition of folic acid to some staple of the regular diet, is likely to be widely effective. So far, experience with such supplements indicates that while beneficial, they are insufficient wholly to eliminate the condition. In the US and Canada, fortifying bread with folate has probably reduced incidence by about 50% (Mills and Signore 2004).[8] The prescription for an individual woman who has experienced an affected birth would be a routine supplementation tenfold higher than the folate bread supplement provides, thereby reducing the risk of recurrence in her case by an estimated 80%.

From the point of view of the public health, one might tentatively agree with Godfrey Oakley (Oakley Jr. 2002) that "these birth defects are as preventable as polio" (neither condition of course would be totally eliminated). The remedy of folate deficiency in the mother during the periconceptional period has not proved easy to achieve across all societies, but neither has universal immunization against polio. The most comprehensive and the least expensive method to achieve universal coverage of folates in the diet is to fortify aspects of the diet, like bread or cereal as is already done, but there is disagreement about the levels likely to be effective in almost entirely preventing neural tube defects. An alternative is to provide all women of childbearing age who are likely to become pregnant with the supplementary folate-containing vitamin pills before pregnancy is commonly recognized, a recommendation difficult to implement.

Some countries have been reluctant to fortify diet (although for many years the United States, for instance, has fortified salt with iodine)[9] and are also concerned to avoid harm by fortifying below some theoretical high level. At the time of writing several European countries have opted instead for education and prenatal supplements. The United Kingdom and the Netherlands, for example, have each reached about half of all pregnancies in

[8] Since these anomalies (if detected with currently available screening tests in the prenatal period) are often aborted, incidence estimates are difficult to obtain with completeness.

[9] An iodine-deficient diet early in pregnancy was long known to relate to cretinism in the offspring. After a clinical trial demonstrated this outcome to be preventable (Pharoah, Buttfield, Hetzel 1971), fortification of diets world-wide has virtually eliminated an important cause of mental retardation.

this way, a seemingly moderate result compared with the likely effectiveness of fortification. However, there remains the possibility that not all affected pregnancies are folate-sensitive.

The preceding discussion emphasizes the influence of prenatal environmental factors on fetal development, and how they may interact with maternal and fetal genotype. Both within and across populations, the risk to the mother of bearing a child with a neural tube defect varies in ways that indicate the likely presence of genetic factors residing in the mother, the father, or the embryo. Were relevant and dominant genes to be confined to the mother, then, given a stable nutritional environment, all her offspring should be at equal risk (as exemplified above by maternal hyperphenylalaninemia).

With neural tube defect that is clearly not the case. Were the genotype of the father or the embryo to be the only risk factor, then maternal nutrition should not influence the frequency of the defect as it does. One may reasonably infer biological interaction between the genetic constitution of the mother and the environment (in this case, specifically, the availability of folates in her diet). Thus the difference between, say, people of Celtic origin with relatively high incidence and those of African origin with relatively low incidence can be taken to reflect the genetic differences.

However, although nutritional supplementation influences both groups, supplementation alone does not eliminate the differential between them. An experimental observation from China emphasizes this point (Berry et al. 1999). Controlled trials of folic acid supplementation beginning with marriage were set up in two areas, one of high incidence (1/100) and one of low incidence (1/1000) of neural tube anomalies among births. Although reduced incidence of these anomalies were effected in both settings, the differences in incidence persisted.

The partial effectiveness of supplementation clearly indicates that maternal environment is a component cause, but also that absolute prevention requires something more besides. Thus it is likely that both maternal and embryonic genetic constitution carry some risk, which may have consequences above and beyond environmental influences or interactions. The embryo has no independent source of folates at 25–29 days, but even with adequate maternal nourishment, failure of the neural tube to close sometimes occurs nonetheless. Thus the possibility arises that the embryonic gene retains some role unconnected directly to maternal nutrition (or that folic acid is not the only element that could be deficient in the maternal diet).[10] Such alternatives might be further explored in an animal model. For example, an intriguing hint of such a case is found in the curly-tailed mouse, a species at high risk for

[10] Maternal vitamin B12 status, even with adequate folate supplementation, has been proposed as an additional risk factor (Ray et al. 2007).

neural tube defects. The frequencies of these defects are higher or lower, however, depending on both higher or lower supplements of nutrients, and the genotype of the offspring (Seller 1994).

Conclusion

These three now well-known examples were all elaborated in well-resourced countries since World War II. They illustrate the transformation in public health approaches that have occurred, so that genetics has become as important to the epidemiologist as the germ theory of the previous century, and as the concurrent developments of individual risk theory. Methodological and statistical advance can and will be harnessed to further understanding of genetic epidemiology, as they will be to infectious disease epidemiology. Both will benefit the cause of public health.

Nevertheless, it is a somber thought that in the short term none of the three conditions discussed are likely to be preventable in countries of limited resources; yet in each case, more affected infants could survive at least until early childhood, laying additional burdens on families and states. Much has yet to be done to narrow disparities in health and disability.

References

Allen G (1974). Aetiology of Down's syndrome inferred by Waardenburg in 1932. *Nature* 250:436–437.

Arn PH (2007). Newborn screening: current status. *Health Aff (Millwood)* 26:559–566.

Berry RJ, Li Z, Erickson JD et al. (1999). Prevention of neural-tube defects with folic acid in China. China-U.S. collaborative project for neural tube defect prevention. *N Engl J Med* 341:1485–1490.

Bickle H, Gerrard J, Hickmans EM (1953). Influence of phenylalanine intake on phenylketonuria. *Lancet* 2:812.

Bleyer A (1934). Indications that mongoloid imbecility is a gametic mutation of degressive type. *Amer J Dis Child* 47:342–348.

Breathnach FM, Malone FD (2007). Screening for aneuploidy in first and second trimesters: is there an optimal paradigm? *Curr Opin Obstet Gynecol* 19:176–182.

Carter CO, Evans K (1973). Spina bifida and Anencephalus in Greater London. *J Med Genet* 10:209–234.

Carter KC (2002). Early conjectures that Down syndrome is caused by chromosomal nondisjunction. *Bull Hist Med* 76:528–563.

Christ SE (2003). Asbjorn Følling and the discovery of phenylketonuria. *J Hist Neurosci* 12:44–54.

Croen LA, Grether JK, Selvin S (2001). The epidemiology of mental retardation of unknown cause. *Pediatrics* 107:E86.

Durkin, MS, Schupf N, Stein ZA, Susser MW (2007). Childhood cognitive disability. In Wallace, ed. *Wallace/Maxcy-Rosenau-Last Public Health & Preventive Medicine*. New York: McGraw-Hill.

Følling A (1934a). Excretion of phenylpyruvic acid in urine as a metabolic anomaly in connection with imbecility. *Nord Med Tidskr* 8:1054–1059.

Følling A (1934b). Über Ausscheidung Von Phenylbrenztraubensäure in Den Harn Als Stoffwechselanomalie in Verbindung Mit Imbezillitat. *Hoppe-Seyl Z* 227:169–176.

Garrod AE (1908). The Croonian lectures on inborn errors of metabolism. Lecture II Alkaptonuria. *Lancet* 2:73–79.

Guthrie R, Susi A (1963). A simple phenylalanine method of detecting phenylketonuria in large populations of newborn infants. *Pediatrics* 32:338–343.

Guthrie R, Whitney S (1964). *Phenylketonuria Detection in the Newborn Infant As a Routine Hospital Procedure: a Trial of a Phenylalanine Screening Method in 400,000 Infants.* Washington, DC: Children's Bureau Publication 419. U.S. Department of Health, Education and Welfare.

Hayes A, Batshaw ML (1993). Down syndrome. *Pediatr Clin North Am* 40:523–535.

Huether CA, Ivanovich J, Goodwin BS, Krivchenia EL, Hertzberg VS, Edmonds LD, May DS, Priest JH (1998). Maternal age specific risk rate estimates for Down syndrome among live births in Whites and other races from Ohio and Metropolitan Atlanta, 1970–1989. *J Med Genet* 35:482–490.

Jervis G (1937). Phenylpyruvic oligophrenia: introductory study of 50 cases of mental deficiency associated with excretion of phenylpyruvic acid. *Arch Neurol Psychiatry* 38:944.

Kline J, Stein Z, Susser M (1989). *Conception to birth*. New York: Oxford University Press.

Lejeune J, Gautier M, Turpin R (1959a). *C R Acad Sci (Paris)* 248:602–603.

Lejeune J, Turpin R, Gautier M (1959b). Mongolism; a chromosomal disease (Trisomy). *Bull Acad Natl Med* 143:256–265.

Leonard H, Petterson B, Bower C, Sanders R (2003). Prevalence of intellectual disability in Western Australia. *Paediatr Perinat Epidemiol* 17:58–67.

MacMahon B, Pugh TF, Ingalls TH (1953). Anencephalus, spina bifida, and hydrocephalus incidence related to sex, race, and season of birth, and incidence in siblings. *Br J Prev Soc Med* 7:211–219.

Malone FD, Canick JA, Ball RH et al. (2005). First-trimester or second-trimester screening, or both, for Down's syndrome. *N Engl J Med* 353:2001–2011.

Mills JL, Signore C (2004). Neural tube defect rates before and after food fortification with folic acid. *Birth Defects Res A Clin Mol Teratol* 70:844–845.

Nevin NC, Johnston WP, Merrett JD (1981). Influence of social class on the risk of recurrence of anencephalus and Spina bifida. *Dev Med Child Neurol* 23:155–159.

Oakley GP Jr. (2002). Global prevention of all folic acid-preventable Spina bifida and Anencephaly by 2010. *Community Genet.* 5:70–77.

Ohara PT (1972). Electron microscopical study of the brain in Down's syndrome. *Brain* 95:681–684.

Penrose LS (1938). *The Colchester Survey: A Clinical and Genetic Study of 1280 Cases of Mental Defect*. London: H.M. Stationery Office, Privy Council of Medical Research Council.

Pharoah PO, Buttfield IH, Hetzel BS. Neurological damage to the fetus resulting from severe iodine deficiency during pregnancy. *Lancet 1971* February 13;1(7694):308–310.

Ray JG, Wyatt PR, Thompson MD, Vermeulen MJ, Meier C, Wong PY, Farrell SA, Cole DE (2007). Vitamin B12 and the risk of neural tube defects in a folic-acid-fortified population. *Epidemiology* 18:362–366.

Richards BW (1962). *Proceedings of the London Conference on the Scientific Study of Mental Deficiency.* Dagenham, UK: May & Baker.

Roubertoux PL, Kerdelhue B (2006). Trisomy 21: from chromosomes to mental retardation. *Behav Genet* 36:346–354.

Seller MJ (1994). Vitamins, folic acid and the cause and prevention of neural tube defects. *Ciba Found Symp* 181:161–173.

Sherman SL, Freeman SB, Allen EG, Lamb NE (2005). Risk factors for nondisjunction of trisomy 21. *Cytogenet Genome Res* 111:273–280.

Smith DS (2001). Health care management of adults with Down syndrome. *Am Fam Physician* 64:1031–1038.

Stein Z, Susser M (1970). Mutability of intelligence and epidemiology of mild mental retardation. *Rev Education Res* 40:29–67.

Stein ZA, Susser M, Saenger G, Marolla F (1975). *Famine and Human Development: the Dutch Hunger Winter of 1944–1945.* New York: Oxford University Press.

Stromme P, Magnus P (2000). Correlations between socioeconomic status, IQ and aetiology in mental retardation: a population-based study of Norwegian children. *Soc Psychiatry Psychiatr Epidemiol* 35:12–18.

Susser E, Hoek HW, Brown A (1998). Neurodevelopmental disorders after prenatal famine: the story of the Dutch Famine study. *Am J Epidemiol* 147:213–216.

Tjio JH, Levan A (1956). The chromosome number of man. *Hereditas* 42:1–6.

Velie EM, Shaw GM, Malcoe LH, Schaffer DM, Samuels SJ, Todoroff K, Block G (2006). Understanding the increased risk of neural tube defect-affected pregnancies among Mexico-born women in California: immigration and anthropometric factors. *Paediatr Perinat Epidemiol* 20:219–230.

Waardenburg PJ (1932). *Das Menschliche Auge Und Seine Erbanlagen*. Martinus, Nijhoff, Den Haag.

Watson JD, Crick FH (2003). A structure for deoxyribose nucleic acid. 1953. *Nature* 421:397–398.

Wilcken B (2007). Recent advances in newborn screening. *J Inherit Metab Dis* 30:129–133.

Wolstenholme GEW, Porter R, eds. (1967). *Mongolism. Ciba Foundation Study Group.* Boston: Little Brown and Co.

Wood W, Lewis EO, et al. (1929). *Report of the Mental Deficiency Committee, Being a Joint Committee of the Board of Education and Board of Control*. London: H.M. Stationery Office.

Yaqoob M, Bashir A, Zaman S, Ferngren H, Von DU, Gustavson KH (2004). Mild intellectual disability in children in Lahore, Pakistan: aetiology and risk factors. *J Intellect Disabil Res* 48:663–671.

22

Infectious Disease Epidemiology: Beyond Bacteria

By the end of the nineteenth century, the significant role of microorganisms in many diseases of humans and other species was incontrovertible. The ground had been prepared for the enormous strides in the understanding of infectious diseases, especially with respect to agent and host, which followed in the twentieth century. The unraveling of the nature of microorganisms and their role in disease, the mechanisms of immunity, recognition of the existence of viruses and their dependence on living matter for survival, led to the scientific development of the virtually new discipline of molecular immunology, to the rising potential for the control of microorganisms, and ultimately to paradigmatic shifts in the practice of public health. Having faced, understood, and at last found means to quell such great scourges as smallpox, the plague, syphilis, cholera, and typhoid, the major threats posed by infectious disease seemed to have subsided. Late in the twentieth century, however, newly emerging diseases, and the human immunodeficiency virus (HIV) in particular, challenged medicine and the public health afresh.

Here a backward glance at Pasteur is informative. As described in Chapter 11, late in his career Pasteur broke new ground in his work on rabies. Unable to visualize the responsible organism under the microscope, but confident that something harmful and capable of infection existed in the material with

which he was working, he coined the term "virus" to describe the unseen yet harmful agent (Pasteur 1882/1996; Debré 1998). Faced with an outbreak of rabies, Pasteur proceeded with his usual methods of experimenting by "passage." With tissue taken from the brains of rabid animals, he proceeded with passage through a range of animal species until, led by trial and error, he found that passage through the monkey had effectively attenuated the effects of the unseen "virus" (Geison 1995). In this venturesome procedure, Pasteur was in a sense following Jenner. Pasteur, however, developed a quantitative rather than a simply qualitative approach. He defined "virulence" by duration of the incubation period: that is the time elapsed between exposure to the virus and the first manifestation of disease. The shorter the incubation period, the greater the virulence (Debré 1998). He found also that the rabies virus, once attenuated by passage through the monkey and injected into dogs, extended the incubation period, and seemed to immunize most dogs against the disease. He showed that attenuation could be achieved with exposure to air and sufficient time. But this immunization of dogs with attenuated virus gave no definite assurance of successful prophylaxis in humans. Pasteur reasoned that for so rare a disease as rabies in humans, the need was not for prophylaxis but for a treatment that would take effect in the interval between the time of initial infection (the bite of a rabid dog) and the disease manifestation. In other words, there was room for a procedure to take advantage of the long incubation period of about 40 days in humans to induce protection (Debré 1998). Such a vaccine would of course need to take effect before the disease became manifest. Pasteur began by injecting a much attenuated form of a highly virulent rabies culture into "clean" dogs. Each day he injected a less attenuated form until finally the fully virulent form proved harmless to the now immunized host. By the time he undertook the initial injections against rabies in the now famous case of a young boy (Joseph Meister) who had been bitten by a rabid dog three days earlier, the protection conferred by the vaccine was clearly evident.[1]

Pasteur seems to have conceptualized his system of gradual desensitization, now the basis of prophylaxis for many allergies, by intuition alone, and well before a science of immunology had begun to develop (Debré 1998). Perhaps the procedure might yet provide a model for postinfection vaccination against HIV.

A virus was first defined as an agent of disease both too minute to identify even under the microscope and small enough to pass through a very fine filter. In 1892, a Russian scientist, Dmitri Ivanovski, had shown that the tobacco

[1] In the case of Joseph Meister, the first inoculation was taken from the spinal cord of a rabbit that had died of rabies, and then suspended in air for 15 days. There followed 13 more injections, each using successively fresher (more virulent) material over the next 10 days (Pearce 2002). Also see Chapter 11.

mosaic disease was transmitted by an agent that could permeate porcelain filters, thus much more diminutive than any known bacillus. A few years later, in 1898, Martinus Beijerinck in Holland repeated Ivanovski's experiments and described the infectious agent as a *contagium vivum fluidum* (a contagious living fluid) (Bos 1995). In the same year, Loëffler and Frosch identified the virus that caused foot and mouth disease (Loëffler and Frosch 1898). A half century later, Macfarlane Burnet (1953a; 1953b) pointed out, however, that the initial definition of viruses solely by "filterability" (a quality standing in for size) would serve neither for smallpox nor rabies (both among the first human viruses studied) since neither was small enough to pass through filters. Rather than a definition based on size (although certainly a virus would be much smaller than a bacillus), Burnet offered a definition that relied on the distinctive biological property shared by all viruses of requiring access to a living cell for their survival.[2] The first direct view of the larger viruses was obtained only in 1925 with improved microscopes using ultraviolet light. In fact, only after the development of the electron microscope in 1938 could most viruses be visualized.

Viruses, like bacilli, attack preferred target cells in particular hosts, each producing effects that may also differ among given host species. Thus in humans, poliomyelitis virus attacks nervous tissue, influenza virus the respiratory system, and, as we now know, the Human Immunodeficiency Virus is named for its attacks on the cells of the immune system. As with other microorganisms, viral invasion induces antibody formation in hosts. Hence, much of the immunological understanding acquired in the study of bacteria applies also to viruses.

During the first half of the twentieth century, the pervasiveness of viruses and their specific effects began to gain general recognition. Yet their presence could not be established by applying Koch's postulates. In 1937, T.M. Rivers (1888–1962), a leading American virologist, pointed out that viruses as a class did not meet Koch's requirements for defining bacteria. Indeed, viruses do not even fulfill all the conditions for life. Thus they can be seen as a bridge between the living and the nonliving, or the animate and inanimate. Rivers noted that viruses consist largely of nucleic acid, and that the conventional light microscope could not visualize them. Moreover, viruses can be grown only on living tissue, and the need for such culture tissue for viruses at once overturns Koch's condition for defining bacteria that a culture medium be "pure" (that is, free of contamination by any other organism, hence lifeless). For example, the influenza virus, cause of the devastating

[2] The mouse skin was the initial choice for growing viruses in laboratory studies, but soon the chicken egg was found to have many advantages, although currently viruses are usually grown in tissue culture.

pandemic of 1918–1919, was first isolated only years later, in 1931. Growth of that virus in culture occurred only in synergy with the concurrent culture of live hemophilic bacteria (Shope 1931; Rivers 1937). At the same time, Rivers issued a proclamation of sorts in these words:

> *"…Virus…has lost its old indefinite meaning and has acquired a new significance similar in exactness to that borne by the words bacterium and spirochete. The terms virus of small pox, Virus variolae, virus myxomatosum (Sanarelli) and virus of poliomyelitis are now as definitive as are the terms bacillus of typhoid, Bacillus typhosus, meningococcus and staphylococcus."* (Rivers 1937)

As an alternative to Koch's formulation for identifying bacteria, Rivers proposed three criteria for a *pathogenic* viral cause: first, the specific viral agent must be present in the body of all affected persons; second, given a specific disease, antibodies to the virus must not be apparent at its onset, but appear only at some time later in its course or else upon recovery; third, in susceptible animals the virus is found capable of causing either disease or death (Rivers 1937).

Understandably, developments spurred by germ theory initially had a narrow focus. Koch's postulates had served well to refute miasma theory and to launch a new phase in bacteriology, medicine, and epidemiology. Before weaknesses could be perceived and demonstrated, the shackles of miasma had first to be shed, and the new formulation had to gain currency. Finally, an ever-increasing pace of development in the relevant sciences posed persistent tests for the postulates formed by Koch. Beyond the very particular critique advanced by Rivers, however, latter-day developments in epidemiology meet further inadequacies.

At an ultramicroscopic level, the newly found "prions" associated with spongioencephalopathies that won Stanley Prusiner the Nobel Prize in 1997 extend the technical critique of the postulates put forward by Rivers. Described as "proteinaceous infectious particles" by Prusiner, prions are distinguished from viruses by their ability to survive boiling, alcohol, and even ultraviolet irradiation; also they have no identifiable nucleic acid. Still, for some critics their unique identity remains under dispute (Oldstone 1998; Chesebro 2003; Priola et al. 2003).

Epidemiologic studies of kuru, the first spongioencephalopathy to be identified in humans, were initiated in the late 1950s. This mysterious neurological epidemic afflicted particular tribal groups in Papua, New Guinea. Manifested clinically after a long inoculation period by ataxia, tremor and dysarthria, it was previously an unknown entity. After several years of on-site investigations, two social anthropologists, Robert Glasse and Shirley Glasse (later Lindenbaum) who were studying the people of the region, observed

that the epidemic appeared to coincide in time with the development of a ritual in which the mourners consumed the remains of the dead (Mathews 1965; Glasse 1967; Mathews et al. 1968; Lindenbaum 1979; Collinge et al. 2006). Moreover, they noted an exception in that portions of human brain were ingested only by women and children, who were also predominantly the victims of this uniformly fatal disease. These close social observations led the researchers to infer that the source of the distinctive physical disorder wreaked upon the victims of the disease might be harbored by the brains of the dead, which was the specific share of women and children. Thus these anthropologists provided an essential clue to the source of what was a rare and newly recognized infectious disease of humans.

In 1959, the veterinarian William Haddow had spelled out the clinical and pathological similarities of kuru to scrapie, an infectious disease of sheep with a very long incubation period (Haddow 1959). Awareness of this zoonosis moved Gajdusek's team to consider the possibility of an infection rather than a genetic condition. With monumental patience in the laboratory, after many attempts Gajdusek and his colleagues eventually succeeded in transmitting the disease to primates; upon exposure, the disease became manifest only several years thereafter. Gajdusek's unique discovery led to the recognition of what was then called a slow virus, now known as an infectious protein or prion responsible not only for kuru but also for Creutzfeld-Jacob disease (CJD) and several other spongioencephalopathies. In 1976, Gajdusek's persistence was rewarded by the Nobel Prize. As he said in his Nobel speech, he considered the infectious agent in Kuru to be a microbial agent, and decided to retain the term "virus," though not without misgivings (Gajdusek 1976).

Current knowledge of emerging infections, discussed later in this chapter, suggests the possibility that the original sufferers from Kuru might have contracted the disease from a local zoonosis or perhaps from a spontaneous case of CJD. Fortunately, the human form has essentially died out, as has the practice of partaking of human brain in that region. The last identified cases still survived in 2004, more than 50 years after the probable end of cannibalism (Collinge et al. 2006).

This winding path to discovery emphasizes once more that current epidemiology, engaged with a broad canvas at several levels of organization, is not well-served for all its purposes by Koch's tightly drawn postulates. As in the examples above, deficiencies first became acute for virologists at the microscopic level. Once viruses could be isolated in culture, the known number of diverse organisms proliferated rapidly; some were specific pathogens, many were not. Increasingly, as the profusion of viruses (and bacteria) that ubiquitously colonize human beings was recognized, the difficulty of establishing causal relations has grown commensurately.

In 1957 Huebner, a leading virologist at the National Cancer Institute in the United States, offered a new critique of the applicability of the postulates in the light of the flurry of new virus discoveries (Huebner 1957). Virologists were beset by a paradox. The new knowledge about viruses bred confusion. Among the many, most viruses were mere passengers without effect. To elicit specific causal relations grew only more difficult. Huebner reached an essentially epidemiological conclusion: in order to link a disease securely to a specific virus in the microcosmic world of viruses, he argued, account must be taken of the environmental context of infection. Full understanding would involve the biological, physical, and social environment external to the virus, as well as other virus types, hosts, transmission, and immunity.

Koch's postulates had contributed greatly to the spectacular early advance of bacteriology. Tests of the legitimacy of multicausality, however, are beyond their reach. An emergent multidimensional approach for the present day is even further out of the reach of Koch's formulation. We take up in later chapters new ways of thinking about causes that bridge infectious and noninfectious diseases (Susser 1991).

Immunology, the Host, and the Environment

In 1945, Macfarlane Burnet encapsulated his view of the key landmarks in immunology, which we summarize as follows (Burnet 1945):

A first landmark was surely the historic realization that survival from a first episode of smallpox confers protection against a recurrence. This well-known phenomenon was a source of speculation first to Rhazes in Persia in the tenth century and later to Fracastoro in Italy, in the sixteenth century. This observation led to the practice of variolation with matter taken from smallpox scabs or pustules, a practice apparently known for centuries before that in India, China, and perhaps earlier yet in the horn of Africa (Dinc and Ulman 2007). Variolation led by analogy to vaccination with Jenner's vaccine (now thought to be of uncertain provenance: see Cook 1996; Cook 1997).

A second landmark for the development of immunology was the modern discovery that with recovery from infection, evidence of the attack usually remains in the blood. Antibodies are produced in response to the specific invading microorganism and thereby form a defense against that organism to facilitate its destruction. The groundwork exploring the nature and identity of antibodies was accomplished mainly in Germany in the 1890s by Paul Ehrlich, E.A. von Behring, and Baron Kitasato Shibasaburō (known as Kitasato). To this day understanding of antibodies continues to deepen.

A third landmark was comprehension of the phenomenon of "tolerance." Both MacFarlane Burnet and Peter Medawar took as their starting points in immunology the disconcerting experience of surgeons whose attempted skin graft operations failed due to the immunological response of their patients. Although one's own skin could be grafted and would take elsewhere on one's body, the skin of another person would be rejected. Similar phenomena had been noted as early as the sixteenth century by Gaspare Tagliacozzi, a famous surgeon of Bologna. Five centuries passed before study of Tagliacozzi's observation was pursued.[3] Research in the mid-twentieth century into graft rejection turned out to have major implications for studies of immunology and genetics, as the Nobel Committee recognized in 1960 in the joint award of the prize to Burnet and Medawar for their work in these fields. These two scientists embarked on an endeavor to track the elements of the phenomenon of "tolerance" observed by Ray Owen in 1945 (Owen 1945). Owen reported that although among bovine siblings identical blood types are rare, the majority of dizygotic bovine twins did have identical blood type, which he attributed to in utero vascular anastomoses. This phenomenon implied immunological consequences that contravened the self/nonself principle set out by Burnet, and that started both Burnet and Medawar on the track of explaining such acquired tolerance.

The fourth landmark in Burnet's chronology of immunology is the clarification of blood groups. In 1901 Landsteiner described the ABO groups (Landsteiner 1901); in 1927 with Levine, the MNP groups (Landsteiner and Levine 1927), and in 1940 with Wiener), the Rh groups (Landsteiner and Wiener 1940).

In 1980 Dausset, Snell, and Benacerraf were together awarded the Nobel Prize for their work on the genetic basis of the Major Histocompatabilty Complex. As Burnet had noted in 1945, the themes and languages of genetics and immunology had then to find common ground. As he put it, immunology is concerned with how and why the body reacts actively against almost everything that is "foreign," where foreign means genetically different from its own substance (Burnet 1945) and thus, in Burnet's words, distinguishing "self" from "not self" (Burnet 1964).

The Agent, the Host, the Environment

With these and later developments in virology and immunology, and their growing engagement with genetics, epidemiologists are now much better equipped to deal with the triad of agent, host, and environment that evolved

[3] Tagliacozzi wrote, "The singular character of the individual entirely dissuades us from attempting this work (tissue transplantation) on another person." (See Duquesnoy 2005.)

in the late nineteenth and early twentieth century. Each of the three elements is now far better understood, the nature of the interactions between them begins to be appreciated, and the complexities of the multilevel causal patterns of disease begin to be unraveled (Susser 1973).

In what follows in this chapter, we first describe interactions between infective agents and host populations in general, without invoking biological processes for these. Next with more specific examples, we illustrate the interactions of infectious agent and host interactions, where certain genetic polymorphisms in the host population enhance defenses against the agent, and thereby favor individuals so endowed over others in the battle for survival. We also illustrate how the effect of one agent, as for example HIV, alters the host response to another agent, namely, tuberculosis. This example illustrates dynamic agent-host-agent interactions. Then, with emerging infections, we consider changes in such other aspects of the environment as vegetation, climate, habitat, culture, and their impact on the host, and how these may influence susceptibility and response to an infection. Some of these factors clearly play a part in the distribution and transmission of these unfamiliar infections. In Chapter 23, we describe how interaction with the host population may even change the genetic nature of the viral agent itself.

Infective Agents, Host populations, and the Environment

As we now appreciate in both host and agent, either individually or en masse, metamorphosis of varying degree is in continual interaction with parallel environmental change. Part of that interactive environmental change is induced by the hosts themselves. Both unknowingly and knowingly, human beings have themselves induced striking changes as the huge rise in populations has pervaded the world in parallel with the growth and power of the technical capacities of modern societies. These interactions create a growing and shifting web of multidimensional interrelationships. To untangle the web calls for the combined efforts of epidemiologists, microbiologists, immunologists, and geneticists, anthropologists, earth scientists, mathematicians, and perhaps other disciplines as well.

As noted above, we can infer that some of the changes in the host evolved in response to exposure to infectious diseases. Thus, the historian William McNeill surveyed such evidence as exists of the effects of epidemic disease—measles as well as smallpox and bubonic plague—on the evolution of populations and civilizations over the millennia. He ascribes to epidemic disease, especially measles and smallpox, the easy conquest of the Aztec Empire of Mexico by a very small contingent of Spanish invaders (McNeill 1989). The Aztecs had never before seen either disease. When Cortés reached

Mexican shores, it seems the might of the imperial Aztec city was already collapsing. The people and their rulers, entirely naïve to the infections carried from Europe by the Conquistadors, and thus completely susceptible, were rendered prostrate by what was for the Aztecs a seeming visitation from the gods (Diamond 2003). On the Spanish side, it is a reasonable historical speculation that most were survivors of such epidemics in Europe and immune.

Within the span of a decade, estimates suggest that the Aztec population of central Mexico declined from 25 million to 16.8 million (Eyler 2003). Within half a century, the susceptible native population may have been reduced by 90% in some regions, though there is an active scholarly debate on the population size, the scale of collapse, and related causes[4] (McCaa 1995; Livi-Bacci 2006). In Peru, and in North America too, colonizing and conquering Europeans provoked similar episodes. Indeed, when naïve populations meet newcomers bringing unfamiliar diseases, decimation and worse is always a likelihood. Examples of populations vulnerable to the smallpox introduced by colonization are numerous: Native Americans, KhoiSan and San in South Africa, the inhabitants of the Amazon jungles, and many others. In 19th century Europe, Panum documented the tragic invasion of the measles virus in the Faro islands. He also provided one of the key early examples of the immunity conferred by infection on those few still alive who had survived a previous epidemic (Panum 1847; see Chapter 11).

From the viewpoint of the epidemiologist both individual host and host populations are central to concepts of immunity. Immunological changes in the host, whatever their origin, will influence the distribution and severity of infectious diseases. Tuberculosis exemplifies a disease in which a genetic element has long been established, but in which environment plays a dominant role. To illustrate, in the United States and other developed countries, by the end of the twentieth century the rates of occurrence of this disease had dwindled to a small fraction of the rates still prevalent in the first half of the twentieth century. But in large swaths of the less-developed world, great change has also taken place in the opposite direction (Dye 2006). Mass HIV infection has engendered widespread compromise of immune competence, since its main targets are the CD4 cells of the immune system. With immune defenses down in populations severely affected, the exposure produced a rapid return of tuberculosis and other secondary infections (Nunn et al. 2005). To this we may add the appearance, first in South Africa and then elsewhere, of the Xdr strain of tuberculosis resistant to all currently available treatment (Gandhi et al. 2006).

[4] Social factors such as the expulsion and forced migration of indigenous peoples would have increased the epidemic's impact.

On the obverse side of the defense against the transmission of disease, since the 1950s equally complex mechanisms and conditions related to genetically determined immune enhancement have come to be understood. Thus, A.C. Allison's study of sickle cell anemia in Kenya gave an early signal of the potential of such immunity and its interaction with environment at the population level (Allison 1954a; Allison 1954b). The sickle cell affliction originates among descendants of Africans who inherit a single recessive gene from each parent. The condition is concentrated in the malarial regions of East, Central, and West Africa. The explanation for this distribution of the gene followed from the discovery of the relative resistance of sickle cell carriers to falciparum malaria. Hence, malarial regions favor the selective survival of those who are carriers of the gene; over long generations the proportion of sickle cell carriers is bound steadily to rise as long as the fatal malarial disease persists (Carter and Mendis 2002), perhaps to be balanced by the relative disadvantage of the double inheritance of the gene for sickle cell anemia.

In this respect, a subsequent study of West African children in Burkina Faso is also worth noting. The authors report that the variant hemoglobin C gives even more protection against malaria than does the sickle cell variant. So far it seems to do so without any parallel of the costs exacted by hemoglobin S in sickle cell disease (Modiano et al. 2001; Verra et al. 2007). Since the discoveries around sickle cell disease were first made, understanding of genetic mechanisms of immunity has been greatly elaborated. Human leukocyte antigens (HLAs) (discussed in chapter 23) have been found to relate to immune mechanisms and hence to the distribution of certain infectious diseases. Because of the multiple variants among the increasing numbers of HLAs detected, these patterns seemed at first to grow inordinately complex. Gradually, however, the application of molecular genetics has established the role of HLAs in an array of disorders (Fellay et al. 2007; Borghans et al. 2007; Tibayrenc 2007).

Changes in the Environment

William Jenner foresaw a possible world that would be free of smallpox. In the world he himself knew, his prediction has indeed been fulfilled. In glaring contrast, two centuries later MacFarlane Burnet foresaw the possibility, very much alive in present day thinking, that an enemy force might make a deliberate effort to spread an epidemic disease. Intuitively, Sir Thomas Browne (1605–82) described the influence of geography and culture on the potentiality for new disease: "New discoveries of the Earth discover new diseases: for besides the common swarm, there are endemial and local infirmities proper

unto certain Regions, which in the whole Earth make no small number; and if Asia, Africa and America should bring in their list, Pandora's Box would swell, and there must be a strange Pathology" (Browne 1971).

Changes in ecology clearly induce change in the patterns of disease. The phenomenon has been demonstrated many times over with malaria, bilharzia (schistosomiasis), river-blindedness (onchocerciasis), and the like. On the East coast of the United States, in a recent example, deer impinge on residential areas and carry minute ticks infected by the spirochete *Borrelia burgdorferi,* the agent of Lyme disease. Focused modifications of the physical niche that maintains the life cycle of such infective agents provide a means of controlling transmission (for instance, limiting vegetation in traversed areas, or using targeted insecticides to control ticks). Bacilli (and other infectious agents) can also lose or gain virulence, often for no known or well understood environmental or genetic reason. We may fairly say that the appearance in a population of previously unknown infectious diseases—now categorized as "emerging infections," fulfill the vision of Sir Thomas Browne. The explanation for their appearance is often far from obvious.

Emerging infectious diseases include those either newly appearing in a population, or rapidly increasing in incidence, or extending their geographic range. Even for the epidemiologist reared on the interactive triadic model of agent, host, and environment, the increasing number and scale of emerging infections is beyond expectation. Whether among animal or human species, some of these infections are novel. Many are caused by microbial vectors carried by animals or birds: Hanta virus pulmonary syndrome, Severe Acute Respiratory Syndrome (SARS), Lyme disease, Rift Valley fever, Hemolytic Uremic Syndrome (HUS), the recent (1999) appearance in the US of mosquito-born West Nile virus, Bovine Spongio-Encephalopathy (BSE), a disease transmitted first from one animal to another and eventually to beef-consuming humans, and most notably, human immunodeficiency virus (HIV).[5] The movements of infections across species has been dubbed "microbial traffic" (Morse 2004). Table 22.1 (reproduced from Morse 2004) provides a succinct overview of the environmental factors, physical, political, and cultural, that form the backdrop to these movements.

For the epidemiologist in our current era, emerging diseases present a global threat no less urgent than does global warming and climate change for the environmentalist. The evidence that both are increasing is inescapable. The connections between them become daily more manifest. As Robert May wrote, "All viral host associations have thresholds that govern the possibility of the virus invading or persisting, and thus, in whatever sense we mean this,

[5] Some scientists posit that cross-species transmission of SIV does not by itself constitute the basis for a zoonosis. See Marx et al. (2004).

Table 22.1 Factors in Infectious Disease Emergence: Some Examples (Categories Should Not Be Considered Mutually Exclusive as Several Factors May Contribute to the Emergence of a Particular Disease)

Factor	Examples of specific factors	Examples of disease emergence
Ecological changes (including those due to economic development and land use)	Agriculture, dams, changes in water ecosystems, deforestation/reforestation, flood/drought, famine, climate change	Rift Valley fever (dams, irrigation), Argentine hemorrhagic fever (agriculture), Hantaan or Korean hemorrhagic fever (agriculture), hantavirus pulmonary syndrome in the southwestern United States of America, 1993 (weather anomalies)
Human demographics, behavior	Societal events: population migration (movement from rural areas to cities), war or civil conflict, economic impoverishment, urban decay, factors in human behavior such as the commercial sex trade, intravenous drug use, outdoor recreation, use of childcare facilities and other high-density settings	Spread of human immunodeficiency virus and other sexually transmitted diseases, spread of dengue (urbanization)
International travel and commerce	Worldwide movement of goods and people, air travel	Dissemination of human immunodeficiency virus, dissemination of mosquito vectors such as *Aedes albopictus* (Asian tiger mosquito), rat-borne hantaviruses, introduction of cholera into South America, dissemination of 0139 (non-01) cholera bacteria (via ships)
Technology and industry	Food production and processing: globalization of food supplies changes in food processing and packaging	Food production processes: hemolytic uremic syndrome (certain *Escherichia coli* strains from cattle contaminating meat and other food products), bovine spongiform encephalopathy, Nipah virus (pigs), avian influenza, severe acute respiratory syndrome (probably)
	Health care: new medical devices, organ or tissue transplantation, drugs causing immunosuppression, widespread use of antibiotics	Ebola and human immunodeficiency virus (healthcare and medical technology, contaminated injection equipment), opportunistic infections in immunosuppressed patients, Creutzfeldt-Jakob disease from contaminated batches of human growth hormone

(*continued*)

Table 22.1 Continued

Factor	Examples of specific factors	Examples of disease emergence
Microbial adaptation and change	Microbial evolution, response to selection in the environment	"Antigenic drift" in influenza virus, possibly genetic changes in severe acute respiratory syndrome coronavirus in humans, development of antimicrobial resistance (human immunodeficiency virus, antibiotic resistance in numerous bacterial species, multi-drug resistance in numerous bacterial species, multi-drug-resistant tuberculosis, chloroquine-resistant malaria)
Breakdown in public health or control measures	Curtailment or reduction in disease prevention programs; lack of, or inadequate sanitation and vector control measures	Resurgence of tuberculosis in the United States of America, cholera in refugee camps in Africa, resurgence of diphtheria in former Soviet republics and Eastern Europe in the 1990s.

Reproduced from Revue scientifique et technique 23(2):443-51, Stephen S. Morse, Copyright (2004) with permission from the International Office of Epizootics, now the World Organisation for Animal Health.

emerging in the population. These thresholds depend on host density and behavior, and they will change from time to time and from place to place as the ecological setting, host behavior, and host density change" (May 1993).

References

Allison AC (1954a). Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J* 4857:290–294.

Allison AC (1954b). The distribution of the sickle-cell trait in east africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Trans R Soc Trop Med Hyg* 48:312–318.

Borghans JA, Molgaard A, de Boer RJ, Kesmir C (2007). HLA Alleles Associated With Slow Progression to AIDS Truly Prefer to Present HIV-1 P24. *PLoS ONE* 2:e920.

Bos L (1995). The embryonic beginning of virology: unbiased thinking and dogmatic stagnation. *Arch Virol* 140:613–619.

Browne T (1971). *A Letter to a Friend Upon Occasion of the Death of His Intimate Friend.* Ed. by G. Keynes. Boston: D.R. Godine.

Burnet FM (1964). 1960 Nobel lecture: immunological recognition of self. *Nobel Lectures, Physiology or Medicine 1942–1962*. Amsterdam: Elsevier Publishing Co, pp. 689–701.

Burnet FM (1945). *Virus As Organism: Evolutionary and Ecological Aspects of Some Human Virus Diseases*. Cambridge, MA: Harvard University Press.

Burnet FM (1953a). *Natural History of Infectious Disease.* 2nd ed. Cambridge, England: University Press.

Burnet FM (1953b). *Viruses and Man.* London, Baltimore: Penguin Books.

Carter R and Mendis KN (2002). Evolutionary and historical aspects of the burden of malaria. *Clin Microbiol Rev* 15:564–594.

Chesebro B (2003). Introduction to the transmissible spongiform encephalopathies or prion diseases. *Br Med Bull* 66:1–20.

Collinge J, Whitfield J, McKintosh E, Beck J, Mead S, Thomas DJ, Alpers MP (2006). Kuru in the 21st century—an acquired human prion disease with very long incubation periods. *Lancet* 367:2068–2074.

Cook GC (1996). The smallpox saga and the origin(s) of vaccination. *J R Soc Health* 116:253–255.

Cook GC (1997). Author's reply. *J Med Biogr* 5:241.

Debré P (1998). *Louis Pasteur.* Baltimore: Johns Hopkins University Press.

Diamond JM (2003). *Guns, Germs, and Steel.* New York: Spark Pub.

Dinc G and Ulman YI (2007). The introduction of Variolation 'A La Turca' to the West by Lady Mary Montagu and Turkey's contribution to this. *Vaccine* 25:4261–4265.

Duquesnoy RJ (2005). Early history of transplantation immunology: Part 1. *ASHI Quarterly (Newsletter)*74–79.

Dye C (2006). Global epidemiology of tuberculosis. *Lancet* 367:938–940.

Eyler JM (2003). Smallpox in history: the birth, death, and impact of a dread disease. *J Lab Clin Med* 142:216–220.

Fellay J, Shianna KV, Ge D Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C et al (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* 317:944–947.

Gajdusek, D (1976). Unconventional viruses and the origin and disappearance of Kuru. In Lindsten J, ed. *Nobel Lectures, Physiology or Medicine 1971–1980.* Singapore: World Scientific Publishing Co, pp. 305–354.

Gandhi NR, Moll A, Sturm AW, Pawinski R, Govender T, Lalloo U, Zeller K, Andrews J, Friedland G (2006). Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet* 368:1575–1580.

Geison GL (1995). *The Private Science of Louis Pasteur.* Princeton, NJ: Princeton University Press.

Glasse R (1967). Cannibalism in the Kuru region of New Guinea. *Trans N Y Acad Sci* 29:748–754.

Hadlow WJ (1959). Scrapie and Kuru. *Lancet* 289–290.

Huebner R (1957). Criteria for etiologic association of prevalent viruses with prevalent diseases; the virologist's dilemma. *Ann N Y Acad Sci* 67:430–445.

Landsteiner, K (1901). On agglutination phenomena of normal human blood (Transl). Republished in S.H. Boyer, ed. *Papers on Human Genetics*. Englewood Cliffs: NJ: Prentice-Hall, 1963, pp. 27–31.

Landsteiner K, Levine P (1927). A new agglutinable factor differentiating individual human bloods. *Proc Soc Exp Biol Med* 24:600–602.

Landsteiner K, Wiener AS (1940). An agglutinable factor in human blood recognized by immune sera for rhesus blood. *Proc Soc Exp Biol Med NY* 48:223–224.

Lindenbaum S (1979). *Kuru Sorcery, Disease and Danger in the New Guinea Highlands.* 1st ed. Palo Alto, CA: Mayfield Pub. Co.

Livi-Bacci M (2006). The depopulation of Hispanic America after the conquest. *Popul Dev Rev* 32:199–+.

Loeffler F, Frosch P (1898). Report of the commission for research on foot-and-mouth disease. *Zentrabl Bacteriol Parastenkunde Infektionkrankh* 23:371–391.

Marx PA, Apetrei C, Drucker E (2004). AIDS as a zoonosis? Confusion over the origin of the virus and the origin of the epidemics. *J Med Primatol* 33:220–226.

Mathews JD (1965). The changing face of Kuru. an analysis of pedigrees collected by R.M. Glasse and Shirley Glasse and of Recent Census Data. *Lancet* 1:1139–1142.

Mathews JD, Glasse R, Lindenbaum S (1968). Kuru and Cannibalism. *Lancet* 2:449–452.

May, RM (1993). Ecology and evolution of host-virus associations. In Morse, eds. *Emerging Viruses*. New York: Oxford University Press, pp. 58–68.

McCaa R (1995). Spanish and Nahuatl views on smallpox and demographic catastrophe in Mexico. *J Interdiscip Hist* 25:397–431.

McNeill WH (1989). *Plagues and Peoples.* New York: Anchor Books.

Modiano D, Luoni G, Sirima BS, Simporé J, Verra F, Konaté A, Rastrelli E, et al (2001). Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature* 414:305–308.

Morse SS (2004). Factors and determinants of disease emergence. *Rev Sci Tech* 23:443–451.

Nunn P, Williams B, Floyd K, Dye C, Elzinga G, Raviglione M (2005). Tuberculosis control in the era of HIV. *Nat Rev Immunol* 5:819–826.

Oldstone MBA (1998). *Viruses, Plagues, and History.* New York: Oxford University Press.

Owen RD (1945). Immunogenetic consequences of vascular anastomoses between bovine twins. *Science* 102:400–401.

Panum P (1847). Observations made during the epidemic of measles on the Faroe Islands in the year 1846. *Bibliothek for Laeger* 1:270–344. See Panum P (1989) *Med Classics* 3:829–886 for recent reprint.

Pasteur L (1882/1996). Pasteur and rabies: an interview of 1882. Introduced and translated by J. Illo. *Med Hist* 40:373–377.

Pearce JM (2002). Louis Pasteur and rabies: a brief note. *J Neurol Neurosurg Psychiatry* 73:82.

Priola SA, Chesebro B, Caughey B (2003). Biomedicine. A view from the top—prion diseases from 10,000 feet. *Science* 300:917–919.

Rivers TM (1937). Viruses and Koch's postulates. *J Bacteriol.* 33:1–12.

Shope RE (1931). Swine influenza III. Filtration experiments and etiology. *J Exp Med* 54:373–385.

Susser M (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* New York: Oxford University Press.

Susser M (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am. J. Epidemiol.* 133:635–648.

Tibayrenc M (2007). Human genetic diversity and the epidemiology of parasitic and other transmissible diseases. *Adv Parasitol* 64:377–422.

Verra F, Simpore J, Warimwe GM Tetteh KK, Howard T, Osier FH, Bancone G, et al (2007). Haemoglobin C and S role in acquired immunity against Plasmodium falciparum malaria. *PLoS ONE* 2:e978.

23

Human Immunodeficiency Virus and the Role of Women: The New Challenge

Origins

In 1980, a new disease that matched no known existing complex of signs, symptoms and course appeared in the American cities of Los Angeles, New York, and San Francisco. Seemingly, the outbreak was confined to communities of homosexual men. Hence, the name first given to this entirely new and complex human affliction, for want of a better, alluded to its most distinctive initial epidemiological characteristics. Patients presented with manifestations of a distressing series of opportunistic infections previously little known, together with manifestly depressed counts of the CD4 cells critical to the protection conferred by the immune system: hence, Gay-Related Immunodeficiency Disease (Shilts 1987; Bayer and Oppenheimer 2000), later described more generally as Acquired Immune Deficiency Syndrome, and colloquially abbreviated as (AIDS).

This new epidemic, later found to be owed to the human immunodeficiency virus (HIV), proved to be a singular, dramatic, and worldwide affliction of our time. Huge in its dimensions and effects over the past three decades, in much of the world HIV/AIDS remains unconquered in 2009, and a continuing potential threat everywhere. We shall treat this devastating new epidemic at some length.

The HIV-related organism, innocuous in the chimpanzee in Central Africa, is thought to have first appeared among humans in the late 1920s (Korber et al. 2000, Korber et al. 2001). Its far from innocuous leap from chimpanzees to human populations was entirely unanticipated. By the mid-twentieth century in developed societies, it was widely assumed that the conquest of infectious diseases was the great feat of the biological sciences and public health. With the diffusion of the technical and scientific culture of Europe and North America across the less-developed world, the ensuing effectiveness of biochemical and antibiotic remedies, the reduction of infant mortality, and a steady rise in age at death, these developments were sufficiently notable for the World Health Organization to predict with confidence that in succeeding decades, health would improve in concert with wealth, living standards, nutrition, and education (Gordon 1976; United Nations and Dept. of Public Information 1979; Stern and Markel 2004).

The HIV/AIDS pandemic undermined that established assumption. During the latter half of the twentieth century, the work of epidemiologists, and hence their focus and their training, had been increasingly absorbed by the study of noncommunicable diseases and the detection of predominant risk factors that might be causal. Concurrently and unsurprisingly, interest and activity on the infectious disease front had declined.[1] But recognition of this new infectious disease, one could not but feel, made mock of the belief then prevailing and so clearly reflected in *The Conquest of Epidemic Disease*, Winslow's grand title for his excellent book published in 1943 and in print for decades thereafter (Winslow 1943).

The closest precedent of the HIV onslaught was the syphilis epidemic in Europe that had broken out more than a half-millennium in the past. Its source—admittedly disputed by some writers—was in all probability the spirochete brought back to Spain from Central America in 1492 by crewmen of Christopher Columbus (Rothschild 2005).[2] The sexually transmitted spirochete ignited an epidemic that within very few years had spread across the whole European continent. In much the same manner, the newly discovered agent of AIDS (HIV: Human Immunodeficiency Virus), was an agent also transmissible sexually from person to person (although less readily than

[1] For example: "There may be some wholly unexpected emergence of a new and dangerous infectious disease, but nothing of the sort has marked the last fifty years." (Burnet and White 1972). "As far as the developing areas of the world are concerned, social advancement, and the activities of the World Health Organization and of governments have greatly decreased the danger of major epidemics.... It would seem, however, that the violent swings in mortality caused either by pandemics or really severe local epidemics are becoming things of the past even in relatively primitive parts of the world." (Gordon 1976). Indeed, in 1979 Richard Doll uses the term "the post-infection era" (Doll 1979).

[2] Fully discussed for many years, the true source of the spirochete and syphilis is still disputed by some authors. Also see (Baker and Armelagos 1988; Meyer et al. 2002).

the spirochete). Its repertoire of defenses challenged therapeutic initiatives and evaded social and behavioral controls at both individual and population levels. Aided by the mobility of modern transport, the new virus spread as quickly and wreaked havoc much as had the spirochete at the turn of the fifteenth century. With a penchant for lodging in and destroying vital cells on which immunity depends, the virus undermined a prime defense of the body against the ravages of infection. The recognition of HIV-related disorder as a newly emerging chronic disease of infectious origin, the devastation it wrought in many countries and its unremitting threat sharpened a worldwide shift in renewed concern with the potential of infectious disease.

The human immune deficiency virus is believed to have emerged in a region of the Northern Congo in Central Africa (Hahn et al. 2000). It was there in 1959 that the first infected human blood specimen was identified.[3] The complex new human retrovirus was without doubt the product of a jump across primate species. Traced back to the chimpanzee *Pan troglodyte troglodyte*, the virus co-existed in its blood as a harmless "resident." Among humans, probably some subtle change in the virus upon the transfer to its new host favored human-to-human transmission. Thereafter, environmental change may then have favored further transmission and dissemination of the virus. Once the virus was recognized and the societal and individual damage entailed by its dissemination came to be understood, virologists, epidemiologists, and others from relevant disciplines immediately set about the work of protection and defense.

Tracing the Advent of the New Virus

Two separate approaches to genomic investigation have yielded biological comprehension and rendered a history of the HIV pandemic. *Phylogenetics* sought to unravel the genome of the virus and track its evolution. *Immunology* sought to unravel markers in the human genome of past exposures to retroviruses.

Phylogenetic studies of this rapidly changing invasive human virus begin with the recognition of a very similar virus among chimpanzees, the source

[3] In 1959, Arno Motulsky from Stanford University and Jean Vandepitte from Belgium, both geneticists interested in blood variants from different populations, journeyed across Africa to Leopoldville and its surrounds in the Congo and secured specimens from 99 apparently healthy adults. In 1985 Andre Nahmias, epidemiologist and geneticist from Emory University, Atlanta, was able to examine these specimens. Nahmias and several others confirmed that one specimen, taken we only know from a male, was HIV positive. This specimen is important in that it ranks as the earliest confirmed human infection with the virus. Further study of this specimen was able to place the molecular position of the virus on the "tree" (Motulsky et al. 1966; Nahmias et al. 1986; Hooper 1999).

from which HIV almost certainly derived. From these studies, an estimate could be made that the viral crossing from chimpanzees to humans—thereby producing a cross-species with perhaps some degree of change in its make-up—occurred during the 20-year interval between World Wars I and II. The site of this crossing seems almost certainly to have been the Congo forest habitat of the suspect chimpanzees (Gao et al. 1999; Hahn et al. 2000; Korber et al. 2000). Some two or three decades thereafter, the virus was identified from the stored blood of the first person known to have been infected (see footnote 3). Doubt remains, however, as to exactly when HIV crossed from chimpanzees to humans and how its dissemination began. One speculation is that native forest dwellers, who shared the forest loci with the chimpanzee carriers, improved their sparse diets with the meat of *Pan troglodyte troglodyte*. Or perhaps mere handling of the "bush-meat" was sufficient to infect the handler. Another suggestion is that used syringe needles, contaminated by lax sterilization or none at all, subsequently facilitated either the transformation of the virus, or its spread, or both (Marx et al 2004).[4] Thus the precise manner of the HIV crossing from chimpanzees to humans remains uncertain.

Among human beings, the effects of HIV are slow to become manifest. In the United States, although the infection must have arisen earlier, it was 1980 before the clinical condition of AIDS was first recognized and as we describe below, it was not until 1983 that the virus responsible for its spread was identified. Soon after, advances in technique enabled recognition of phylogenetic changes in viruses. The main subtypes were identified and two viruses, HIV 1 and HIV 2, have been distinguished as causing different diseases. HIV 1 has been further subdivided into subtypes or clades. Subsequent observation made it evident that in an infected individual, not only might more than one subtype be at work, but that these variants could again combine to form new variants. Such combinations would add to those presumed either to result from random mutations or, perhaps, from the effects of environmental factors as yet unknown. Figure 23.1 displays the assumed course of the phylogenetic history of HIV 1 (Korber et al. 2000). HIV 2, a virus found among several monkey species has been identified in the main in West Africa. Compared with HIV 1, it is much less pathogenic,

[4] For several reasons, it is useful to distinguish HIV, the virus, from AIDS, which is the disease it causes. Thus Marx et al. argue with some urgency that while we know that the virus was transmitted to humans from other primates infected with simian immune-deficiency virus (SIV) there is no evidence of disease among infected chimpanzees. Why typical manifestations of the virus should appear in humans but not in some other primates is not well understood. Marx et al. emphasize that what is novel and unique is the human response to HIV/AIDS, and not the transmission of HIV (Marx et al. 2004). In a somewhat similar vein, vaccinologists are now concentrating more on attempts to modify the response to the infection, and less so on preventing the initial encounter with HIV.

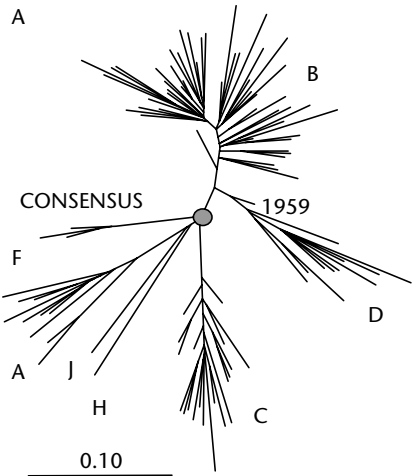


Figure 23.1 The maximum-likelihood tree, by Korber et al. (2000). Korber et al. (2000) reconstructed the evolutionary history of HIV-1 as a phylogenetic tree. Subtypes are indicated by letters. Branch and leaf lengths are plotted against time. Reprinted with permission. Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000) *Science* 288, 1789–1796.

slower to cause disease, and less likely to be transmitted to others (Mayer and Meese 2005).

Immunological data, from humans as well as chimpanzees and from the genomes of both individuals and groups among them, provide some evidence of past infections. It is a reasonable supposition that some of the genomic differences between these primate species appeared only when the human species began to diverge from its closest primate relative. For example, the chimpanzee genome contains DNA residues thought to be the result of past infections. Similarly, in human beings also genomic traces of past endogenous retroviruses can be found. In general, the greater the geographic distance separating human beings, and the further in historical time they are removed from each other, the more distinctive are such traces of endogenous retro-viruses as remain within the genomic structure. Human endogenous retroviruses (HERVS) in the human genome are currently under active investigation (Griffiths 2001).

Such studies suggest that past exposures to pathogens have shaped subsequent human immunological responses. Frank Macfarlane Burnet and also Peter Medawar, joint Nobel prizewinners in 1960, inferred that in human evolution the nature of such biological defenses against microbial invaders were crucial. The adumbration of the human genome continues to add precision and hence better understanding to this evolutionary perspective. Thus, in a recent paper, researchers could describe the quantitative makeup

of the human genome in the following terms: "It is estimated that approximately 8% of the human genetic sequence is derived directly from retroviral elements." Genetic variation in the human host, it is clear, plays a key role in the impact, progress, and treatment of HIV disease among individuals and populations (Lander et al. 2001; Mayer and Meese 2005).

The Spread of the Virus

The HIV/AIDS syndrome, first described among homosexual men in Los Angeles, New York, and San Francisco, next manifested in hemophiliac men. The unmistakable inference was that the new disease must be blood-borne, a critical observation for epidemiologists. Evidently, the blood derivatives on which hemophiliac men depended for survival must be responsible. So, too, other patients with conditions that called for blood transfusions contracted the disease. Soon thereafter, an epidemic was recognized also among intravenous drug users.

Given the presence of infected bisexual men, it was of course inevitable that in time they would transmit the virus to the women who were their partners (Flam and Stein 1986). Nevertheless, the realization that women could both contract the disease and transmit it heterosexually required a further shift in thinking. In the US, the presence of such heterosexual transmission became apparent from two observations: first, infants were manifesting immunodeficiency, having contracted the disease from their infected mothers. Second, reports began to flow from Africa that AIDS (i.e., the full blown autoimmune deficiency syndrome) had taken hold not only among men, but among women and if less often, among children, too.

As the epidemic spread across the world, the World Health Organization (WHO) began periodically to map and enumerate the accumulating data on incidence and prevalence. At the time of writing (2009), prevalence is highest in sub-Saharan Africa. In some regions, as in parts of southern Africa, prevalence in the population of reproductive age reached as high as 40%. Infection spread also to parts of India, Thailand, Myanmar, Vietnam, and most recently, China. The range of the epidemic is bound to extend from all these sites, given the dominant mode of transmission of the virus by heterosexual spread. In countries of Eastern Europe such as Russia and Poland and also in Central Asia, however, the main initial focus of infection and consequent transmission seemed to arise among intravenous drug users (Aceijas et al. 2004; Rosinska 2006; Rhodes et al. 2006). The considerable differences across countries can be attributed in the main to variation in social practices. Given the many mutations of the virus itself, these too may have a role (Abdool Karim et al. 2007).

Transmission of the disease from mother to infant was first described by pediatricians in New York and in New Jersey (Rubinstein et al. 1983; Rubinstein 1983; Oleske et al. 1983). These pediatricians identified the condition at birth and, soon after, specifically among the infants of immigrant women from Haiti. Haiti was a popular vacation resort for US residents, not least for the male homosexual population. At once stigmatized as a source of infection, the numbers of vacationers in Haiti dropped precipitously, exacerbating the poverty of the islanders. In the event, the march of the epidemic in Haiti advanced in synchrony with that among target groups in the US, leaving in doubt just where the disease first took hold in the Americas. What seems likely is that in Haiti and other Caribbean resorts, bisexual married men were infected during homosexual encounters with vacationing "gay" men, and thereafter transmitted the virus not only to other men but also to their wives.

Molecular Battles: Invasion and Defense

Epidemiologists have learned many new lessons from the HIV pandemic. Indeed, the more than two decades of that experience can be seen as virtually defining an era in its own right. At the biological level, the human immunodeficiency virus introduces new complexity into the concept of interaction between agent and host (Nolan et al. 2004). As noted previously (see Chapter 22), viruses survive as parasites on the living cells of the host. Different viral types each tend to select the cells of a particular system of the body as targets or hosts, for instance, respiratory, neurological, or gastrointestinal. HIV, an RNA virus, is unusual in its targets and its effects: it chooses for attack and destruction the CD4 lymphocytes, the very cells that are a principal component of the human immune defense system. To gain entry into these cells the virus, in a process known as "fusion," first binds to receptors on the cell surface. The virus then penetrates the cell, and by means of its enzyme *reverse transcriptase* it is able to appropriate and adapt to its own use the nuclear DNA of the invaded cell. During this so called provirus stage, the viral enzyme protease is produced. Viral proteins cut by protease are assembled in the parasitized cell, and reconstitute the RNA form of the virus. In this new form the virus escapes from the cell into the bloodstream.

New understanding of these processes has guided the development of therapeutic drugs. Thus, some among these drugs target the reverse transcriptase enzyme. Especially in combination, these drugs provide a powerful weapon against HIV. Another group of drugs aim to disable the protease enzyme, in this way disrupting the reconstitution of the provirus within the

CD4 cell. Other drugs still to be tested aim to block entry of the virus into human cells.

Among such other primate species as for example macaques and African monkeys, HIV may or may not choose the same cellular targets and it is only in some primate species that the virus causes devastation equal to or greater than in humans. Thus, as noted above, in its African chimpanzee host the virus seems to live as an innocent parasite, and elicits no manifestations of disease. Yet in the Asiatic macaque, never previously exposed, the infection tends to be rapidly fatal. Among humans, the virus undermines immune defenses and thereby opens the way for opportunistic infection to provoke active disease by organisms that ordinarily coexist in the body without ill effect. Previous to HIV, such infections were known to inflict harm only where the immune system had been compromised, as for instance after treatment by irradiation or with drugs designed to suppress the rejection of such foreign tissue as skin grafts or organ replacement. Infection sometimes also shifts the balance between partial defense and active disease, as it may do when tuberculosis is also present. We should note, too, that a rare genetic type of immunodeficiency may be present at birth. Hence the need for a distinction from the genetic condition by the cumbersome descriptive name of Acquired Immunodeficiency Disease Syndrome (AIDS) caused by the Human Immunodeficiency Virus.

With the passing of time, drugs have been rendered more effective and treatment continues to improve. Current drugs are also more palatable and less costly than initial treatments. They reduce and repress the viral population to a low level, and thereby allow recovery of the defensive CD4 cells to adequate levels of both numbers and function. The ultimate object of "cure," however, has, at the time of writing, not yet been achieved, and for the present, treatment must be lifelong. Pockets of virus not eliminated "hide" in the body. Sheltered thereby from the drugs in the circulating blood, the virus remains ready to emerge at any drug-free opportunity.

HIV also has the capacity to evade destruction by changing its form quickly and often. All active cells mutate, but HIV does so more often than do most known invaders. In the long run, the task of addressing the ability of the virus to mutate is even more daunting than its capacity to evade or hide. Although errors in the process of cell division are not unusual among viruses, HIV is exceptionally error-prone. Described innocently as "poor copy editing," the indirect and daunting effect of such "errors" has been to enhance virulence. With every defense put up by the human host, whether innate in the immune system or extrinsic as with therapeutic drugs, the attack of the virus can be effectively stemmed only until new mutations induce resistance, which will in time reduce the effectiveness of the drugs.

The innate defenses of an individual harboring the virus will often induce a period of quiescence during which the untreated infection is tolerated. Such a "latency period" can sometimes persist for a decade or more. Although factors affecting the duration of such latency are not well understood, the constitution of the host does play some part. As one example, a variant of the gene CCR5 apparently blocks viral entry into the CD4 cell. Possession of the variant would reduce susceptibility (Huang et al 1996).

More universal forms of defense reside in the polygenic elements of the Human Leucocytic Antigen (HLA) system. Everyone possesses a set of HLAs in diverse combinations unique to each individual. The virus, upon entry into the body either via the mucosal surface or the bloodstream, first encounters specialized cells that attach the viral antigen (the foreign "non-self") to their surface. These specialized cells carry the antigen to a site on the HLA system, and thereby trigger an appropriate defense against the antigen (such as the production of specific cytotoxic "killer" cells). Since the inherited HLA system shapes the form of individual responses, genetic variability endows some infected persons with defenses against an HIV attack stronger than those of others. Such protective mechanisms are thought to account for the prolonged HIV seronegative state sustained by certain highly exposed groups of sex workers in Kenya. Among some persons already seropositive, these mechanisms perhaps account also for unusually long latency periods and slow progress of the infection. Maternal infant transmission has provided a natural experimental site for observing and interpreting HIV-host interaction. Thus either paternal or maternal inheritance of HLA defenses may obstruct transmission of the virus both from infected pregnant women to their fetuses, and from nursing mothers to their infants at the breast (Kuhn et al. 2004a; Tiemessen and Kuhn 2007).

Epidemiologists Respond

From the outset epidemiologists and clinicians alike, puzzled as to how to meet the new unknown epidemic threat of HIV, agreed on the need for epidemiological study. Since at least the 1930s, two now familiar approaches to research design—the case-control snapshot at a point in time, and the long-term cohort prospective view—had been well studied, tested, and applied (as discussed in Chapter 17). In the face of the epidemic, both designs had been essayed in the field before any specific cause of the new manifestation of immune-deficiency had been recognized in the laboratory. Like John Snow before the time of Robert Koch, epidemiologists recognized the pattern of transmission of an infectious organism before the virus was identified.

While these projects proceeded, microbiologists set to work on molecular studies in search of microorganisms. The almost simultaneous discovery of an as yet unnamed virus in affected men in Paris by Luc de Montagnier at the Pasteur Institute (Barre-Sinoussi et al. 1983), and by Robert Gallo at the National Institutes of Health in Bethesda (Popovic et al. 1984) put the question of priority and credit in dispute. A distinguished refereeing committee set up to resolve this *cause celebre* recommended that credit be shared. Both contenders agreed, the matter was resolved, and research studies continued unimpeded. From this stage on, an infected person could be quickly identified, at first from antibodies to the virus circulating in the blood, and later by direct study and counts of the viral load.

Several cases of Kaposi's Sarcoma (KS), a rare tumor first described decades before among African and Mediterranean peoples, emerged afresh among gay men in association with HIV infection both in California and New York City. At first taken to be an integral symptom of the new disease of HIV, the sarcoma was later recognized as a separate condition owed to an opportunistic infection (by the virus Herpes 8) attendant on immune-deficiency (Moore and Chang 1995a; Moore and Chang 1995b). Before either the virus causing AIDS or Kaposi's Sarcoma had been identified, a case-control study of Kaposi's Sarcoma had been conducted among gay men in New York City to explore a possible causal role for the recreational drugs known to be commonly used among them. No support for this hypothetical cause was found (Marmor et al. 1982). One may conclude that such associations as were found were most likely owed to confounding (Morabia 1995).

To pursue further the search for risk factors, the National Institutes of Health funded two longitudinal cohort studies among populations presumed to be at high-risk. In the city of San Francisco, one study recruited openly "gay" men. A second and larger study (assigned the acronym MAC) recruited self-identified gay men from the cities of Los Angeles, Chicago, Pittsburgh and Baltimore. Since both studies recruited "convenience samples," by any definition neither could be truly representative. The choice of such samples was perhaps justifiable, however, in that sexual orientation among homosexual men was the single risk factor for HIV so far apparent. Less justifiable was the skewing of the study populations. In an apparently unintended lapse, men of color were excluded. But women were also excluded, in this instance with intent, although some were assuredly at risk from gay marital partners. So, too, the study omitted men from the epicenter of New York City. Despite doubtful justification for those omissions, the follow-up yielded much relevant information to then sketchy understanding (Kaslow et al. 1987).

To plot the onset and probable clinical course of the still mysterious disease, some statisticians undertook a model-building approach (Zeger et al.

1989). To project the incidence of any disease, a starting point is a necessary datum. Yet with HIV, given both the prevailing high frequencies among gay men of sexual contact and multiple sex partners, those infected could seldom provide even a rough estimate of the date of infection. The difficulty of determining duration was the greater because of the variable and potentially long-lasting latency period of the virus. *"Back-calculation"* was a device invented to cope with the problem of the missing starting date. This calculation approximates the starting date of infection inferred from many successive observations of the march of the disease in the study population, and from these, projects the estimated shape of the curve of advancing disease towards full-blown AIDS. Subsequently, technical advances have yielded better estimates of duration of infection deduced from biological markers (e.g., Janssen et al, 1998).

Later, models perhaps ultimately of more lasting value and importance for public health have sought to predict the likely spread and expansion of HIV disease in populations under differing conditions (Anderson and May 1991; Blower et al. 2005). These models gave some guidance in the devising of public health strategies. Appropriately, the relevant formulas and their interpretation have entered epidemiological parlance.

In the following example (Figure 23.2), one might weigh the efficacy of one public health strategy against another. This analytic model, using available data from the seriously affected province of KwaZulu-Natal, South Africa, predicted the impact of alternative strategies for allocating antiretroviral treatments (Wilson et al. 2006). The model was built on detailed and specific clinical and demographic data to answer the question: should treatment strategies give priority to treatment equity, or to utilitarian efficacy? The researchers found that concentrating treatments in the large city of Durban, as opposed to spreading it spatially across large thinly populated rural areas, would as expected prevent the greatest number of new HIV infections and AIDS-related deaths. Surprisingly, it would also generate the lowest level of transmitted resistance and, in the result, maximize utilitarian efficacy. However, the urban concentration would accentuate disparities in health care between urban and rural areas and thereby undermine overall equity in treatment. Figure 23.2 shows the urban-rural differences between infections prevented, transmitted resistance, and AIDS deaths averted, should treatments be concentrated in Durban.

In any group under study with regard to chronic infectious disease such as HIV, an essential is to establish incidence frequencies both for predicting prevalence and for monitoring the effect of preventive strategies. Despite major advances at the molecular and clinical levels, to establish incidence still presents a challenge for public health agencies. The episode in which

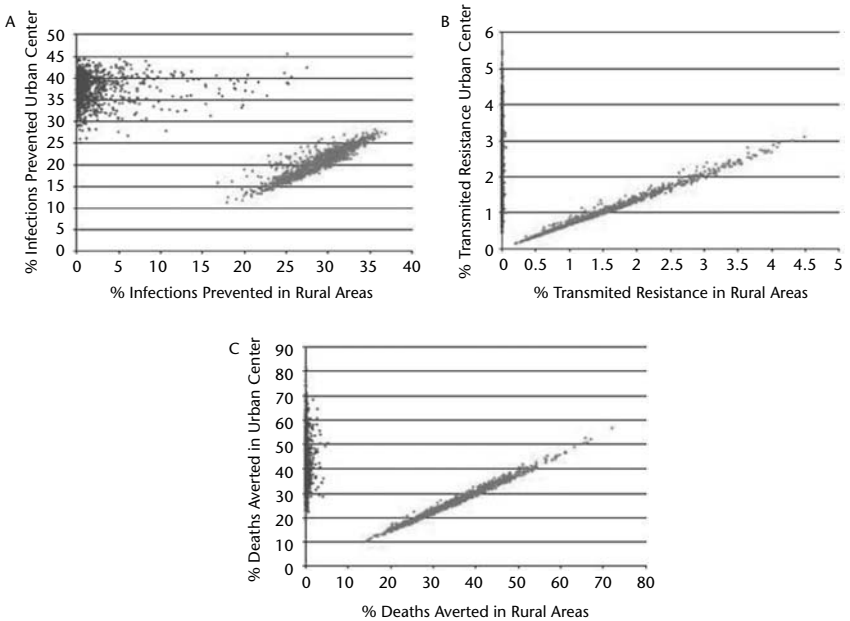


Figure 23.2 Impact of Drug Allocation Strategies: Urban and Rural (Wilson et al. 2006). The figure summarizes a complex set of comparisons, given the assumption that infected people living in Kwazulu/Natal were treated 2004–2008. The black dots refer to a utilitarian urban-only strategy; the gray dots to an egalitarian rural-urban mixed strategy. The vertical axis shows the outcomes for urban residents; the horizontal axis those for rural residents. *Graph a* of the figure summarizes outcomes of prevention: new infections averted as the result of reduced transmission of infection among the treated. *Graph b* indicates the potential for treatment–resistant viruses that would be transmitted. *Graph c* indicates deaths that would be averted with treatment. For each of the outcomes, the "utilitarian" (black dots) urban-only strategy does better than the "egalitarian" (gray dots) urban/rural strategy. Thus "utilitarian" trumps "egalitarian" in terms of societal benefit to the province, although at the cost of greater rural-urban inequities. Reprinted with permission and adapted from: Wilson DP, Kahn J, Blower SM (2006). Predicting the Epidemiological Impact of Antiretroviral Allocation Strategies in KwaZulu-Natal: the Effect of the Urban-Rural Divide. Proc Natl. Acad Sci 103 p. 14231.

infection is acquired is often asymptomatic, excepting perhaps some nonspecific and ill-defined reactions. Moreover, the interval between acquiring the infection and the manifestation of the typical clinical symptoms of HIV infection tends to be highly variable. Still, at such strategic sites as prenatal clinics, or treatment facilities for sexually transmitted diseases and emergency rooms, routine anonymous testing can contribute to maintaining some degree of epidemic surveillance. Thus these data can yield

prevalence rates maintained over time and controlled for age, sex, and location. They can also yield useful although indirect estimates of a rise or decline in incidence. For instance, well-controlled prevalence data for HIV from South India point to a credible decline in incidence among the affected Southern Indian states (Kumar et al. 2006). Such a decline could arise either with an antecedent decline in incidence or, in the grim circumstance where mortality rates exceed incidence, with a rise in deaths. In this instance, the happier inference would be that where preventive strategies began early and were energetically sustained, these data really do reflect gains on the epidemic by means of public health efforts. By contrast, in North India where prevention began later and was less intense, prevalence showed no decline.

Nowadays, a positive test for HIV can help distinguish a recent infection from one acquired further back in time (Janssen et al. 1998). Use of this "denatured" test procedure enables surveillance data to approximate incidence more closely. Those most susceptible to HIV and at highest risk of infection are, as one would expect, likely to be infected sooner. Moreover, viral load and hence transmission rates of the virus are higher among those recently infected. This may well explain the rapid rate of spread when sexually active young men unattached and previously unexposed are recruited for military ventures at distant sites with convenient access to commercial sex workers; or, as in South Africa, where men have long-term commitments to labor in the gold mines far from their homes and families; or, in those societies where men customarily maintain a number of concurrent sexual partnerships, all circumstances conducive to spread (Morris and Kretschman 1994, 2000). As the epidemic matures, the rapid spread of infection in the early phase of HIV epidemics has tended, if not everywhere, to decline. One explanation is that with growing awareness of risks, many men and women have decreased the number of their successive and concurrent partnerships (for a discussion, see Epstein 2007).

The highest risk of sexual transmission of the HIV infection is in the early stages of the disease, a state seldom obvious so that antiretroviral treatment, which undoubtedly reduces transmission, will ordinarily not be instituted until the appearance of symptoms at later stages. Thus acute early infection provides an inapparent route for the spread of infection to others. Not surprisingly, even with widespread treatment no decline in incidence is likely to be sustained: hence preventive efforts must be pursued widely, persistently, and intensively. Without such sustained effort, even if incidence rates appear to level, they are unlikely to remain low. Tenacious HIV epidemics can persist unless active prevention programs are sustained (Shelton et al. 2006). Some options for such programs are discussed below.

Susceptibility

In the United States, early studies of heterosexual encounters reported that women were more susceptible to HIV infection than were men (Guinan and Hardy 1987; Johnson 1992; People with AIDS Coalition of New York 1998; Burger and Weiser 2001). In Rakai, Uganda, however, in a large, defined, and closely monitored population under regular surveillance, among new cases rates of transmission from infected partners, whether man or woman, were roughly equal (Gray et al. 2001). The source of this difference between studies remains unexplained. A difference among viral subtypes in different populations proposed as an explanation seems not to have been confirmed. Small numbers, or unreported differences between men and women in sexual exposures to multiple partners, are other possible sources of difference among studied populations. Contributing factors could be the difference in prevalence of other infections of the uro-genital tract. Thus Bacterium Vaginalis, which among southern African women seems more prevalent than elsewhere, has been recognized as a risk factor for HIV among them (Myer et al. 2005). However, differences in prevalence between women and men may well be owed less to the biology of sex than to the gender inequality described later in this chapter.

Variations in prevalence rates among younger and older people have also been attributed to age differences governed either by biological susceptibility to infection or by immune deficiency. Given the present state of knowledge, neither biological nor social differences can be ruled out. Susceptibility to the virus undoubtedly varies with exposure at different mucosal sites. Thus risk of infection is greater with rectal than with vaginal exposure, and least with oral exposure. Rough sexual encounters that damage the vaginal or anal mucosa will raise the risk of infection still further. Younger women are thought to be more susceptible than older, because the entrance to the cervix acquires more protection with maturity. As discussed below, among adult men circumcision has been shown to provide for them a degree of protection from HIV infection.

Prevention, Trials, and Impact

Research aimed at preventing the further spread of HIV infection by means of the technical defense afforded by vaccines continues, although so far without effect. Despite the many previous successes with vaccines against other infections, such a development against HIV must overcome many obstacles. Even the longstanding efforts to curb such organisms as malarial parasites and tubercle bacilli have not as yet succeeded in overcoming all the obstacles

to such enterprise. With HIV, a particular difficulty is the inherent facility of the virus in mutating. Another is the fact that HIV directs its attack specifically against the defensive cells of the immune system.

Current exploratory efforts with vaccines aim to stimulate both humoral responses by antibodies and cellular responses by so-called "killer lymphocytes." Some facts serve to sustain a degree of mild optimism in this research field, however. Thus a few individuals have been found to resist infection even in the face of repeated exposure to HIV (described as "*highly exposed persistently seronegativ*e"); others, infected and seropositive, yet remain asymptomatic (described as "*longtime non-progressors"*) (Markel 2005; Cohen 2005; Letvin 2006). Effort is increasingly and appropriately focused on the nature of the immune response. Thus researchers pursue the means to induce antibody production, to provoke cellular immunity and, ultimately, to control if not destroy the ever-changing and elusive virus.

Apart from vaccines, attempts are under way to prevent infection by the use of known therapeutic drugs taken either shortly before or shortly after exposure. A now considerable body of data supports the effectiveness of the latter procedure, as for instance, when treatment follows immediately after a needle-stick contaminated by infected blood. Preexposure protection conferred by therapeutic drugs has been observed in animal experiments, and taken either orally or applied locally to the genital tract, such drugs are presently under test among healthy adults.

For the present, however, the available means for preventing the sexual transmission of HIV are limited to counselling, either about the number or selection of partners, or about the use of physical barriers. The male condom has of course long been widely available, although the promotion of its routine use has often met resistance.[5] The female condom, although too little used so far, provides women with a much needed barrier and a means of self-protection (Susser I and Stein 2000; Hoffman et al. 2004). Health authorities face important tasks in promoting such barrier methods, in ensuring their availability, and in encouraging their use. Counselling, whether about behavior or the use of protective devices, must of course be informed by knowledge of local community mores as well as the technicalities of self-protection.

Trials of barriers as a preventive against HIV illustrate some of the difficulties that face tests of efficacy. These include ethical issues, research design, implementation, and data analysis, whether the test involves chemical microbicides, or physical devices in the form of diaphragms, traditional or modified (Lagakos et al. 2008).

[5] The male condom is purportedly an eighteenth-century invention, its origin buried in history and its name attributed by some (if doubtfully) to an Englishman by the name of Condom.

First, in regard to the design in all such trials, ethical imperative demands of research teams that they instruct participants to use male condoms in all sexual encounters. While this "imperative" is understandable in that it highlights the need to protect women in the trial from infections, it must be said at the same time that it undermines the intended objective of testing and expanding resources for their protection. Thus, the prime purpose of such a trial is to provide an alternative means of protection for those women who cannot persuade their partners to use condoms. If indeed those women so instructed do act as advised and use a condom, then given the protection rendered thereby, there is no way in which the effect of any additional barrier under trial can be assessed. In practice, of course, seldom will women have been able to ensure the invariable use of the condom. For researchers, in such circumstances doubt must reside in reliance on the woman's report alone as to whether, and how often, the condom as well as the barrier under test was used. That condom use is certainly not invariable is evident from the rates of pregnancy in nearly all such trials (for example, 13% in a diaphragm trial in South Africa and Zimbabwe; (Padian et al. 2007). Sexual congress takes place in private, and use of either the condom or any other barrier under trial cannot ordinarily be known with certainty.

A second issue, this one regarding the implementation of a trial has been controversial. The issue is whether to depart from the rule that both participants and researchers remain strictly "blind" as to whether the vaginal gel randomly assigned to each subject is the putative microbicide or the placebo. Thus one investigator proposed adding to the trial a "condom only" arm as a third unblinded component (van de Wijgert et al. 2005). Such an addition of course not only sacrifices the requisite blindness but adds complexity, expense, and time to the execution and analysis of trials. Nevertheless, this addition has been undertaken in one trial of a vaginal microbicide. Whether the departure from the blinding of participants and researchers as well as added cost and time are justified, remains unresolved.

Still other issues of analytic procedures have aroused controversy. These followed from the direction of "intention to treat" established as standard therapeutic trial practice. This requirement means that every eligible participant, once having given consent, and been recruited to the trial, must be included as a subject in the analysis of results. Whether she drops out of the trial, and whether she reports appropriate use of either the relevant treatment or preventive, she remains a subject to be accounted for in the trial. Should many participants default, one might perhaps consider the end result a test of both behavior and the barrier under test combined, but not of the efficacy of the barrier under trial. In a recent test of the vaginal diaphragm, the reported behavior of the participating women made clear that adhering to the analytic procedure of Intention to Treat

did not provide a satisfactory test of efficacy (Padian et al. 2007; Stein and Glymour 2007).

The use of microbicides was first proposed as a protective barrier against sexually transmitted disease (Stein 1990). Still under test almost two decades later, however, efficacy has not yet been established (Harrison and Lamphear 2005). For the present, indeed, the efficacy of any physical or chemical barriers applied either vaginally or by means of a cervical ring remains unproven. Although the advent of an effective chemical or physical barrier less obtrusive than the male or female condom will surely be a major asset for prevention, it will not be without problems. For instance, clandestine use of a barrier, if discovered, could strain a sexual relationship. This would perhaps be more likely to be the case if the barrier is applied at the time of intercourse and less likely if, like the cervical ring, it is more or less permanently in place. Vaginal gels enhanced by therapeutic drugs like Tenofovir must be absorbed if they are to be effective, hence will be put in place some time before or after intercourse. In fact, a Tenofovir gel probably has more in common with pre- and postexposure preparations taken orally than with other chemical barriers. The behavioral implications are clearly different for a visible physical barrier, or for an invisible one discretely applied as a deterrent to infection that is not actually present in the vagina at the time of intercourse (as with the cervical ring), a product that depends on absorption from the vaginal tract but is or is not inserted at the time of intercourse. None of these methods is likely to confer absolute protection and as with all partially effective strategies, intensive and widespread health education would retain its key role in prevention.

Circumcision among adult men has long been thought to offer some protection against sexually transmitted disease. In three separate controlled trials among adults, the procedure has now been shown to protect the man to some degree. Each trial reported a surprising and substantial reduction (of up to 60%) in risk of HIV infection in the follow-up period among the circumcised as compared with those uncircumcised (Hargreaves 2007; Bailey et al. 2007; Newell and Barnighausen 2007; Gray et al. 2007). Among the women partners of the circumcised men in the trials, a hoped-for reduction remains to be established. Among HIV positive men, however, intercourse before the circumcision wound was fully healed raised the risk of HIV transmission to the female partner (Wawer 2008). Nonetheless, where a reduction of infection rates is achieved among men, reduced transmission of HIV to women partners is bound to follow. Although male circumcision apparently provides a potent preventive strategy against HIV transmission, difficulties in its application on the social, economic, and medical fronts are to be expected.

In the absence of an effective vaccine against human immunodeficiency disease, of health education in sexual matters, and of a norm of communication

between partners, lacking in many societies, testing for infection becomes a critical factor essential to sustaining the public health. Widespread testing may well discover many who are unaware of their infected state. Systematic search, testing and identification could steer them into appropriate services to manage their treatment, and ameliorate the danger they constittue especially to their intimates. Meanwhile, among the infected antiretroviral therapy can play a part in reducing viral load and thereby the risk of transmission, as well as in halting the advance of the disease. It is an irony that potential risk of spreading the disease resides in the fact that effective treatment of those infected by HIV prolongs their lives but also adds to their capacity further to spread the disease. If population incidence is to be reduced, treatment needs to be accompanied by widespread and vigorous education, communication, and behavior change that promotes the conscientious use of protective barriers.

Prevention of Infection in the Infant

As noted above, in the United States a surprising 10 years elapsed before spread of the new disease began to be recognized in women. Soon thereafter, a further important mode of the spread of the deadly new virus was recognized: infected women transmitted the virus whether to the fetus during pregnancy, or intrapartum, or to the infant at the breast. Among pregnant women infected by HIV and untreated, some 35% among them might transmit the organism to the fetus either in utero, or intrapartum, or by breastfeeding. A number of circumstances favor or inhibit transmission from mother to infant. Thus the risk of such transmission is greater in mothers with either high viral blood counts of the virus or low counts of protective CD4 blood cells. During delivery, risk is raised by early intrapartum rupture of the membranes and the prolonged exposure of the infant to the maternal birth passages that is likely to follow. The prematurely born infant is also at raised risk.

Steps to prevent the transmission of HIV from mother to infant have evolved as new research findings emerged, although as always their application is determined by the available socio-economic and technical resources. Thus in developed countries, with comprehensive health services, and readily available antiretroviral therapy, effective intervention and management has greatly reduced mother to child transmission. Testing, diagnosis and antiretroviral treatment of an HIV-infected mother during pregnancy, delivery by planned Caesarian section, and elimination of breast feeding has almost entirely prevented viral transmission to the newborn infant (Townsend et al. 2008).

Early studies of maternal-infant transmission of HIV in New York, Paris, Spain, Italy, and the Congo soon began to report an added risk for infection from breast-feeding. In the US and Europe, therefore, the World Health Organization (WHO) recommended formula feeding only for infants of HIV infected mothers. In sub-Saharan Africa, however, WHO recommended breast-feeding. This was reasonable advice in anticipation of the social, practical, and economic obstacles to formula feeding, taken together with the potential ill-effects of the lack of hygiene likely to attend poor and primitive conditions for food and feeding. WHO later relaxed this recommendation to include such alternatives to breast-feeding as were "feasible, affordable, accessible, safe and sustainable."

According to at least two studies so far, however, within the first four months of life *exclusive* breast-feeding (as compared with mixed (breast milk plus other foods or fluids) can substantially reduce although not entirely eliminate maternal transmission to the infant born uninfected (Kuhn et al. 2004b; Coovadia et al. 2007). Where resources are lacking, as in southern Africa, exclusive breast-feeding of infants avoids the ever likely bacterial infection from contaminated water, food and utensils, and the more so in the primitive circumstances of those mothers most at risk. These infections in turn may, enhance susceptibility to HIV.

On the other hand, for the HIV-afflicted mother the risk of viral transmission to the infant in breast milk can be raised by a reduced count of protective CD4 cells, or by mastitis and other infections of the breast (Kuhn et al. 2005; Coovadia et al. 2007). Research into antiretroviral drugs under way at the time of writing gives some promise of protection for the breast-feeding infant by treating either mother or infant or both (Abrams and Kuhn 2008).

In low-resource countries at this juncture, what the best feasible practices for infant feeding might be aside from the cost, allows no simple injunction. Breast milk is not only everywhere affordable, but undeniably the healthiest option for the infant, nutritionally and because of the immunological properties it deploys. Moreover a high risk of infection attends the use of formula prepared with impure water. In Africa south of the equator, nonetheless, along with breast-feeding in the early months of life, a strong tradition persists that is not easily overcome of adding water as well as cooked cereals or even adult food. In rural areas, of necessity, the water is often scooped from contaminated pools or streams.

To effect prevention of maternal transmission of HIV to the infant before birth essentially requires all infected pregnant women to receive care, and to be tested for HIV, syphilis and tuberculosis, and both to comprehend and accept the diagnosis of their seropositivity. In countries with few resources, however, in face of the HIV epidemic the task of counselors assigned to

assist women and families in countering both stigma and tradition is often far from simple. Departure from tradition becomes all the more difficult when a practice newly advocated may at once identify and stigmatize an HIV-infected mother.

Thus, although policies related to breast-feeding have generated much controversy, at least a rough consensus does now exist (Fowler 2008). In the less-developed world, given the extent of the HIV epidemic and the extreme poverty of the many afflicted women, in order to secure infant survival it is well to ensure antiretroviral treatment during pregnancy and parturition for every infected woman. After parturition, excepting those mothers with advanced disease, exclusive breast-feeding is advisable for at least the initial six months of infant life. Thereafter, weaning should be gradual. Some feeding at the breast could be continued into the second year but will need to be supplemented (Kuhn 2008).

Prevention of maternal transmission to the infant before birth essentially requires that all pregnant women receive care, are tested for HIV as well as for other antenatal routiness, and comprehend and accept the diagnosis of their seropositivity. In countries with few resources, however, the task of counselors assigned to assist women and families in countering stigma, tradition and low resources, is often far from simple. Some women seemingly lacking the skills and resources for successful formula feeding will elect to do so nonetheless. Others, obliged to work and earn outside the home, may find themselves unable by circumstance (which in favorable situations could, in fact, be modified) to sustain breast feeding (though many workplace environments could be and some are easily modified to accommodate breast feeding mothers). Indeed, counselors themselves become torn and anxious about their roles (Buskens and Jaffe, 2007). As with so much in dealing with the affliction of HIV, the personal and behavioral demands on personnel serving afflicted populations can be both stressful and demanding.

Social Factors

In the mid-twentieth century, the broad sweep of public health and medicine attracted study by a growing corps of sociologists, social anthropologists, and psychologists (see Chapters 15 and 20). In the US and Britain several joined epidemiologists and socially oriented physicians in studies of the effect on health of societal forces, cultural adaptations, and human behavior. In the last two decades of the twentieth century the advent of HIV reinforced this connection.

As indicated, many societal factors contribute to the HIV epidemic (See Table 23.1). Some causes of disease are embedded in the structure and

culture of societies: among these are socioeconomic development, customary behavior and social institutions, governments and social policies, the nature of occupations that demand absence from the home, the ties that bind families and the factors that fragment them, the educational and religious influence of upbringing, and the availability of adequate health services. All these and more have been invoked as factors underlying both the spread of HIV and its impact on societies and individuals. Here we comment on only two among these, neither of which is entirely independent of the other. First we describe influences related to the family, next the question of gender inequality.

A crucial factor in sexually transmitted disease arises with population movements that loosen family bonds. Sailors, soldiers, commercial travelers, itinerant craftsmen, and migrant workers have long been known for their high rates of sexually transmitted diseases, as have the women with whom they consort. Sidney Kark, in his classic paper on the epidemiology of syphilis in the native rural population in South Africa, clarified the social pathology underlying its epidemic source (Kark 1949; Kark 2003). Through the twentieth century, huge numbers of rural black men were recruited as laborers for annual nine-month working spells in the depths of the gold mines a mile underground. There they lived in barracks, dependent for sex on women who had likewise migrated to the gold mines surrounding the city of Johannesburg, a social process that created an epicenter of sexual disease. In recent decades in these same South African mines, HIV has replicated the history of older sexually transmitted diseases. The social and economic forces behind these demographic movements are various (Campbell and Williams 1999, Lurie 2003). While to resolve such processes is perhaps outside the familiar domains of public health, still their contribution to incidence and spread of serious disease needs to be documented, understood, and contained.

Table 23.1 Societal Factors that have Influenced the Global HIV Epidemic

SES inequalities
Gender inequalities and lack of empowerment among women and sexual minorities
Migrant labor practices and impact on family life and relationships
Access to health information, sex education, and methods of protection
Male circumcision (decreases transmission to the man)
Stigma and discrimination against vulnerable groups
Age of sexual debut
Number and concurrence of sex partners
Prevalence of other sexually transmitted diseases
Prevalence of injection drug use
Societal norms for HIV screening, testing and treatment programs, including antenatal, infants and children, adult men and women
Cultural and religious attitudes and influences on sexual behaviors

A second ubiquitous factor is captured by the phenomenon of gender inequality. In Africa women, both the young and the married, are those at highest risk of infection. Men and women who have many and especially concurrent partnerships are obviously at greatest risk, often compounded by the like presence of other sexually transmitted diseases. The commonest route of HIV transmission is from one partner to another. In both modern and traditional societies, and both within and without marriage, men tend to play the dominant role in sexual encounters. Hence the commonly reiterated advice, "use a condom," refers directly to the man. And if, as seems true in many cultures, married men have more sexual partners than do married women, whether within traditional polygamous marriages, or extramaritally when monogamy is the legal norm, then their role in bringing infection to wife and children is much the greater. Women, even those who are strictly monogamous, may find themselves infected, may then infect their offspring, and will generally bear the main brunt and burden of stigma. Then, too, women who take many partners, sometimes as an essential prop to the household economy, will be exposed and infected with tragic frequency and at younger ages. For as long as they are able women carry the main burden of caring for sick and dying family members and for orphaned grandchildren. Thus gender, the social allocation of male and female sex roles, takes on wider meaning in the face of the HIV epidemic. Poverty chains many women to their homes, and limits both their capacity to fulfill traditional family responsibilities of maintaining shelter, providing food, and sustaining the traditional roles and duties thrust upon them (Susser I 2002; Billson and Fluehr-Lobban 2005; Doyal 2006; Mantell et al. 2008; Susser I 2009).

By contrast with the historic past, in modern times women have risen to leadership roles even in the face of disadvantage, although most often from the ranks of the better educated and the higher social classes. In many countries, modern women are a growing force of advocacy for those in need of advice and care. In an Africa much afflicted by HIV, women have been ready volunteers in trials both of microbicides and of female condoms as preventives against infection. The devastation of the current epidemic will have altered and depleted many societies, but may yet also build up the strength of women to survive and to serve as a critical link for families in gaining access to and using health services.

Health, Education, and Behavior Change

Education about health aims to enlighten society and raise consciousness of avoidable risks to health, about the means available for the prevention

of ills, and how best to improve and sustain fitness. An encouraging example was the decades-long and ultimately successful public health campaign against the pervasive smoking of tobacco initiated by Dr. Thomas Parran, US Surgeon General in the 1950s and 1960s (US Department of Health 1964). The campaign faced the heavy odds posed by the widespread and then socially acceptable addiction to smoking tobacco, effectively reinforced by a rich and powerful tobacco industry. The industry stalled, braked, and countered all efforts at prevention and control, subtly enticed the young to join the mass of habitual and addicted users, and enlisted powerful political allies. Nonetheless, once the ravages to health in the form of lung cancer, bronchitis, and other lung disease caused by smoking were understood, national campaigns mobilized many sectors of society in the effort against the tobacco industry and the addictive smoking habit. Over several decades and in many countries, the campaign has effected a substantial reduction in cigarette smoking and, consequently, also in epidemic lung cancer, bronchitis, vascular and coronary heart diseases. These achievements, aided in the United States by a legal clampdown on advertising cigarettes and forceful health education, brought about a substantial reduction in both smoking and its serious health effects (Surgeon General 2000; Kean-Cowdin et al. 2000).

Heartening as is this example of behavior change with smoking, with attempts to limit the current HIV epidemic, very different issues arise. These issues include sexual orientation and behavior, behavior differentiated by gender, and issues of gender inequality. In those societies and those social strata that most lack education, material resources, and any degree of sexual equality, the effort to contain the HIV epidemic remains a prime individual and public health responsibility. The voluminous literature on these complex issues provides no simple, uniform path to effective action. Creative initiatives, depending on the array of favorable or unfavorable local forces and their evaluation, should spur wider understanding and improved defenses against the pandemic.

In the face of the HIV epidemic, the intimacies of sex and sexual behavior called for much more than simple primary observations of the distributions of the new plague. Sexual proclivities and behavior had to be identified and described, never an easy undertaking although, as late as the 1940s, the remarkable biologist Alfred Kinsey had measured and shown what could be done about these essentially social questions of behavior (Kinsey et al. 1948). With the advent of HIV, as the virus rapidly penetrated the geographic bounds of Central African forests and the ensuing disease spread from its origin among the chimpanzees of the Congo to cross all continents, the need to extend the epidemiological endeavor and follow the course of the new international scourge took on critical importance.

Sexual orientation and sexual roles have assumed new significance in furthering both the understanding of an unprecedented viral disease and the strategies for its prevention and treatment. With the advent of HIV, effective international work could not be of that narrow kind where studies rely mainly on biological specimens grown in culture and studied under the microscope. The culturally appropriate investigations of social anthropologists were an essential resource both in discovering the origin and in tracking the spread of the newly discovered plague. Thus, those epidemiologists who would launch controlled trials aimed at either prevention or treatment needed to comprehend and consider more than the routine considerations of "intention to treat," random assignment, the blinding of participants and researchers to the condition and circumstances of their subjects, and statistical issues of power and appropriate analysis.

While major opposition to the anti-smoking campaign of the 1960s had emanated from the tobacco industry as the source of that great plague of smoking and the attendant lung cancer, with HIV opposition to sex education generally stems from traditional beliefs and values. The rules are set thereby. Thus virginity testing, common in some sub-Saharan communities, stems from built-in social inequalities between the sexes: the general rule, incompatible with reality, is that boys but not girls are free to engage in premarital sex. On the other hand, those same communities may well encourage pregnancy in young women as a guarantee of the fecundability that is a prerequisite for marriage. Nor are these apparently contradictory requirements by any means limited to less-developed societies. Thus, in the twenty-first century of the Christian era, in the United States the regime has sought to narrow education about sex for youth to no more than the demand for abstinence before marriage. Undoubtedly, this Comstock-like[6] endeavor contributes to a relatively high rate of teenage pregnancies in the US as compared, for instance, with sexually liberal Sweden, even though age at sexual debut is not much different across these two countries.

With regard to effecting behavior change as protection against HIV, the influence of beliefs can operate subtly through stigma, or openly through either state or religious education as, for instance, it does in the US. Nor are the pressures of belief and culture necessarily confined to the localities in which they take origin. Again, as with the historic example of tobacco, global factors play a part. Thus the generous funds donated by the US and intended to counter HIV in low-resource countries may carry with them

[6] Anthony Comstock (1849–1915) was a "moral" crusader whose efforts were devoted to suppression of obscene literature, and founder of the New York Society for the Suppression of Vice. He was responsible for the destruction of 160 tons of literature and pictures (1927).

ideologically-driven restrictions, for example, in emphasizing abstinence, and in the disapproval of commercial sex. Such restrictions run counter to the overall recognition that the aims of changing behavior must start from understanding and respect for the local cultural setting. These same restrictive ideologies oppose funding of such newer preventive methods as, for instance, the female condom that either a woman or a man who engages in sex might initiate. Given the lack of major support for the female condom from foundations and governments, both the female condom and its male counterpart "AEGIS" specifically designed for rectal sex, have so far failed to make the impact on prevention of sexually transmitted disease that otherwise they might do (Mantell, Stein, Susser 2008; Susser I 2009).

Two connected but separable elements bound up with this disease have been fundamentally challenging to public health responses. These are "stigma" on the one hand, and "exceptionalism" on the other. Stigma, and the consequent "spoiled identity" of social disgrace (Goffman 1963), quickly attaches to the infected person, whether among family, social groups, or even health providers. In the early days of the HIV epidemic, especially in the US and Europe, stigma stemmed from the association of the disease with two categories of "disgraceful" behaviors, namely intravenous drug use and gay sex among men. As the HIV epidemic spread, extramarital sex too enhanced the attached disgrace. The ominous and rising risk of transmitting to others a new disease that was then uniformly untreatable and fatal enhanced the degree of stigma all the more.

In the pretreatment era, in different times and places, stigma varied in intensity and form. Health providers in the US, for instance, assumed from the beginning that knowledge of the infected status of individuals needed to be kept strictly confidential. HIV/AIDS was to be both a secret and a dread disease. Given the long and variable incubation period of such an irremediable and stigmatized disease, nothing but harm for the victim could follow from a diagnosis made, and then made known to others. Disclosure threatened job, home, family, and insurance, as well as health and life. Thus the new rule required that blood could only be tested for HIV infection given formal procedures of counseling both before and after testing, nor could the professional health advisor inform the partners of infected persons of the risk of infection. To describe the implications of these new rules from the perspective of both health departments and personal physicians, Ronald Bayer coined the term of "HIV exceptionalism" (Bayer 1991). In time, as effective medication became available and somewhat curbed the spread and the fear of the disease, its exceptionalism and the attached stigma declined in importance. Regardless of such social impediments, to obtain needed knowledge remained an imperative from the personal, social, epidemiological, and public health perspective.

Conclusion

A paradigmatic change in epidemiology was wrought by the advent of the HIV epidemic, which provides the matter of this chapter. In the light of the many individual and social elements encompassed, it is aptly described as holistic or better, as eco-epidemiology. No single concept suffices: neither individual humors nor the environment (airs, waters, places in the formulation of Hippocrates), nor the mode of transmission or contagion. Physicians first learned of a sexual mode of transmission of disease from Fracastoro in Italy in the last decade of the fifteenth century when confronted, upon the return of Columbus from his trans-Atlantic voyaging, by the onslaught of syphilis from country to country across Europe. Earlier in the face of the thirteenth-century plague and the conflict it generated between citizen pressure groups and the City Fathers, the Venetians had invented the 40-day quarantine. At the turn of the eighteenth century, we acquired from the French revolutionaries the concept of health as a human right, independently reinforced in Germany later in the century by Jacob Henle and then by Rudolph Virchow. Early in that century, Mary Wollstonecraft had fought for the vindication of the rights of women as, in the late nineteenth century, Olive Schreiner followed her with the equally cogent message that the progress of society depends on both men and women working together; from Villermé and Parent-Duchatelet, Edwin Chadwick and his ally Florence Nightingale, as also cogently from John Simon, we learned the potential power of conferring defined local responsibility for the public health; from Sidney Kark in twentieth-century South Africa we learned to observe the complex interaction of the needs of industry, migrant labor, the family, and sexual health. Cicely Williams, in confronting malnutrition in the West African toddler, recognized the critical role of the mother-infant bond in sustaining child health under the most adverse conditions. From experiences with controlled trials, we learned to appreciate the ethical constraints on research and interventions: whether the problem relates to diet, adherence to medical and health advice, or sexual behavior.

On the larger societal scale, we must ask how social inequalities in health arise and what their impact is both within and across societies. We need to study the constraints and benefits on health of global and local economies. At the same time, outside of science, we need to understand the effects or implications that reside in traditional, faith-based, or universal interpretations of health and disease. An ecological perspective entails, for epidemiologists, an approach to the study of disease in all its dimensions from individual to societal, and that aims to comprehend each component as part of the whole (see Chapters 24, 25, and 26).

References

Abdool Karim SS, Abdool Karim Q, Gouws E, Baxter C (2007). Global epidemiology of HIV-AIDS. *Infect. Dis. Clin. North Am* 21:1–17, vii.

Aceijas C, Stimson GV, Hickman M, Rhodes T (2004). Global Overview of Injecting Drug Use and HIV Infection Among Injecting Drug Users. *AIDS* 18:2295–2303.

Anderson RM, May RM (1991). *Infectious Diseases of Humans Dynamics and Control.* Oxford: Oxford University Press.

Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CF, Campbell RT, Ndinya-Achola JO (2007). Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* 369:643–656.

Baker BJ, Armelagos GJ (1988). The origin and antiquity of syphilis: paleopathological diagnosis and interpretation. *Curr Anthropol* 29:703–738.

Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J et al. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for Acquired Immune Deficiency Syndrome (AIDS). *Science* 220:868–871.

Bayer R (1991). Public health policy and the AIDS epidemic. An end to HIV exceptionalism? *N Engl J Med* 324:1500–1504.

Bayer R, Oppenheimer GM (2000). *AIDS Doctors Voices From the Epidemic.* Oxford: Oxford University Press.

Billson JM, Fluehr-Lobban C (2005). *Female Well-Being Toward a Global Theory of Social Change.* London: Zed Books.

Blower SM, Bodine EN, Grovit-Ferbas K (2005). Predicting the potential public health impact of disease-modifying HIV vaccines in South Africa: the problem of subtypes. *Curr Drug Targets Infect Disord* 5:179–192.

Burger H, Weiser B (2001). Biology of HIV-1 in women and men. *Clin Obstet Gynecol* 44:137–143.

Burnet FM, White DO (1972). *Natural History of Infectious Disease.* 4th ed. Cambridge England: University Press.

Campbell C, Williams B (1999). Beyond the biomedical and behavioural: towards an integrated approach to HIV prevention in the Southern African mining industry. *Soc Sci Med* 48:1625–1639.

Cohen J (2005). Is an effective HIV vaccine feasible? *Science* 309:99.

The Columbia Encyclopedia (1927). New York: Columbia University Press.

Coovadia HM, Rollins NC, Bland RM, Little K, Coutsoudis A, Bennish ML, Newell ML (2007). Mother-to-child transmission of HIV-1 infection during exclusive breastfeeding in the first 6 months of life: an intervention cohort study. *Lancet* 369:1107–1116.

Coutsoudis A, Pillay K, Kuhn L, Spooner E, Tsai WY, Coovadia HM (2001). Method of feeding and transmission of HIV-1 from mothers to children by 15 months of age: prospective cohort study From Durban, South Africa. *AIDS* 15:379–387.

Doll R (1979). The pattern of disease in the post-infection era: national trends. *Proc R Soc Lond B Biol Sci* 205:47–61.

Doyal L (2006). How well are women worldwide? *Lancet* 367:1893–1894.

Epstein H (2007). *The Invisible Cure: Africa, the West, and the Fight Against AIDS.* New York: Farrar, Straus and Giroux.

Flam R and Stein ZA (1986). Behavior, infection and immune response: an epidemiological approach. In Feldman and Johnson, eds. *The Social Dimension of Aids: Methods and Theory.* New York: Praeger Press, pp. 61–76.

Fowler MG (2008). Further evidence that exclusive breast-feeding reduces mother-to-child HIV transmission compared with mixed feeding. *PLoS Med*. 5(3):e63.

Gao F, Bailes E, Robertson DL *et al* (1999). Origin of HIV-1 in the chimpanzee pan troglodytes troglodytes. *Nature* 397:436–441.

Goffman E (1963). *Stigma: Notes on the Management of Spoiled Identity*. Englewood Cliffs, NJ: Prentice-Hall.

Gordon D (1976). *Health, Sickness, and Society Theoretical Concepts in Social and Preventive Medicine*. St. Lucia, Q: University of Queensland Press.

Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F et al. (2007). Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *Lancet* 369:657–666.

Gray RH, Wawer MJ, Brookmeyer R, Sewankambo NK, Serwadda D, Wabwire-Mangen F et al.(2001). Probability of HIV-1 transmission Per Coital Act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *Lancet* 357:1149–1153.

Griffiths DJ (2001). Endogenous retroviruses in the human genome sequence. *Genome Biol* 2:REVIEWS1017.

Guinan ME, Hardy A (1987). Women and AIDS: the future is grim. *J Am Med Womens Assoc* 42:157–158.

Hahn BH, Shaw GM, De Cock KM, Sharp PM (2000). AIDS As a zoonosis: scientific and public health implications. *Science* 287:607–614.

Hargreaves S (2007). 60% Reduction in HIV risk with male circumcision, says WHO. *Lancet Infect Dis* 7:313.

Harrison PF, Lamphear TL (2005). Microbicides. In Mayer and Pizer, eds. *The AIDS Pandemic: Impact on Science and Society*. London: Elsevier Academic Press, pp. 190–235.

Hoffman S, Mantell J, Exner T, Stein Z (2004). The future of the female condom. *Perspect Sex Reprod Health* 36:120–126.

Hooper E (1999). *The River: A Journey to the Source of HIV and AIDS*. 1st ed. Boston, MA: Little, Brown and Co.

Huang Y, Paxton WA, Wolinsky SM *et al* (1996). The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat Med* 2:1240–1243.

Janssen RS, Satten GA, Stramer SL *et al* (1998). New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA* 280:42–48.

Johnson AM (1992). Epidemiology of HIV infection in women. *Baillieres Clin Obstet Gynaecol* 6:13–31.

Kark SL (2003). The social pathology of syphilis in Africans. 1949 (Reprinted). *Int. J Epidemiol*. 32:181–186.

Kark S (1949). The social pathology of syphilis in Africans. *S Afr Med J* 23:77–84.

Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR, Jr (1987). The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol* 126:310–318.

Kean-Cowdin R, Feigelson HS, Ross RK, Pike MC, Henderson BE (2000). Declining Cancer Rates in the 1990s. *J Clin Oncol* 18:2258–2268.

Kinsey AC, Pomeroy WB, Martin CE (1948). *Sexual Behavior in the Human Male*. Philadelphia: W.B. Saunders Co.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T (2000). Timing the Ancestor of the HIV-1 Pandemic Strains. *Science* 288:1789–1796.

Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *Br Med Bull* 58:19–42.

Kuhn L, Abrams EJ, Palumbo P, Bulterys M, Aga R, Louie L, Hodge T (2004a). Maternal versus paternal inheritance of HLA class i alleles among HIV-infected children: consequences for clinical disease progression. *AIDS* 18:1281–1289.

Kuhn L, Kasonde P, Sinkala M, Kankasa C, Semrau K, Scott N et al. (2005). Does severity of HIV disease in HIV-infected mothers affect mortality and morbidity among their uninfected infants? *Clin Infect Dis* 41:1654–1661.

Kuhn L, Stein Z (1997). Infant survival, HIV infection, and feeding alternatives in less-developed countries. *Am J Public Health* 87:926–931.

Kuhn L, Stein Z, Susser M (2004b). Preventing mother-to-child HIV transmission in the new millennium: the challenge of breast feeding. *Paediatr Perinat Epidemiol* 18:10–16.

Kumar R, Jha P, Arora P, Mony P, Bhatia P, Millson P et al. (2006). Trends in HIV-1 in young adults in South India from 2000 to 2004: a prevalence study. *Lancet* 367:1164–1172.

Lagakos SW, Gable AR, Institute of Medicine Commmittee on Methodological Challenges in HIV Prevention Trials (2008). *Methodological Challenges in Biomedical HIV Prevention Trials.* Washington, DC: National Academies Press.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Letvin NL (2006). Progress and obstacles in the development of an AIDS vaccine. *Nat Rev Immunol* 6:930–939.

Lurie MN, Williams BG, Zuma K, Mkaya-Mwamburi D, Garnett GP, Sweat MD et al. (2003) Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa. AIDS 17(15):2245–2252.

Mantell JE, Stein ZA, Susser I (2008). Women in the time of AIDS: barriers, bargains, and benefits. *AIDS Educ Prev* 20:91–106.

Markel H (2005). The search for effective HIV vaccines. *N Engl J Med* 353:753–757.

Marmor M, Friedman-Kien AE, Laubenstein L, Byrum RD, William DC, D'onofrio S, Dubin N (1982). Risk factors for Kaposi's sarcoma in homosexual men. *Lancet* 1:1083–1087.

Marx PA, Apetrei C, Drucker E (2004). AIDS as a zoonosis? Confusion over the origin of the virus and the origin of the epidemics. *J Med Primatol* 33:220–226.

Mayer J, Meese E (2005). Human endogenous retroviruses in the primate lineage and their influence on host genomes. *Cytogenet Genome Res* 110:448–456.

Meyer C, Jung C, Kohl T, Poenicke A, Poppe A, Alt KW (2002). Syphilis 2001—a Palaeopathological Reappraisal. *Homo* 53:39–58.

Moore PS, Chang Y (1995a). Detection of herpesvirus-like DNA Sequences in Kaposi's sarcoma in patients with and without HIV infection. *N Engl J Med* 332:1181–1185.

Moore PS, Chang Y (1995b). Kaposi's sarcoma findings. *Science* 270:15.

Morabia A (1995). Poppers, Kaposi's sarcoma, and HIV infection: empirical example of a strong confounding effect? *Prev Med* 24:90–95.

Morris M, Kretzschmar M (1997). Concurrent partnerships and the spread of HIV. *AIDS* 11:641–648.

Morris M, Kretzschmar M (2000). A Microsimulation Study of the Effect of Concurrent Partnerships on the Spread of HIV in Uganda. Working Paper. Population Research Institute, Pennsylvania State University.

Motulsky AG, Vandepitte J, Fraser GR (1966). Population genetic studies in the Congo. I. Glucose-6-phosphate dehydrogenase deficiency, hemoglobin S, and malaria. *Am J Hum Genet* 18:514–537.

Myer L, Kuhn L, Stein ZA, Wright TC Jr., Denny L (2005). intravaginal practices, bacterial vaginosis, and women's susceptibility to HIV infection: epidemiological evidence and biological mechanisms. *Lancet Infect Dis* 5:786–794.

Nahmias AJ, Weiss J, Yao X, Lee F, Kodsi R, Schanfield M et al. (1986). Evidence for human infection with an HTLV III/LAV-like virus in Central Africa, 1959. *Lancet* 1:1279–1280.

Newell ML, Barnighausen T (2007). Male circumcision to cut HIV risk in the general population. *Lancet* 369:617–619.

Nolan D, Gaudieri S, John M, Mallal S (2004). Impact of host genetics on HIV disease progression and treatment: new conflicts on an Ancient Battleground. *AIDS* 18:1231–1240.

Oleske J, Minnefor A, Cooper R Jr., Thomas K, dela CA, Ahdieh H, Guerrero I, Joshi VV, Desposito F (1983). Immune deficiency syndrome in children. *JAMA* 249:2345–2349.

Padian NS, van der SA, Ramjee G, Chipato T, de BG, Blanchard K et al. (2007). Diaphragm and lubricant gel for prevention of HIV acquisition in Southern African women: a randomised controlled trial. *Lancet* 370:251–261.

People with AIDS Coalition of New York (1998). Some Ailments Found Guilty of Sex Bias. *Newsline People AIDS Coalit N Y* 35.

Popovic M, Sarngadharan MG, Read E, Gallo RC (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* 224:497–500.

Rhodes T, Platt L, Maximova S, Koshkina E, Latishevskaya N, Hickman M et al. (2006). Prevalence of HIV, Hepatitis C and syphilis among injecting drug users in Russia: a Multi-City Study. *Addiction* 101:252–266.

Rosinska M (2006). Current trends in HIV/AIDS epidemiology in Poland, 1. *Euro Surveill* 11.

Rothschild BM (2005). History of syphilis. *Clin Infect Dis* 40:1454–1463.

Rubinstein A (1983). Acquired immunodeficiency syndrome in infants. *Am J Dis Child* 137:825–827.

Rubinstein A, Sicklick M, Gupta A, Bernstein L, Klein N, Rubinstein E et al. (1983). Acquired immunodeficiency with reversed T4/T8 ratios in infants born to promiscuous and drug-addicted mothers. *JAMA* 249:2350–2356.

Shelton JD, Halperin DT, Wilson D (2006). Has global HIV incidence peaked? *Lancet* 367:1120–1122.

Shilts R (1987). *And the Band Played On: Politics, People, and the AIDS Epidemic*. New York: St. Martin's Press.

Stein ZA (1990). HIV prevention: the need for methods women can use. *Am J Public Health* 80:460–462.

Stein Z, Glymour MM (2007). Diaphragms and lubricant gel for prevention of HIV. *Lancet* 370:1823–1824.

Stern AM, Markel H (2004). International efforts to control infectious diseases, 1851 to the present. *JAMA* 292:1474–1479.

Surgeon General (2000). Reducing tobacco use. A Report of the Surgeon General. Executive Summary. *MMWR Recomm Rep* 49:1–27.

Susser I (2002). Health rights for women in the age of AIDS. *Int J Epidemiol* 31:45–48.

Susser I (2009). *Aids, Sex and Culture: Global Politics and Survival in Southern Africa*. New York: Blackwell.

Susser I and Stein Z (2000). Culture, sexuality, and women's agency in the prevention of HIV/AIDS in Southern Africa. *Am J Public Health* 90:1042–1048.

Tiemessen CT, Kuhn L (2007). CC chemokines and protective immunity: insights gained from mother-to-child transmission of HIV. *Nat Immunol* 8:219–222.

Townsend CL, Cortina-Borja M, Peckham CS, de RA, Lyall H, Tookey PA (2008). Low rates of mother-to-child transmission of HIV following effective pregnancy interventions in the United Kingdom and Ireland, 2000–2006. *AIDS* 22:973–981.

US Department of Health Education and Welfare, Public Health Service (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Washington: US Government Printing Office.

United Nations and Dept. of Public Information (1979). *Everyone's United Nations*.

van de Wijgert J, Jones H, Pistorius A, de Kock A, Sebola M, Friedland B, Hoosen A, Coetzee N (2005). Phase III microbicide trial methodology: opinions of experienced expanded safety trial participants in South Africa. *SAHARA. J* 2:311–319.

Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, Laeyendecker O et al. (2005). Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis* 191:1403–1409.

Wawer M, Kigozi G, Serwadda D, Makumbi F, Nalugoda F, Watya S, et al (2008) Trial of male circumcision in HIV+ men, Rakai, Uganda: Effects in HIV+ men and in women partners. Proceedings of the 15th Conference on Retroviruses and Opportunistic Infections. Boston, February 3–6, 2008. Abstract 33LB CROI, Alexandria, VA.

Wilson DP, Kahn J, Blower SM (2006). Predicting the epidemiological impact of antiretroviral allocation strategies in KwaZulu-Natal: the effect of the urban-rural divide. *Proc Natl Acad Sci U S A* 103:14228–14233.

Winslow CEA (1943). *The Conquest of Epidemic Disease*. Princeton, NJ: Princeton University Press.

Zeger SL, See LC, Diggle PJ (1989). Statistical methods for monitoring the AIDS epidemic. *Stat Med* 8:3–21.

24

Choosing a Future for Epidemiology: I. Eras and Paradigms

ABSTRACT

To inform choices about the future of epidemiology, the present condition of epidemiology is examined, in terms of its evolution through three eras, each demarcated by its own paradigm: (1) the era of *sanitary statistics* with its paradigm, *miasma;* (2) the era of *infectious disease epidemiology* with its paradigm, the *germ theory;* and (3) the era of *chronic disease epidemiology* with its paradigm, the *black box*. The historical context in which these eras arose is briefly described. In each era, the public health was at the center of the concerns of the founders and early protagonists of the prevailing paradigm. Around this intellectual development we weave a further theme. We argue that in the present era, the public health has become less central a concern. At the same time, in epidemiology today the dominant black box paradigm is of declining utility and is likely soon to be superseded. (*Am J Public Health*. 1996;86:668–673)

Mervyn Susser, MB, Bch, FRCP(E), DPH,
and Ezra Susser, MD, DrPH

Introduction

The present era of epidemiology is coming to a close. The focus on risk factors at the individual level—the hall-mark of this era—will no longer serve. We need to be concerned equally with causal pathways at the societal level and with pathogenesis and causality at the molecular level.[1]

This paper prepares the groundwork for the argument that choices have to be made about the future of epidemiology. To look forward, we do well to look backward for guidance. Part I of this article sketches in brief outline the evolution of modern epidemiology in three successive eras. Following Kuhn,[2] we set the bounds of these eras in terms of dominant paradigms.[3] In Part II of this article, we advocate a paradigm for a fourth emergent era of "eco-epidemiology."

Susser M, Susser E. (1996) Choosing a Future for Epidemiology. I. Eras and Paradigms. *American Journal of Public Health* 86: 668–673.

The Evolution of Modern Epidemiology

The underlying idea that marked the beginnings of quantitative epidemiology in the 17th century was concern for the public health and disparities in mortality across society. John Graunt the haberdasher, in his book *Natural and Political Observations Made upon the Bills of Mortality* (1662),[4] reported on the social distribution of death in London and especially on the mortal consequences of plague. In his book *Political Arithmetick* (1667),[5] the physician William Petty, Graunt's friend and sponsor in the Royal Society, was the first to provide a method to quantify the costs of mortality.

The utilitarian approach that Graunt, Petty, and others adopted was entirely in accord with the justifications prevailing over the beginnings of modern science in the 15th and 16th centuries. Driven by the twin forces of capitalism and the Protestant ethic, science was sanctioned (in Robert Merton's words)[6] by "economic utility" and "the glorification of God." This ideology fostered discoveries with immediate technical application in astronomy, navigation, firearms, optics, and many other fields.

With the accelerating flow of discovery over the centuries, science as an institution abandoned its utilitarian roots to become an end in itself. For some time, however, that was not true for epidemiology. That field retained a central concern with the public health and its social distribution.

Thus, in the face of the miseries of 19th-century England—the advance guard of industrialization and rapid urbanization—modern epidemiology gradually took shape and then burst into activity with the Sanitary Movement.[7–9] Thereafter, one can discern at least three eras in epidemiology, each with its own dominant paradigm: (1) the era of *sanitary statistics* with its paradigm, *miasma;* (2) the era of *infectious disease epidemiology* with its paradigm, the *germ theory;* and (3) the era of *chronic disease epidemiology* with its paradigm, the *black box.* Each of these eras is described in historical context below (Table 24.1).

Sanitary Statistics and Miasma

Miasma was the prevailing theory of the Sanitarians for the greater part of the 19th century. Sanitary statistics made plain the toll of sickness and death in the city slums of England, France, Germany, Scandinavia, and the United States (forerunners of the squatter camps, favelas, or barrios in today's less developed world). For the conditions in these slums, the Sanitarian hypothesis of miasma impugned poisoning by foul emanations from the soil, water, and environing air. The environmental causes were thought to have broad and multiple manifestations in morbidity and mortality, and the sanitary statistics that were collected as evidence were largely undifferentiated; that is, they were related more to overall morbidity and mortality than to specific

Table 24.1 Three Eras in the Evolution of Modern Epidemiology

Era	Paradigm	Analytic Approach	Preventive Approach
Sanitary statistics (first half of 19th century)	Miasma: poisoning by foul emanations from soil, air, and water	Demonstrate clustering of morbidity and mortality	Drainage, sewage, sanitation
Infectious disease epidemiology (late 19th century through first half of 20th century)	Germ theory: single agents relate one to one to specific diseases	Laboratory isolation and culture from disease sites, experimental transmission, and reproduction of lesions	Interrupt transmission (vaccines, isolation of the affected through quarantine and fever hospitals, and ultimately antibiotics)
Chronic disease epidemiology (latter half of 20th century)	Black box: exposure related to outcome without necessity for intervening factors or pathogenesis	Risk ratio of exposure to outcome at individual level in populations	Control risk factors by modifying lifestyle (diet, exercise, etc.) or agent (guns, food, etc.) or environment (pollution, passive smoking, etc.)

diseases. Only in 1839 in England did William Farr begin to use specific diagnostic classifications for national mortality statistics.[10]

Closed drainage and sewage systems, supplemented by garbage collection, public baths, and housing, were the remedies that would disperse miasma, reduce mortality and morbidity (as indeed they did), and dispel the poverty of the new urban poor (as indeed they did not). A foremost proponent—and in some cases, the originator—of these innovations was Edwin Chadwick.[11] Chadwick was a reformist who argued that disease engendered by the physical environment caused poverty. Friedrich Engels, his contemporary, was a revolutionary who, in documenting the ills of Manchester factory workers, understood poverty to be the cause rather than the consequence of their ills.[12] But both agreed that the issues were societal and that the appropriate measures thus had to be applied across society.

To emphasize the underlying public health and social values of the sanitary pioneers, it is worth noting that statistics began literally as the study of the state and of the pertinent data. The newly formed London Statistical Society was chiefly concerned with assembling that data.[13] Louis René Villermé in France and William Farr in England,[10,14] founding figures of epidemiology as we know it today, are only two among many who worked to advance the public health in this fashion.

Epidemiologists, largely autodidacts, were often medical heroes in this era.[15,16] Young physicians were excited by the challenge of emergent patterns of disease that seemed rooted in a horrendous environment of urban misery. Beginning in 1858 John Simon, as chief medical officer of the national Board of Health in England, was able to draw around him over a few years a brilliant team—17 in all, no fewer than 8 of whom gained election to the Royal Society on the strength of their work. These epidemiologists mapped excess mortality across the country by district and in relation to housing, infant care, and specific diseases; studied a wide range of industries and occupations; detected many hazards from dusts, heavy metals, and general working conditions; and conducted national surveys of diet, parasite-infested meat, and food contamination.

Unfortunately, these high points of the era closely preceded its culmination. Unmodified, the miasma paradigm could not survive advances in microbiology, and its demise brought an end to the Sanitary Era. The tenacity of some of the brilliant figures of the movement, such as Edwin Chadwick and Florence Nightingale, in resisting revision of their theory rather than subsume the new biology invited ridicule from medical scientists that has hardly yet been dispelled, and the broad perspective for which they stood gradually faded. The drama of the new microbiology was not to be gainsaid.

An irony of the history of public health is that, while the sanitarians were mistaken in their causal theory of foul emanations, they nonetheless demonstrated how and where to conduct the search for causes in terms of the clustering of morbidity and mortality. The reforms they helped to achieve in drainage, sewage, water supplies, and sanitation generally brought major improvements in health. Their mistake lay in the specifics of biology rather than in the broad attribution of cause to environment.[17]

Infectious Disease Epidemiology and the Germ Theory

In 1840, Jakob Henle published a tightly argued treatise that hypothesized (as a few beginning with Fracastorius and others had done before him) that infection by minute organisms was a major cause of disease.[18–20] Despite John Snow's founding work of 1849 to 1854 in analytic epidemiology on the organismic cause of cholera[21] and his advances on Henle's formulation,[22] 25 years passed before Henle was vindicated. Louis Pasteur's demonstration of a living organism as the agent in an epidemic afflicting silkworms culminated in 1865.[23] Studies of infection and contagion in human disease—for instance, tuberculosis, anthrax, and leprosy[24–26]—followed. Finally, in 1882, Henle's one-time student Robert Koch established a mycobacterium as "the cause" of tuberculosis.[27] Henle, Snow, Pasteur, and Koch can well stand as symbolic founding figures of the new era.

Snow and Koch faced directly the most acute public health problems of the time. Although Henle had no means of intervention at hand and Pasteur worked first on the commercial problems of diseases that threatened the silk industry and viticulture, they too declared and shared a public health perspective on the prevention of disease. Despite these origins, the new paradigm of disease that followed from their work, the *germ theory,* led in the end to the narrow laboratory perspective of a specific cause model[28]—namely, single agents relating one to one to specific diseases.

The germ theory and its attendant view of specific cause dominated medical and public health sciences from the last quarter of the 19th century through at least the mid-20th century. Single agents of disease were sought by the isolation and culture of microorganisms from disease sites, the experimental transmission of these microorganisms, and the reproduction of lesions. The appropriate responses were to limit transmission by vaccines, to isolate those affected, and, ultimately, to cure with chemotherapy and antibiotics. Laboratory-based diagnosis, immunization, and treatment gained precision with every new advance. The miasma theory was relegated to the same oblivion as phlogiston.

At the same time, the epidemiology of populations and environmental exposures, and the social dynamics of disease that had flowed from the miasma theory, went into decline, replaced by a focus on control of infectious agents. Epidemiology was often a derivative pursuit rather than a creative science in its own right. The new era scarcely maintained, let alone matched, the epidemiological advances of the 19th century in the design and conduct of field surveys, the construction of national statistical systems for vital data, and the statistical analyses of large numbers.[17] The adherents of the traditional philosophy of public health lost prestige and power in the medical hierarchy and, indeed, were disparaged in ways that in many places continue in the present.

The search for other than microbiological causes of disease in the environment stumbled if it did not altogether cease. Thus, in the United States, Joseph Goldberger, in his work on pellagra from 1914 through the 1920s,[29,30] ran against the tide of belief when he established nutritional deficiency as a cause of infection. This was even more the case when, in the rural South, he and Edgar Sydenstricker showed the dietary deficiency to be consequent on the poverty of sharecroppers and other workers trapped by the economic structure of the cotton fields.

In the same period, the search for a viral cause for the growing scourge of poliomyelitis was of course ultimately justified. But the concentration of resources in the laboratory search for an organism led to the neglect of key epidemiological findings and rendered futile the preventive approaches attempted. As early as 1905, Ivar Wickman in Sweden[31] and, a decade later,

Wade Hampton Frost in the United States[32] had concluded from epidemiological data that widespread transmission of silent infection by some unknown agent was the underlying factor in the summer epidemics that were devastating the children of the better-off classes in particular.

The irony of the Sanitary Era was here reversed. While, within their limited frame of reference, the germ theorists were accurate in their causal attributions for many diseases, their narrow focus retarded the creative use of bacterial discoveries to advance the science of epidemiology. Some have argued that the decline of infectious diseases in the developed countries in the first part of this century, at the height of the germ theory paradigm, owed very little to science (including the use of vaccines and antibiotics) and much to nutrition or improved living standards.[33,34] While closer analysis does not sustain the argument against the role of science,[35] the primary role of economic development and social change is not in doubt.

Whatever the causes, the great scourges of communicable disease did come under control in the developed countries. Once the major infectious agents seemed all to have been identified and communicable disease no longer overwhelmed all other mortal disorders, the force of the germ theory paradigm faded. With notable exceptions such as René Dubos,[36] few anticipated the recrudescence of communicable disease or new global epidemics. With the emerging predominance of chronic disease of unknown cause, under any credible causal paradigm the social and physical environment had now to be reckoned with once more.

Chronic Disease Epidemiology and the Black Box

World War II serves as a convenient watershed for the beginning of the Chronic Disease Era and the black box paradigm. Shortly after the war ended in 1945 it was clear that, in the developed world, rising chronic disease mortality had overtaken mortality from infectious disease. The rise was not owed to the aging of populations alone. In middle-aged men specifically, the rises in peptic ulcer disease, coronary heart disease, and lung cancer were in each case fast and frightening enough to earn place and title as epidemics.[37]

By this time, also, chemotherapy and antibiotics had been added to the medical armamentarium. Their overwhelming therapeutic effects seemed to give tangible evidence that the major infectious diseases had been conquered. Only later was it discerned that these treatments were neither the only nor the primary factor in the steady decline of these diseases in the first half of the 20th century.[33]

The prevailing epidemiology of our day expressed the effort to understand and control the new chronic disease epidemics. Again the era was, at the

outset, driven by public health concepts. The problems selected for investigation were the chronic diseases that most visibly threatened the public health, and the groups studied were those at manifest risk—namely, middle-aged men.

Chronic disease epidemiology took firm hold with the first undeniable successes in this endeavor. British epidemiologists Richard Doll, Austin Bradford Hill, Jeremy Morris, Thomas McKeown, and others were key figures. The case-control and cohort studies on smoking and lung cancer, and the early cohort studies on coronary heart disease that established serum cholesterol and smoking as risk factors, demonstrated the power of the observational method and established its credentials.[2]

These studies carried the invisible imprimatur of the black box paradigm ("black box" being the general metaphor for a self-contained unit whose inner processes are hidden from the viewer). This paradigm related exposure to outcome without any necessary obligation to interpolate either intervening factors or even pathogenesis (although not all neglected such interpolation). Epidemiologists were faced once more, as in the Sanitary Era, with major mortal diseases of completely unknown origin. At the outset, of necessity they resorted to straightforward descriptive studies of disease distribution and exploratory sweeps for possible factors that enhanced risk.[37] As they moved on to test the emergent observations, these epidemiologists relied upon ingenuity in design and the seizing of opportune circumstance to reach their conclusions. They seldom resorted to complex statistical analysis.

The studies of lung cancer were particularly influential in giving the new paradigm credibility. Pathogenesis was by-passed. Thus, the best biological support to be found for the smoking–lung cancer relationship was quite indirect, residing in the demonstration by the Kennaways and their colleagues that tars applied to the skin of mice were carcinogenic.[38] Indeed, for another 4 decades, no direct analogy with the epidemiological studies of smoking existed in animal experiments.

Step by step, the complexities of chronic disease epidemiology emerged, first in matters of design and causal inference and, in parallel a little later, in matters of statistical analysis.[2] The incipient thinking on design of previous decades was developed and systematized.[39,40] The structure of designs was clarified, the necessity for statistical power and the advantages of large numbers understood.[41]

Epidemiologists were obliged to depart from the specific-cause model of the germ theory. The metaphor of a "web of causation"[40] characterized the multi-causal nature of public health problems, particularly those of chronic disease. After this beginning, one of us tried for his own part to give systematic form to the problems of inference that arose in the nascent epidemiology of a multivariate world.[17]

Later, analytic issues and statistical refinement became a driving force. The sharpening of technique led to a cycle of continual refinement. Epidemiologists began to explore in depth the subtleties of confounding, misclassification, survivorship, and other such issues. This labor is represented in the elegant and unifying concept of the fourfold table and the case–control and cohort designs as alternative methods of sampling the population disease experience to estimate risk ratios or odds ratios.[17,41–44]

The black box paradigm remains the prevailing model, and virtually all contemporary epidemiologists including ourselves have made use of it. It can still yield findings of public health significance. Neural tube defects provide a recent example. Typical black box approaches eventually led to the major discovery of the role of folate deficiency in neural tube defects. Early studies found variations with social class, geography and ethnicity, and economic cycles.[45] Further studies found exposure to famine early in pregnancy to be associated with an increased risk of congenital neural defects and prenatal vitamin intake to be associated with a decreased risk.[46,47] Finally, going beyond the black box, animal studies followed by clinical trials of supplementation established that periconceptional folic acid can prevent a large proportion of neural tube defects.[48,49]

Momentum for a New Era

The climax and, in all likelihood, the culmination of the black box as dominant paradigm is already upon us. Two forces, characteristic of our time and much written about, are blunting the black box paradigm: (1) a transformation in global health patterns and (2) new technology.

Health Patterns

With regard to health patterns, none has had more impact than the human immunodeficiency virus (HIV) epidemic. Although epidemiology has made some notable contributions to understanding the epidemic, black box epidemiology is ill equipped to address epidemic control. The causative organism as well as the critical risk factors are known, so prevention is theoretically possible. Yet the HIV epidemic has demonstrated that both developing and developed countries remain vulnerable to devastation by infectious disease.

Analysis of mass data at the individual level of organization alone, as implied by the black box paradigm, does not allow us to weigh at which points in the hierarchy of levels intervention is likely to be successful.[50,51] No vaccine now in prospect seems likely to achieve the efficacy level that could

also achieve epidemic control. Absent such efficacy, the failure to control the disease resides in our lack of understanding of transmission and illness in the social context. We know which social behaviors need to change, but we know little about how to change them, even when entire societies are at stake.

In retrospect, our confidence during the Chronic Disease Era about the control of infectious diseases seems naive and also blind to the less developed world. For the majority of the world's population, chronic infections—tuberculosis, syphilis, malaria, and many others—were never under control. As with HIV infection, the immediate causes and the risk factors were known, but this knowledge could not be translated into protection of the public health.

Similarly, our confidence in our ability to control chronic noncommunicable diseases themselves by modifying behavior that carries risk has been shaken. Again, knowledge of risk factors and interventions directed solely at changing the behavior of individuals, even across several communities, have proven insufficient.[52,53]

Health problems driven by societal problems point to the location of the underlying difficulties. The black box paradigm alone does not elucidate societal forces or their relation to health. The focus on populations is generally directed at the individuals within them. Prevention at the societal level, conceptualized as intervening with individuals en masse, is often nullified when the target is a social entity with its own laws and dynamics.

Technology

With regard to technology, the developments that will drive research and that can lead epidemiology to a new paradigm reside primarily in biology and biomedical techniques on the one hand, and in information systems on the other. These advances have begun to reshape all health disciplines.

Biological techniques such as genetic recombination and imaging have transformed the ability of epidemiologists to comprehend human disease at the microlevel. For example, the methods of recombinant DNA have led to recognition of both viral and genetic components in insulin-dependent diabetes[54]; to the definitive tracking from person to person of HIV, tuberculosis, and other infections through the molecular specificity of the organisms[55]; to the discovery of a herpes virus as almost certainly the agent in Kaposi's sarcoma[56]; and to the drama of the familial tracking and marking of the first breast cancer gene.[57] Imaging has undermined the notion of schizophrenia as functional psychosis and given backing to the existence of environmental factors.[58] It has also allowed us to discover a frequency of brain lesions in the premature newborn that was unsuspectedly high overall and concentrated in the earliest hours of life.[59] Learning from the new technology has

only begun. Once unimaginable possibilities follow from the mapping of the human genome for specifying the role of heredity in disease, and no less from the visualization of physiological processes for interpreting human function.

The potential contribution of these advances to epidemiology is an exquisite refinement of the definition and measurement of susceptibility, exposure, and outcome. Such refinement also clarifies the intervening pathways and so elucidates with precision causal processes and not merely causal factors. We can be confident that new techniques, properly applied, can help dig epidemiology out of the slough of marginally significant risk estimates.[60,61]

In parallel, technology at the societal level in the form of the global communication network has opened new possibilities for understanding and controlling disease. Information networks can provide instant access to—and enable the continuous assemblage of—existing stores of vital statistics and other relevant health and social data[62] across the world. Such data have myriad uses for newly empowering public health. They promise a capacity for devising and testing well-directed interventions at a societal level. Stores of data can be mined to describe distributions across societies, to make comparisons of strata and groups nationally and internationally, to generate and test hypotheses, and to serve as sampling frames. Continuous accumulation of data over time can serve for overall surveillance of health states, the detection of nascent epidemics and new diseases, the response to disasters, and the evaluation of interventions. This technology thus brings comprehension of large-scale phenomena and even systems within our grasp; it places at our command the ability and the necessity to recognize broad dynamic patterns and, not least, disease in its social context.

Conclusion

When research under the current black box paradigm in its pure form relies on risk ratios that relate exposure to outcome with no elaboration of intervening pathways, it forfeits the depth offered by our new biological knowledge. In addition, because of an implicit and sometimes explicit commitment to analyzing disease solely at the individual level, research under this paradigm also dispenses with the potential breadth offered by new information systems in placing exposure, outcome, and risk in societal context.

The apogee of the black box paradigm is heralded by epidemiology texts of the 1980s.[63,64] These mark two trends. They move away from the public health orientation of the pioneers of the Chronic Disease Era. At the same time, analysis edges out design as the central focus. At the extreme we find an epidemiology untrammeled by the call to address disease in social groups, communities, and other formations of the social structure. Thus,

a widely used modern text endorses a pithy definition of epidemiology as the study of disease occurrence,[64(p17)] implicitly setting aside public health ends. Epidemiology in this view is akin to the physical sciences in sharing a search for the highest levels of abstraction in universal laws.[65] Research in this universalist vein cannot take advantage of the extraordinary shifts and opportunities opened by new dynamics of disease and new technology.

In the evolution of modern epidemiology, dominant paradigms have been displaced by new ones as health patterns and technologies have shifted. As happened with previous paradigms, the black box, strained beyond its limits, is soon likely to be subsumed if not superseded entirely by another paradigm. This paradigm reflects a particular era in our development as a discipline. In our view, we stand at the verge of a new era.

Acknowledgments

A first version of this paper and its companion served as the keynote (given by M. S.) for the Pan-American Epidemiology Congress, Salvador, Brazil; April 28, 1995. Thanks are due to Mauricio Barreto and his organizing committee for posing the challenge. The current version was the basis for the Thomas Francis Memorial Lecture (given by M. S.) at the University of Michigan, Ann Arbor, February 23, 1996.

For their careful reading and comments on earlier versions, we also owe thanks to Elizabeth Fee, Nancy Krieger, Bruce Link, John McKinlay, Gerald Oppenheimer, Nigel Paneth, Charles Poole, and Zena Stein.

References

1. Susser M. Epidemiology today: "a thought-tormented world." *Int J Epidemiol.* 1989;18: 481–488.
2. Kuhn TS. *The Structure of Scientific Revolutions.* 2nd ed. Chicago, Ill: University of Chicago Press; 1970.
3. Susser M. Epidemiology in the United States after World War II: the evolution of technique. *Epidemiol Rev.* 1985;7:147–177. Reprinted in: Susser M, ed. *Epidemiology, Health, and Society: Selected Papers.* New York, NY: Oxford University Press; 1987:22–49.
4. Graunt J. *Natural and Political Observations Made upon the Bills of Mortality.* London, England: T. Roycraft; 1662. Reprinted: Baltimore, Md: The John Hopkins Press; 1939.
5. Petty W; Hull CH, ed. *The Economic Writings of Sir William Petty, together with the Observations upon the Bills of Mortality More Probably by Captain John Graunt.* 2 vols. Cambridge, England: Cambridge University Press; 1899.
6. Merton RK. *The Sociology of Science: Theoretical and Empirical Investigations.* Chicago, Ill: University of Chicago Press; 1973.

7. Simon J. *English Sanitary Institutions.* 2nd ed. London, England: John Murray; 1887.
8. Rosen G. *A History of Public Health.* Reprinted with an introduction by Fee E. Baltimore, Md: The John Hopkins University Press; 1993.
9. Fee E, Acheson RM, eds. *A History of Education in Public Health.* New York, NY: Oxford University Press; 1991.
10. Farr W; Humphreys NA, ed. *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr, 1885.* Reprinted with an introduction by Susser M and Adelstein A. Metuchen, NJ: Scarecrow Press for the New York Academy of Medicine; 1975.
11. Chadwick E. *1842 Report on the Sanitary Condition of the Labouring Population of Great Britain.* Reprinted: Edinburgh, Scotland: Edinburgh University Press; 1965.
12. Susser M. Ethical components in the definition of health. *Int J Health Serv.* 1974;4:539–548. Reprinted in: Susser M, ed. *Epidemiology, Health, and Society: Selected Papers.* New York, NY: Oxford University Press; 1987:186–193.
13. Cullen MF. *The Statistical Movement in Early Victorian Britain.* Sussex, England: Harvester Press; 1975.
14. Villermé LR. Rapport fait par M. Villermé, et lu a l'Academie de medecine, au nom de la Commission de statistique, sur une serie de tableau relatifs au mouvement de la population dans les douze arrondissements municapaux de la ville de Paris pendant les cinq annees 1817, 1818, 1819, 1890, et 1821. *Archives generales de medecine.* 10:216–245.
15. Brockington CF. *Public Health in the Nineteenth Century.* Edinburgh, Scotland: E & S Livingstone; 1965.
16. Lambert R. *Sir John Simon (1816–1904) and English Social Administration.* London, England: MacGibbon and Kee; 1963.
17. Susser M. *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* New York, NY: Oxford University Press; 1973.
18. Rosen G. Social aspects of Jacob Henle's medical thought. *Bull Inst Hist Med.* 1937;5: 509–537.
19. Henle J; George Rosen, trans. *On Miasmata and Contagia.* Baltimore, Md: The Johns Hopkins Press; 1938.
20. Shryock RH. Germ theories in medicine prior to 1870: further comments on continuity in science. *Clio Med.* 1972;7:81–109.
21. Snow J. *On the Mode of Communication of Cholera.* 2nd ed, much enlarged. London, England: J. Churchill; 1855. Reprinted as: *Snow on Cholera.* New York, NY: Commonwealth Fund; 1936.
22. Winkelstein W Jr. A new prospective on John Snow's communicable disease theory. *Am J Epidemiol.* 1995;142(suppl):S3–9.
23. Vallery-Radot R; Devonshire RL, trans. *The Life of Pasteur.* London, England: Constable; 1901.
24. Villemin JA. Cause et nature de la tuberculose: son inoculation del l'homme au lapin. *Compt Rend Acad Sci.* 1865;61:1012–1015. Cited by Dubos R, Dubos J. *The White Plague, Tuberculosis, Man and Society.* Boston, Mass: Little Brown; 1952:98–99.
25. Hansen GA, Looft C, eds. *Leprosy: In Its Clinical and Pathological Aspects.* Bristol, England: John Wright & Co; 1895.
26. Irgens LM, Bjerkedal T. Epidemiology of leprosy in Norway: the history of the National Leprosy Registry of Norway from 1856 until today. *Int J Epidemiol.* 1973;2:81–89.

27. Koch R. Die aetiologie der Tuberkulose. 1882. Reprinted in: Schwalbe J, ed. *Gesammelte Werke von Robert Koch*. Leipzig, Germany: Georg Thieme, Verlag; 1912;1:428–455.
28. Evans AS. Causation and disease: the Heule Koch Postulatis revisited. *Yale J Biol Med*. 1976;49:175–195.
29. Goldberger J, Wheeler GA, Sydenstricker E. A study of the diet of nonpellagrous and pellagrous households in textile mill communities in South Carolina in 1916. *JAMA*. 1918;71:944–949.
30. Terris M, ed. Goldberger on Pellagra. Baton Rouge, La: Louisiana State University Press; 1964.
31. Wickman I. Weitere Studine uber Poliomy-elimitis acute; ein Beitrag zur Kenntnis der Neurophagen und Kornchenzellen. *Dtsche Z. Nervenheik*. 1909/10;38:396–437. Beitrage zur kenntnis der Heine-Medinschen Krankheit (Poliomyelitits acuta und verwandter Erkrankungen). Berlin, Germany: Karger; 1907. Both cited in: Paul JR. *History of Poliomyelitis*. New Haven, Conn: Yale University Press; 1971; chap 10.
32. Frost WH. Poliomyelitis (infantile paralysis): what is known of its cause and modes of transmission. *Public Health Rep*. 1916;51:1817–1833.
33. McKeown T. *The Modern Rise of Population*. London, England: Edward Arnold; 1976.
34. McKinlay JB. From "promising report" to "standard procedure": seven stages in the career of a medical innovation. *Milbank Mem Fund Q*. 1981;59(3):374–411.
35. Susser M. Health as a human right: an epidemiologist's perspective on the public health. *Am J Public Health*. 1993;83:418–426.
36. Dubos R. *The Mirage of Health: Utopias: Progress and Biological Change*. New York, NY: Harper & Row; 1959.
37. Morris JN. *Uses of Epidemiology*. London, England: Churchill Livingstone; 1957.
38. Burrows H, Heiger I, Kennaway EL. The experimental production of tumors of connective tissue. *Am J Cancer*. 1932;16:57–67.
39. Witts LJ, ed. *Medical Surveys and Clinical Trials*. London, England: Oxford University Press; 1959:64–90.
40. MacMahon B, Pugh TF, Ipsen J. *Epidemiological Methods*. Boston, Mass: Little, Brown & Co, 1960.
41. Fleiss JL. *Statistical Methods for Rates and Proportions*. 1st ed. New York, NY: John Wiley & Sons; 1974.
42. Miettinen O. Estimability and estimation in case-referent studies. *Am J Epidemiol*. 1976;103:226–235.
43. Breslow NE, Day NE, eds. *Statistical Methods in Cancer Research*. Vol 1. *The Analysis of Case–Control Studies*. Lyon, France: International Agency for Research on Cancer; 1980.
44. Greenland S, Thomas DC, Morgenstern H. The rare-disease assumption revisited. A critique of "estimators of relative risk for case-control studies." *Am J Epidemiol*. 1986;124:869–876.
45. Elwood JM, Little J, Elwood JH. *Epidemiology and Control of Neural Tube Defects*. New York, NY: Oxford University Press; 1992.
46. Stein Z, Susser M, Saenger G, Marolla F, eds. *Famine and Human Development. The Dutch Hunger Winter of 1944–45*. New York, NY: Oxford University Press; 1975.
47. Milunsky A, Jick H, Jick SS, et al. Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects. *JAMA*. 1989;262:2847–2852.

48. Smithells RW, Nevin NC, Seller MJ, et al. Further experience of vitamin supplementation for prevention of neural tube defect recurrences. *Lancet.* 1983;i:1027.
49. Medical Research Council Vitamin Study Research Group. Prevention of neural tube defects: results of the medical research council vitamin study. *Lancet.* 1991;338:131–137.
50. Koopman JS, Longini IM, Jacquez JA, et al. Assessing risk factors for transmission of infection. *Am J Epidemiol.* 1991;133:1199–1209.
51. Blower SM, McLean AR. Prophylactic vaccines, risk behavior change, and the probability of eradicating HIV in San Francisco. *Science.* 1994;265:1451–1454.
52. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): II. Changes in adult cigarette smoking prevalence. *Am J Public Health.* 1995;85:193–200.
53. Susser M. Editorial: the tribulations of trials—interventions in communities. *Am J Public Health.* 1995;85:156–158.
54. Solimena M, De Camilli P. Coxsackie viruses and diabetes. *Nature Med.* 1995;1(1): 25–26.
55. Alland D, Kalkut GE, Moss AR, et al. Transmission of tuberculosis in New York City: an analysis by DNA fingerprinting and conventional epidemiologic methods. *N Engl J Med.* 1994;330:1710–1716.
56. Chang Y, Cesarman E, Pessin MS, et al. Identification of herpesvirus-like DNA sequence in AIDS-associated Kaposi's sarcoma. *Science.* 1994;266:1865–1869.
57. Hall JM, Lee MK, Newman B, et al. Linkage of early onset familial breast cancer to chromosome 17Q21. *Science.* 1990;250:1684–1689.
58. Andreasen NC, Arndt S, Swayze II, et al. Thalamic abnormalities in schizophrenia visualized through magnetic resonance image averaging. *Science.* 1994;266:294–298.
59. Paneth N, Pinto-Martin J, Gardiner J, et al. Incidence and timing of germinal matrix/intraventricular hemorrhage in low birth weight infants. *Am J Epidemiol.* 1993;137:1167–1175.
60. Stein Z, Hatch M. Biological markers in reproductive epidemiology: prospects and precautions. *Environ Health Perspect.* 1987; 74:67–75.
61. Rothman N, Stewart WF, Schulte PA. Incorporating biomarkers into cancer epidemiology: a matrix of biomarker and study design categories. *Cancer Epidemiol Biomarkers Prev.* 1995;4:301–311.
62. Friede A, Reid JA, Ory HW. CDC WONDER: a comprehensive on-line public health information system of the Centers for Disease Control and Prevention. *Am J Public Health.* 1993;83:1289–1294.
63. Kleinbaum DG, Kupper LL, Morgenstern H, eds. *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, Calif: Lifetime Learning Publications, Wadsworth Inc; 1982.
64. Rothman KJ. *Modern Epidemiology.* Boston, Mass: Little, Brown & Co; 1986.
65. Rothman KJ, ed. *Causal Inference.* Chestnut Hill, Mass: Epidemiology Resources Inc; 1988.

25

Choosing a Future for Epidemiology: II. From Black Box to Chinese Boxes and Eco-Epidemiology

ABSTRACT

Part I of this paper traced the evolution of modern epidemiology in terms of three eras, each with its dominant paradigm, culminating in the present era of chronic disease epidemiology with its paradigm, the black box. This paper sees the close of the present era and foresees a new era of eco-epidemiology in which the deployment of a different paradigm will be crucial. Here a paradigm is advocated for the emergent era. Encompassing many levels of organization—molecular and societal as well as individual—this paradigm, termed *Chinese boxes,* aims to integrate more than a single level in design, analysis, and interpretation. Such a paradigm could sustain and refine a public health–oriented epidemiology. But preventing a decline of creative epidemiology in this new era will require more than a cogent scientific paradigm. Attention will have to be paid to the social processes that foster a cohesive and humane discipline. (*Am J Public Health*. 1996;86:674–677)

Mervyn Susser, MB, Bch, FRCP(E), DPH, and
Ezra Susser, MD, DrPH

> *Whether the provocative factors are inherent in the population through their genes, their cells and their parasites; whether they are failures to meet environmental influences of social, physical or biological nature they constitute the concern of the epidemiologist.... Consequently, epidemiology must constantly seek imaginative and ingenious teachers and scholars to create a new genre of medical ecologists who... can interpret the interplay of forces which result in disease.*
>
> Thomas Francis[19]

Susser M, Susser E. (1996) Choosing a Future for Epidemiology. II. From black box to Chinese boxes and eco-epidemiology. *American Journal of Public Health* 86: 674–677.

Introduction

In this paper on the choices before epidemiology, we advocate a paradigm for an emergent era of eco-epidemiology. To connote the inclusion of systems at different levels, we term the paradigm *Chinese boxes.* This paradigm stems from a particular distinction between the "universalism" of the physical sciences and the "ecologism" of the biological sciences. It places epidemiology on the track of ecologism, a perspective we aim to explain and justify below.

The practical implication of a localizing ecological paradigm for the design of epidemiological research is that an exclusive focus on risk factors at the individual level within populations—even given the largest numbers—will not serve. We need to be equally concerned with causal pathways at the societal level and with pathogenesis and causality at the molecular level. Here we note that investigations at all these levels are found in the history of medicine and epidemiology since early times. Hippocrates was concerned with the effects of broad environmental conditions on health.[1] Later Galen, who emphasized the individual host in the form of the four humors, did not neglect the interaction of susceptibility with lifestyle. Paracelsus, in the 16th century, aimed to grasp multiple levels.[2] He tried to apply chemistry to medicine, and he also studied the influence of the stars on physiology.

The Need for a New Paradigm

The necessity and the potential of a new paradigm can be illustrated for the infectious disease of human immunodeficiency virus (HIV) and the chronic disease of peptic ulcer. While these two diseases were selected to represent infectious and chronic diseases of our time, each of them also shows a blurring of the distinction between infectious and chronic disease. This is itself a hallmark of the new era.

To understand and contain the global epidemic of HIV requires causal thinking at several levels of analysis. At the molecular level, the precision of molecular biology is required to determine the means and the timing of transmission and to find a way to interrupt it. At an intermediate level, specific social behavior of individuals fosters sexual and other forms of transmission of the virus. At the population level, the dynamics of the epidemic are governed by the prevalence of the infection itself as well as by other characteristics of the population—for instance, patterns of sexual relationships and of breast-feeding, prevalence of other sexually transmitted diseases, and nutritional factors such as maternal vitamin A levels. At the global level, the interconnections between societies determine the path of the infection. As investigators, we are naturally constrained by our capabilities

and by the necessary reductionism entailed in firmly establishing the connection between one thing and another and, more especially, their causal relationships. Yet the best hopes for containing the epidemic rest upon a coherent strategy that can address all these levels.

Peptic ulcer similarly illustrates the limitations of a narrow frame of reference for a chronic disease.[4] The causal framework of the gastrophysiologist is likely to focus on the wall of the stomach; that of the neurophysiologist, on the autonomic nervous system. The psychosomaticist expands the framework to include internal and environmental stressors, the human geneticist considers familiality in blood groups and secretor status, and the microbiologist brings the recent discoveries about *Helicobacter pylori* to bear. The epidemiologist includes all the above and adds smoking as an individual risk factor.

But the mystery and the challenge of peptic ulcer for epidemiology lies at the ecological level of major secular change. We still have to unravel the factors that caused the peptic ulcer complex first to wax and then to wane. This condition (or complex of conditions) mysteriously reached a peak in the 1950s and then, no less mysteriously, began to decline. This was a cohort phenomenon that began its rise in cohorts born before the turn of the 19th century, with a steady decline in cohorts born thereafter.[3] A fully adequate causal model for public health must explain the disease at the ecological level as well as at lesser and more refined levels of organization. This remains so even if the best explanation turns out to be the historical behavior of *Helicobacter* micro-organisms.

Universalism vs Ecologism

The road is now open for epidemiologists to work at the same time at the molecular and the societal levels. To do so, we need to be guided by appropriate causal concepts, a matter already under discussion in epidemiology.[4–10]

Like all the sciences, epidemiology seeks generalizing concepts to explain the causes of things. In the history of science, however, one can trace not one but two conceptual tracks. The well-described universalism of the physical sciences must be complemented by the often unacknowledged ecologism of the biological sciences. In contrast with universalism, ecologism entails localization and attention to the bounds that limit generalizations about biological, human, and social systems.

A concept of causality based on universal laws is pervasive in the sciences. Most philosophers of science have confined the enterprise almost entirely to the universalist framework, although, of course, exceptions exist.[11] We believe that epidemiologists among others have been misled by standard interpretations of the nature of science.

The search for universal laws of the material world must deal with a paradox. The smaller the interacting microcosmic elements that such laws explain, the more likely those elements are to be universal. Universality implies a view of space and time expanding outward across the boundaries and horizons of our world and others, unimpeded by the local accretions and characteristics of intervening structures such as planets, continents, or our biological world, including people.

Some laws may hold across our planet for species and the evolutionary processes that produced them. But above the level of molecules, no biological entity can conform entirely to universal laws because of the overarching contexts and the interactions between levels within a biological structure. And the banal fact is that each society is influenced by its economic, political, and cultural circumstances as well as by its mix of peoples, climate, and topography. What is most universal is least biological and, most of all, least human.

It follows that universalism is not universally applicable to the scientific endeavor. Thus, when we enter the physical, biological, and social realms of the human world, we need a parallel set of ideas interwoven with the search for generality. In epidemiology, the poor fit of universalism with human reality is better replaced by a contrasting construct of ecologism. Ecological constructs try to deal with the true complexity of the biological world. Such constructs must in varying degree be localized; they must be bounded if they are to encompass all of the biological world's less-than-universal levels and their particular interactions.

Chinese Boxes: A Paradigm for Eco-Epidemiology

In proposing a paradigm in the vein of ecologism, we draw on and develop an earlier formulation of agent and host esconced in an environment that comprises systems at multiple levels.[4] Our concept envisages interactive systems. A system is a set or assembly of factors connected with each other in some form of coherent relationship. Thus, a system is an abstraction that allows a set of related factors to be described in terms of a coherent structure or coherent function. We speak properly of anatomical (structural) systems and physiological (functional) systems—circulatory, nervous, or reproductive. The human body is in itself a system that encompasses all these. Societies comprise much more complex systems of persisting and ordered relations. The universe is a system of vast scale, a molecule one of minuscule scale.

Each system can be described in its own terms. Each defines the limits of a particular level of organization and the structure within those limits. Hence, a set of factors that make up a system can be identified. Their

coherence implies a degree of persistence and stability. This stability coexists, however, with the capacity for change. Because the factors contained in a system relate in some fashion, change and activity in one sector impinges on and affects other sectors.

Systems also relate to one another; they do not exist in isolation. A metaphor may serve to illuminate this ecological perspective. We liken it to Chinese boxes—a conjurer's nest of boxes, each containing a succession of smaller ones. Thus, within localized structures, we envisage successive levels of organization, each of which encompasses the next and simpler level, all with intimate links between them.

Within each level, a relatively bounded structure such as a nation or society or community may be characterized by lawful relations that are localized to that structure and can be discovered. At any given level within the hierarchy of scale and complexity, these lawful relations are generalizable, but only to the extent that they hold for other similar structures, whether they are societies, cities, local communities, or individuals.

The paradigm represented by the metaphor of Chinese boxes could be suited to a new eco-epidemiology (Table 25.1). This paradigm treats relations within and between localized structures that are bounded socially, biologically, or topographically. The appropriate epidemiological approach is to analyze determinants and outcomes at different levels of organization. Such contextual analysis would draw on new information systems both within and across levels to achieve breadth. It would draw on new biomedical techniques to achieve depth. The action that follows would need to find leverage at the most efficacious levels, whether contextual or molecular or both together.

The metaphor of Chinese boxes is perhaps not apt in every dimension, in that levels exist in a hierarchy not only of scale but also of complexity, with multiple interactions between and within levels.[14,15] The outer box might be the overarching physical environment which, in turn, contains societies and populations (the epidemiological terrain), single individuals, and individual physiological systems, tissues and cells, and finally (in biology) molecules.

To study even ecological systems in depth, we still have to use the basic methodological procedures of science and limit the fields of observation. Epidemiology can never aspire to the reductionism that Freeman Dyson defines as the "effort to reduce the world of physical phenomena to a finite set of fundamental equations."[12] Steven Weinberg calls this "grand reductionism," which entails a view of nature. However, epidemiologists must of necessity live with and use what Weinberg calls "petty reductionism," which entails only a research procedure or program.[13] But these approaches should not be allowed to obscure the contextual structure of enveloping systems.[14,15] To deal with such a hierarchy of enveloping systems, the need for a new paradigm is patent.

Table 25.1 Eras in the Evolution of Modern Epidemiology and an Emergent Era

Era	Paradigm	Analytic Approach	Preventive Approach
Sanitary statistics (first half of 19th century)	Miasma: poisoning by foul emanations from soil, air, and water	Demonstrate clustering of morbidity and mortality	Introduce drainage, sewage, sanitation
Infectious disease (late 19th century through first half of 20th century)	Germ theory: single agents relate one to one to specific diseases	Laboratory isolation and culture from disease sites, experimental transmission and reproduction of lesions	Interrupt transmission (vaccines, isolation of the affected through quarantine and fever hospitals, and ultimately antibiotics)
Chronic disease epidemiology (latter half of 20th century)	Black box: exposure related to outcome, without necessity for intervening factors or pathogenesis	Risk ratio of exposure to outcome at individual level in populations	Control risk factors by modifying lifestyle (diet, exercise, etc), agent (guns, food, etc), or environment (pollution, passive smoking, etc)
Eco-epidemiology (emerging)	Chinese boxes: relations within and between localized structures organized in a hierarchy of levels	Analysis of determinants and outcomes at different levels of organization: within and across contexts (using new information systems) and in depth (using new biomedical techniques)	Apply both information and biomedical technology to find leverage at efficacious levels, from contextual to molecular

Choosing the Future

Although we hear stirrings, we have yet to adopt, develop, and apply this type of paradigm in epidemiology. What we present here is no more than a skeletal framework. As this embryonic paradigm is tested in the field, no doubt its simplifications and inadequacies will emerge, and some of its deficiencies will be repaired.

The paradigm is bound to evolve and change as the constraints of existing thought are broken, and one can expect it to confer new power on epidemiology. Such a paradigm will require a slew of sophisticated methods—borrowed, adapted, and created anew—that enable epidemiologists to test models at levels from the molecular to the social.

At this time, the task will seem daunting, even hopeless, to many of us. Few epidemiologists are equipped to undertake it. At the beginning of this century, however, Ronald Ross pioneered an analogous approach.[16] In 1902

he won the Nobel prize for establishing (in the 1890s by painstaking microscopy) that mosquitoes transmit malaria. He thereafter took an epidemiological approach to eradicating the disease. Epidemiology and a mathematical bent led him to multivariate modeling to determine what the efficacy might be of interventions of different sorts.

We draw further on historical precedent to justify optimism. A study of the literature of the early Chronic Disease era[17] confirms direct experience of the elementary design and analytic tools in use at the opening of the era. Design principles were only just taking form, and multivariate analysis was almost inaccessible. The contrast with the powerful designs and sophisticated analyses of the latter years of the era could scarcely be greater. Many such precedents give us reason to believe that the requisite analytic tools are within reach, provided that the attention of epidemiologists is focused on their development and use.

Here one must recognize that a molecular paradigm taken on its own is hugely attractive because of its explanatory power. Without conscious countervailing effort, that paradigm will very likely come to dominate epidemiology no less than did the germ theory in its time. In that event, with the sacrifice of conceptual and analytic breadth, epidemiology could again be reduced to a derivative pursuit of laboratory science, and the mainstream of our subject could be lost to creative science. A countervailing force, which at the same time restores public health to epidemiology, resides in a developed version of the Chinese boxes paradigm.

One must also take heed of another emergent paradigm. Information systems combined with systems analyses might well lead into a systems paradigm, with its own attractions for mathematically minded epidemiologists. Standing alone, this paradigm would sacrifice biological depth and the direct address to health disorders. To avoid constriction, both the emergent themes of biology and information as well as the black box of our era need to be subsumed into a broader paradigm such as the Chinese boxes proposed here for eco-epidemiology.

A cogent scientific paradigm alone is not enough to anchor epidemiologists to public health, however. So what more is needed to accomplish the linkage, one may ask, beyond simple evangelism for an epidemiology inviolably tied to the public health?

Socialization

At the least, a practical program must be devised to ensure that, in the course of their education, epidemiologists are socialized in a manner that keeps alive the idea of improving the public health as a primary value. Epidemiologists

must be scientific but also in some degree professional in the sense traditional to medicine, the law, and the clergy. That is, society accords them a privileged and autonomous function founded on special training. That autonomy carries reciprocal and primary ethical obligations for service to individuals or society.

To maintain such an ethic, we shall have to choose and act accordingly. The power of the socialization process to imbue values is well documented in the work in medical education pioneered by Robert Merton and his colleagues[18] and in much that followed.

In this respect, epidemiology and public health face ambiguities of role and status. As emphasized above, the public health function has been to serve populations and, informed by notions of social equity, to prevent and control disease in those populations. Yet the historic origins of epidemiology are predominantly if not exclusively in medicine. And for millennia, the medical function, enshrined in ethics and teaching, has been primarily to serve sick individuals.

In this century, epidemiology and public health have often withered in a medical environment that almost inevitably must give primacy to the individual care of sick persons who solicit care. It follows that autonomous schools of public health among others can have a crucial role in socialization.

The diversification of public health professions has resulted in further role ambiguities. In addition to the doctors and sanitarians who were its original mainstay, the public health corps now comprises statisticians, economists, social scientists, professional administrators, organization and other specialists and epidemiologists without medical training. This diversification has centrifugal force. To imbue these diverse groups with the values of public health, schools of public health will have to give due weight to the process of socializing their students to common values.

Socialization of students to public health will require conscious induction through learning about its traditions and its history. They will need exposure to faculty and others who understand and embody public health values. They will need learning experiences in community situations as vivid and telling as those provided for medical students by clinicians at the bedside. They will need to comprehend the hurt and waste of deprived or disordered communities. They will need to recognize the true scale of the effects that a few percentage points in a cogent indicator can have on a nation's health.

Without intense socialization and learning, we may well find—because of the natural momentum and narrow focus that specialization generates—that the links between the values of public health and its specialized disciplines dissolve as we watch. In this respect, epidemiology is most certainly at risk.

Acknowledgments

A first version of this paper and its companion served as the keynote (given by M. S.) for the Pan-American Epidemiology Congress, Salvador, Brazil, April 28, 1995. Thanks are due to Mauricio Barreto and his organizing committee for posing the challenge. The current version was the basis for the Thomas Francis Memorial Lecture (given by M. S.) at the University of Michigan, Ann Arbor, February 23, 1996.

For their careful reading and comments on earlier versions, we also owe thanks to Rodney Ehrlich, Elizabeth Fee, Nancy Krieger, Bruce Link, John McKinlay, Gerald Oppenheimer, Nigel Paneth, Charles Poole, and Zena Stein.

References

1. Hippocrates; Chadwick J, Mann WN, trans. *Epidemics, Books I & III: The Medical Works of Hippocrates.* Oxford, England: Blackwell Scientific Publications; 1950:29–80.
2. King LS. *The Growth of Medical Thought.* Chicago, Ill: University of Chicago Press; 1963.
3. Susser M, Stein Z. Civilization and peptic ulcer. *Lancet.* 1962;1:115–119.
4. Susser M. *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology.* New York, NY: Oxford University Press; 1973.
5. Smith A. The epidemiological basis of community medicine. In: Smith A, ed. *Recent Advances in Community Medicine.* Vol. 3. Edinburgh, Scotland: Churchill-Livingston; 1985:1–10.
6. Krieger N. Epidemiology and the web of causation: has anyone seen the spider? *Soc Sci Med.* 1994;39(7):887–903.
7. Wing S. Limits of epidemiology. *Physicians Soc Respons Q.* 1994; no. 1:74–86.
8. Evans RG, Barer ML, Marmor TR, eds. *Why Are Some People Healthy and Others Not? The Determinants of Health of Populations.* New York, NY; Aldine de Gruyter; 1994.
9. Pearce N. Traditional epidemiology, modern epidemiology, and public health. *Am J Public Health.* 1996;86:678–683.
10. Link BG, Phelan JC. Editorial: understanding sociodemographic differences in health—the role of fundamental social causes. *Am J Public Health.* 1996;86:471–473.
11. Mayr E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance.* Cambridge, Mass: Belknap Press of Harvard University Press; 1982.
12. Dyson F. The scientist as rebel. *NY Rev Books.* May 25, 1995;42:31–33.
13. Weinberg S. Reductionism redux. *NY Rev Books.* October 5, 1995;42:39–42.
14. Susser M. The logic in ecological: I. the logic of analysis. *Am J Public Health.* 1994;84:825–829.
15. Susser M. The logic in ecological: II. the logic of design. *Am J Public Health.* 1994;84:830–835.

16. Ross Sir R. *The Prevention of Malaria*. 2nd ed. New York, NY: EP Dutton; 1910.
17. Susser M. Epidemiology in the United States after World War II: the evolution of technique. *Epidemiol Rev.* 1985;7:147–177. Reprinted in: Susser M, ed. *Epidemiology, Health, and Society: Selected Papers*. New York, NY: Oxford University Press; 1987: 22–49.
18. Merton RK, Reader GC, Kendall PL. *The Student Physician: Introductory Studies on the Sociology of Medical Education*. Cambridge, Mass: Harvard University Press; 1957.
19. Francis T. Quoted by: Wegman M. Thomas Francis, Jr.: an appreciation. *Arch Environ Health*. 1970;21:230–233.

26

The Eco- in Eco-Epidemiology

In 1996, Mervyn Susser and Ezra Susser proposed breaking the constraints of the risk factor paradigm. They envisioned for the future an eco-epidemiology that would explicitly recognize multilevel causation and emphasize the ties that bind epidemiology to public health (Susser M and Susser E, 1996a, 1996b; Schwartz et al. 1999). Others also advocated for a transition in epidemiology and articulated various perspectives on what it should be (Krieger 1994; Kuh and Ben-Shlomo 1997; Perera 1997; Rapport et al. 1998; McMichael 1999). Since then, these perspectives have begun to merge into a common vision of a broad and integrative epidemiology, in which studies designed to identify risk factors would be balanced by studies designed to answer other questions equally vital to public health. These would include studies of trajectories of health and disease over the life course, the effects of social contexts broad and narrow, the spread of communicable diseases and behaviours through populations, genetic causes, and historical trends. By fully incorporating these elements, epidemiology would be rooted in the investigation of the pathways by which biological and social experiences generate health

Reprinted from *International Journal of Epidemiology*, Vol. 35 (6) 1379–1383, Dana March and Ezra Susser, Copyright (2006) with permission from Oxford Journals.

and disease, and would be equipped to identify the impact of biological and social changes on the health of populations.

Developments in epidemiology over the ensuing decade attest to the movement in this direction (Susser E 2004). In the International Journal of Epidemiology, these elements are represented and embraced as being integral to the discipline (Ben-Shlomo 2005; Davey Smith 2005). The trend is also evident, to varying degrees, in other journals, conferences, and research underway. It is quite apparent that the identification of risk factors will not be sufficient for epidemiologists to confront some of the most pressing public health challenges of our time, such as those emanating from the AIDS pandemic, social inequalities, and movements of and changes in populations. Other strategies are being applied and refined, reminiscent of the progress in previous decades during which the designs for studying risk factors were honed.

With the aim of nurturing this transition, we offer an historical perspective on the "eco" in eco-epidemiology, and on the recent resurgence of interest in the elements encompassed by it. We suggest that this rubric carries forward a line of thinking that has deep roots in previous eras and is well suited to the challenges of the present era in epidemiology. It is our hope that ultimately the eco- in eco-epidemiology, still embryonic in form, will no longer be required for the sake of differentiation from the dominant paradigm.

Transitions in Epidemiological Thinking

Thomas Kuhn's notion of the scientific paradigm is widely known and useful for recognizing how certain ideas have achieved dominance in successive eras of epidemiology (Kuhn 1962/1996). Kuhn focuses on the normative constraints of scientific cognition and activity, which he terms "paradigms." Although he analyzes the conditions under which scientific paradigms change, he does not explain how science moves from one paradigm to the next. In contrast, Ludwik Fleck, who preceded and anticipated Kuhn, illuminates the underlying intellectual and social processes that provoke scientists to change their thinking (Fleck 1935). In this sense, his work addresses more directly the way in which paradigms shift in epidemiology.

For Fleck, all scientific facts develop within the context of "thought collectives." These exist "wherever two or more people are actually exchanging thoughts . . . [which] creates a condition in which each utters thoughts he would not have been able to produce by himself or in different company" (Fleck 1935). The broader scientific culture represents a "thought style." Thought styles both shape and are shaped by thought collectives.

Fleck illustrates the process by which ideas develop within a given thought style. He analyzes the discovery of the biological agent responsible for syphilis and the development of the Wasserman reaction. Both resulted from the confluence of several lines of collective thought, and thus cannot be attributed to a single individual. In the words of Fleck as translated by F Bradley in the 1979 edition:

> [A]n interaction exists between that which is known and the act of cognition. What is already known influences the particular method of cognition; and cognition, in turn, enlarges, renews, and gives fresh meaning to what is already known. Cognition is therefore not an individual process of any theoretical ''particular consciousness.'' Rather it is the result of a social activity, since the existing stock of knowledge exceeds the range available to any one individual. (Fleck 1935)

We propose that the ideas of Fleck and Kuhn can be meaningfully integrated to offer a perspective on transitions in epidemiological thinking. Thought styles in epidemiology can be seen as emerging from a critical concentration of collective thought, and a paradigm as crystallizing a given thought style. From this vantage point, thought collectives peripheral to the mainstream of the discipline in one era may establish a thread for a new paradigm in a subsequent era. Indeed, the concepts of eco-epidemiology originated in a thought collective that coalesced in the decades after World War II, and its roots can be traced to even earlier antecedents.

The Conceptual Origins of Eco-Epidemiology

The intellectual foundation of eco-epidemiology was cemented, for the most part, in the same years as the risk factor paradigm. Jerry Morris and Mervyn Susser contributed significantly to an epidemiological understanding of multiple causation and other concepts central to the risk factor approach. However, they situated these ideas within a broader rubric.

In *Uses of Epidemiology*, first published in 1957, Morris advanced the notion that populations have unique properties representing more than just the amalgamation of their constituent properties (Morris 1957; Davey Smith 2001). He also recognized the individual as a dynamic system in which small changes might have large consequences for health, and in which intervention might have unintended consequences in the context of 'multiple causality':

> [I]f the causes are related in a dynamic way to each other, and if one or more perform a homeostatic function, it must not be assumed that simple interference will

> have simple results. Thus it is supposed to be a common experience that men who give up smoking may put on weight. (Morris 1957)

Nearly two decades later, Mervyn Susser formally introduced the concept of levels of organization into epidemiology. He emphasized that the determinants of health on the individual level differ from those on the population level, despite the fact that populations comprise individuals, and that there are difficulties inherent to analytical movement between them. He gave equal weight to the hazards of the "ecological fallacy" (inferring causation at the individual level from population level comparisons) and the "atomistic fallacy" (inferring causation at the population level from individual level comparisons) (Susser M 1973).

Thus, Morris and Susser viewed health in terms of dynamic states influenced by factors on multiple levels, such as the cellular, the individual, the community, and the population. As intimated by Morris and explicitly articulated by Susser, this constituted an "ecological" perspective. According to Susser, "[E]cology embraces the interrelations of all living things. Epidemiology could be described as human ecology, or that large part of human ecology relating to states of health." (Susser M 1973) (To avoid confusion, we note that the terms ecologic and eco-epidemiology have been used in different ways by other epidemiologists [Last and Abramson 1995].)

Morris and Susser participated in a thought collective that considered the elements central to eco-epidemiology: multiple levels of organization; the individual life course, as reflected by consideration of early antecedents and concepts of embodiment; equal consideration of communicable and noncommunicable diseases, which entailed understanding infectious causes of chronic diseases, as well as risk factors for infectious diseases; the imprint of historical time; and, the dynamic relationship between macro causes (e.g., societal change) and micro causes (e.g., genetic mutations). This thought collective was situated in an ecological thought style, evident in the work of many other epidemiologists, including John Gordon (1952), Gunnar Inghe (1958), Thomas Francis Jr. (1959), Alexander Leighton (1959), Manfred Pflanz (1962), René Dubos (1965), and John Cassel (1976). An ecological thought style was also apparent in other disciplines during this era.

Although this was perhaps the first thought collective to articulate the perspective we identify as eco-epidemiology, one can find examples of multilevel thinking in each of the previous eras of epidemiology (Susser E 2004). The best known are John Snow's work on cholera during the Sanitary era (Davey Smith 2002; Vinten-Johansen et al. 2003), and Joseph Goldberger's work on pellagra during the Infectious Disease era (Terris 1964; Kraut 2003). Each challenged the dominant ideas of the time by considering the changes in social and biological context that gave rise to an epidemic, the individual

behaviors that increased the risk of disease or its transmission, and the cellular and molecular processes underlying the pathophysiology. As early as the 1930s, Goldberger's co-worker, social scientist Edgar Sydenstricker, came close to outlining an ecological perspective on public health, if not specifically epidemiology (Sydenstricker 1933).

The Risk Factor Storm and the Shoring of Ecology

An approach like that of eco-epidemiology has the ability to conceptually subsume various branches of epidemiology within a more capacious rubric. However, such an approach did not, at the time, gain traction in the discipline.

By the mid-twentieth century, the decline in infectious disease mortality in developed countries and the rapid discoveries of effective antibiotics and vaccines seemed to herald an imminent triumph over diseases caused by microorganisms. The thought style that crystallized into the dominant paradigm in epidemiology largely relegated communicable diseases to the discipline's periphery, signaled in 1964 by the declaration of the US Surgeon General that "the time has come to close the book on infectious diseases' " (United States Department of Health, Education and Welfare 1964).

Indeed, the time was ripe for the risk factor paradigm, tailored to noncommunicable diseases, which were referred to as chronic diseases. Risk factor studies aimed specifically at estimating the effect of an exposure on the disease risk of individuals within a given population. In these studies, the broad social context was held constant and was not within the frame of the investigation (Schwartz et al. 2006). In the cohort, case-control, and other study designs crafted during this era, it proved possible in some instances to infer a causal relationship and draw implications for prevention without understanding the biological processes that linked the exposure to the disease and without understanding the social context that gave rise to them (Savitz 1994). By the mid-1980s, an armamentarium of elegant and readily applicable study designs and analytical methods had been significantly developed (Miettinen 1985; Rothman 1986; Greenland 1987).

Thus, in a particular sociohistorical moment characterized by concern with chronic diseases, the risk factor framework achieved the status of a paradigm, forming what Kuhn would term the normal science of modern epidemiology. The ecologically informed ideas of Morris and Susser and others like them became, in essence, background noise to the overture of risk factor epidemiology. Nonetheless, at the periphery of the discipline, the theoretical germ line for eco-epidemiology continued to develop, evident in

the work of Geoffrey Rose (1985), Richard Levins and Lewontin (1985), and James Koopman et al. (1991), among others.

Turning Tides

Despite the continuing dominance of risk factor methods and their successes with chronic diseases, critiques of this approach began to proliferate in the 1990s. Among the public health problems driving that intellectual discourse were AIDS and social inequalities in health. The limited utility of the dominant risk factor methods to explain and address both began to be seen as handicapping the field.

In the subsequent decade, the discipline has witnessed remarkable developments. A particularly salient example is the response to the AIDS pandemic. The 2000 International AIDS Conference, held in Durban, South Africa—the first international AIDS conference organized in the developing world—was emblematic of the transition taking place in epidemiology.

First, the conference made explicit that AIDS was a global public health emergency that would necessarily drive much of the development of epidemiology and public health. This countered a notion in risk factor epidemiology that the scientific endeavor should be separated from public health advocacy (Rothman 1986). The conference and the social movement connected to it legitimated and invigorated what soon became a landmark in public health history—the global effort to provide treatment to AIDS patients in resource poor countries. Women's rights and family relationships were made central to prevention and treatment strategies. To their credit, epidemiologists played a major role in both of these epic changes.

Second, the conference captured what is meant by eco-epidemiology. An overarching theme of the conference was to recognize causes on multiple levels and advance both the qualitative and quantitative methods required to investigate them. Ecological concepts of illness were evident in discussions of mutations in the virus, host resistance, and social phenomena, and the interactions among them. Indeed, the conference reflected the transmission of the ecological perspective across generations of epidemiologists. The successful bid to host the 2000 conference in Durban was led by the Head of South Africa's AIDS Control Programme at the time, Quarraisha Abdool Karim; the scientific chair of the conference was Salim Abdool Karim. Both had studied epidemiology with Mervyn Susser and Zena Stein. Susser and Stein originated in South Africa and had been strongly influenced by Sidney Kark (Oppenheimer and Rosner 2002), who applied multilevel thinking in a classic analysis of the forced migratory labor system as an underlying cause of epidemic syphilis in Natal (now Kwazulu/Natal) in the 1940s (Kark 2003;

Myer et al. 2003). Among many other manifestations of this continuity, parallel sessions were organized to ensure that affected communities and individuals were represented and engaged, an approach advocated by Kark, Susser, and Stein.

Third, analyses of the dynamics of HIV transmission inserted an innovative quantitative strand. In the keynote speech, Roy Anderson drew attention to the notion of the reproductive rate and its relevance to the AIDS pandemic. Thus, nonlinear quantitative means of examining ecological phenomena and ecological concepts of illness shared a high-profile platform in public health. The implicit challenge for the coming era was to move from the juxtaposition to the integration of these two threads.

The quantitative strand reflected other intellectual continuities. Anderson had developed this concept with Robert May, a mathematician who entered the realm of epidemiology in the late 1970s by examining the dynamic behavior of several diseases, both chronic and infectious (May 1978). May's work grew out of a thought collective of mathematicians concerned with complex systems (May 1972, 1973). May had partly propelled the development of the chaos paradigm in mathematics, which seeks to understand patterns of stability and complexity of large systems (Gleick 1987). Thus, he was equipped to provide the quantitative companion to Anderson's expertise in parasitology. While their initial work addressed the regulation of populations by infections and diseases, it increasingly focused on the movement of infectious diseases through populations. They published on the transmission dynamics of AIDS (Anderson et al. 1986; May and Anderson 1987; May 1988), and a groundbreaking textbook on infectious disease epidemiology (Anderson and May 1991).

Their ideas, too, had earlier antecedents: epidemiologists and physical scientists concerned with the dynamics of disease, beginning some six decades prior, had laid the foundation for their modeling efforts. At the beginning of the twentieth century, Sir Ronald Ross, an epidemiologist with a keen interest in mathematical modeling, began to characterize the transmission dynamics of malaria (Ross 1910, 1911, 1915), although at the time, epidemiologists largely eschewed this work (Hardy and Magnello 2002). On the heels of Ross, in the 1920s, William Ogilvy Kermack and Anderson McKendrick examined epidemic transmission and, like Ross, championed mathematical epidemiology (Kermack and McKendrick 1927; Davey Smith and Kuh 2001).

Other Developments

In the same year that the AIDS conference took place in the southern hemisphere, the first draft of the human genome was completed largely in the

northern hemisphere. Advances in genomics make it easier to see that the ecological perspective is relevant to understanding processes at the cellular and molecular level (McClellan et al. 2006). Conceptual and methodological developments allow epidemiologists to study more thoroughly genetic and nongenetic causes alike (Davey Smith and Ebrahim 2004).

Other developments over the past decade also suggest an important transition in epidemiology. Consideration of the life course is increasingly well developed and defined as an approach (Kuh and Ben-Shlomo 2004), building upon early studies of the fetal origins of adult disease and health trajectories over the life span (Herbst et al. 1971; Stein et al. 1975; Wadsworth 1991). Psychiatric epidemiology, with its rich ecological tradition—Morris addressed in *Uses of Epidemiology* what he termed an "ecology of mental disorders" (Morris 1957)—is being integrated into mainstream epidemiology (Susser E et al. 2006). Moreover, social epidemiology continues to develop in order to better address the implications of inequalities—rife within high and low income societies alike—for health (Berkman and Kawachi 2000; Phelan et al. 2004); examination of contextual factors by means of multilevel analyses is central to such endeavors (Diez-Roux 2000). In addition, the field is rediscovering the central importance of analyzing historical trends (Leon et al. 1997), and of improving the methods for doing so, such as age-period cohort analyses (Chen et al. 2003).

New Depths

There is no turning back toward a more circumscribed epidemiological science that is less connected to public health applications. These advances represent what Kuhn would call "scientific achievements… supplying the foundation for… further practice" (Kuhn 1962/1996).

Still, the risk factor paradigm remains the dominant if not exclusive focus of training in epidemiology, especially in the United States, as reflected in textbooks and curricula. Thus, the training of new epidemiologists lags somewhat behind the development of the field. It is true that the study of risk factors remains the most elaborated and widely applied strategy. Other approaches have not yet achieved the discipline-defining salience of the successes of risk factor epidemiology at the height of the chronic disease era. It is also true, however, that without advancing and integrating these other elements, epidemiology will diminish in importance as a central science of public health in the current era.

We envision for the future an epidemiology that—informed by its rich history—adopts, develops, and teaches methods to understand both biological and social complexities, thereby extending its range of application.

The charge of the next generation of epidemiologists is to bring about these changes—in the words of Fleck, to enlarge, renew, and give fresh meaning to what is already known. For those already established in the field, a priority is to provide this next generation with the conceptual and methodological tools that will enable them to do so. Engendering awareness of the historical roots from which they can grow and develop, essentially creating an intergenerational and international thought collective, is at once a continuation and a beginning.

Acknowledgments

We thank several members of our thought collective—Sarah Conover, Gerald Oppenheimer, Sharon Schwartz, Zena Stein, and Mervyn Susser—for their insights.

References

Anderson RM, May RM (1991). *Infectious Diseases of Humans: Dynamics and Control.* Oxford: Oxford University Press.

Anderson RM, Medley GF, May RM, Johnson AM (1986). A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J Math Appl Med Biol* 3:229–263.

Ben-Shlomo Y (2005). Real epidemiologists don't do ecological studies? *Int J Epidemiol* 34:1181–1182.

Berkman L, Kawachi I (eds) (2000). *Social Epidemiology.* New York: Oxford University Press.

Cassel J (1976). The contribution of the social environment to host resistance: the fourth Wade Hampton Frost lecture. *Am J Epidemiol* 104:107–123.

Chen X, Li G, Unger J, Liu X, Johnson C (2003). Secular trends in adolescent never smoking from 1990 to 1999 in California: an age-period-cohort analysis. *Am J Public Health* 93:2099–2104.

Davey Smith G (2001). The uses of "Uses of Epidemiology." *Int J Epidemiol* 30:1146–1155.

Davey Smith G (2002). Commentary: Behind the Broad Street pump: aetiology, epidemiology and prevention of cholera in mid-19th century Britain. *Int J Epidemiol* 31:920–932.

Davey Smith G (2005). Equal, but different? Ecological, individual and instrumental approaches to understanding determinants of health. *Int J Epidemiol* 34:1179–1180.

Davey Smith G, Ebrahim S (2004). Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 33:30–42.

Davey Smith G, Kuh D (2001). Commentary: William Ogilvy Kermack and the childhood origins of adult health and disease. *Int J Epidemiol* 30:696–703.

Diez-Roux AV (2000). Multilevel analysis in public health research. *Annu Rev Public Health* 21:171–192.

Dubos R (1965). *Man Adapting*. New Haven, CT: Yale University Press.

Fleck L (1935). Genesis and development of a scientific fact. (See 1979 edition with comment by kuhn).

Francis T (1959). The epidemiological approach to human ecology. *Am J Med Sci* 237:677–684.

Gleick J (1987). *Chaos*. New York: The Penguin Group.

Gordon JE (1952). The Twentieth Century—yesterday, today, and tomorrow (1920-). In: Top FH (ed). *The History of American Epidemiology*. St. Louis: The C.V. Mosby Company, pp. 114–167.

Greenland S (1987). Interpretation and choice of effect measures in epidemiologic analysis. *Am J Epidemiol* 125:761–768.

Hardy A, Magnello ME (2002). Statistical methods in epidemiology: Karl Pearson, Ronald Ross, Major Greenwood and Austin Bradford Hill, 1900–1945. *Soz Praventivmed* 47:80–89.

Herbst AL, Ulfelder H, Poskanzer DC (1971). Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *New Engl J Med* 284:878–881.

Inghe G (1958). Mental and physical illness among paupers in Stockholm. *Acta Neurol Scand Suppl* 121:1–316.

Kark SL (2003). The social pathology of syphilis in Africans. *Int J Epidemiol* 32:181–186.

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proc Royal Soc London* 115:700–721.

Koopman JS, Longini IM, Jacquez JA, Simon CP, Ostrow DE, Martin WR, Woodcock DM (1991). Assessing risk factors for transmission of infection. *Am J Epidemiol* 133:1199–1209.

Kraut A (2003). *Goldberger's War: The Life and Work of a Public Health Crusader*. New York: Hill and Wang.

Krieger N (1994). Epidemiology and the web of causation: has anyone seen the spider? *Soc Sci Med* 39:887–903.

Kuh D, Ben-Shlomo Y (eds.) (1997). *A Life Course Approach to Chronic Disease Epidemiology*. Oxford: Oxford Medical Publications,.

Kuh D, Ben-Shlomo Y (eds) (2004). *A Life Course Approach to Chronic Disease Epidemiology*. New York: Oxford University Press.

Kuhn TS (1962/1996). *The Structure of Scientific Revolutions*. 3rd ed. Chicago: University of Chicago Press.

Last JM, Abramson JH (1995). *A Dictionary of Epidemiology*. New York: Oxford University Press.

Leighton AH (1959). *My Name is Legion: Foundations for a Theory of Man in Relation to Culture*. New York: Basic Books.

Leon DA, Chenet L, Shkolnikov V et al. Zakharovs, Shapiro J, Rakhmanoua G, Uassin S, Mckee M. (1997). Huge variation in Russian mortality rates 1984–94: artefact, alcohol, or what? Lancet 350:383–388.

Levins R, Lewontin R (1985). *The Dialectical Biologist*. Cambridge, MA: Harvard University Press.

May RM (1972). Will a large complex system be stable? *Nature* 238:413–414.

May RM (1973). Stability and complexity in model ecosystems. *Monogr Popul Biol* 6:1–235.
May RM (1978). Dynamical diseases. *Nature* 272:5655–5673.
May RM (1988). HIV infection in heterosexuals. *Nature* 331:655–666.
May RM, Anderson RM (1987). Transmission dynamics of HIV infection. *Nature* 326:137–142.
McClellan J, Susser E, King M (2006). Maternal famine, de novo mutations, and schizophrenia. *JAMA* 296:582–584.
McMichael AJ (1999). Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. *Am J Epidemiol*149: 887–897.
Miettinen OS (1985). *Theoretical Epidemiology*. New York: Wiley.
Morris JN (1957). *Uses of Epidemiology*. Edinburgh: Livingstone.
Myer L, Morroni C, Susser E (2003). Commentary: The social pathology of the HIV/AIDS pandemic. *Int J Epidemiol* 32:189–192.
Oppenheimer GM, Rosner D (2002). Two lives, three legs, one journey: a retrospective appreciation of Zena Stein and Mervyn Susser. *Int J Epidemiol* 31:49–53.
Perera FP (1997). Environment and cancer: who are susceptible? *Science* 278: 1068–1073.
Pflanz M (1962). *Sozialer Wandel und Krankheit (Social Change and Disease)*. Stuttgart: Ferdinand Enke Verlag.
Phelan J, Link B, Diez Roux A, Kawachi I, Levin B (2004). Fundamental causes of social inequalities in mortality: a test of the theory. *J Health Soc Behav* 45:265–285.
Rapport D, Costanza R, Epstein P, Gaudet C, Levins R (eds) (1998). *Ecosystem Health*. Malden, MA: Blackwell Science.
Rose G (1985). Sick individuals and sick populations. *Int J Epidemiol* 14:32–38.
Ross R (1910). *The Prevention of Malaria*. London: John Murray.
Ross R (1911). Some quantitative studies in epidemiology. *Nature* 87:466–467.
Ross R (1915). Some a priori pathometric equations. *BMJ* 1:546–547.
Rothman KJ (1986). *Modern Epidemiology*. Boston: Little, Brown and Co.
Savitz DA (1994). In defense of black box epidemiology. *Epidemiology* 5:550–552.
Schwartz S, Susser E, Susser M (1999). A future for epidemiology? *Annu Rev Public Health* 20:15–33.
Schwartz S, Diez Roux AV, Susser E (2006). Causal explanation outside the black box. In: Susser E, Schwartz S, Morabia A, Bromet E. *Psychiatric Epidemiology: Searching for the Causes of Mental Disorders*. New York: Oxford University Press, pp. 441–460.
Stein ZA, Susser M, Saenger G, Marolla F (1975). *Famine and Human Development: the Dutch Hunger Winter of 1944–1945*. New York: Oxford University Press.
Susser E (2004). Eco-epidemiology: thinking outside the black box. *Epidemiology* 15:519–520.
Susser E, Schwartz S, Morabia A, Bromet EJ (2006). *Psychiatric Epidemiology: Searching for the Causes of Mental Disorders*. New York: Oxford University Press.
Susser M (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology*. New York: Oxford University Press.
Susser M, Susser E (1996a). Choosing a future for epidemiology: I. Eras and paradigms. *Am J Public Health* 86:668–73.
Susser M, Susser E (1996b). Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *Am J Public Health* 86:674–77.

Sydenstricker E (1933). *Health and Environment*. New York: McGraw Hill.
Terris M (ed) (1964). *Goldberger on Pellagra*. Baton Rouge: Louisiana State University Press.
United States Department of Health, Education and Welfare (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General*. Washington, DC: Public Health Service.
Vinten-Johansen P, Brody H, Paneth N, Rachman S, Rip MR (2003). *Cholera, Chloroform and the Science of Medicine: A Life of John Snow*. New York: Oxford University Press.
Wadsworth MEJ (1991). *The Imprint of Time: Childhood, History and Adult Life*. Oxford: Clarendon Press.

This page intentionally left blank

Index

NOTE: Page references in italics refers to illustrations and tables.