Understanding the Fundamentals of Epidemiology

an evolving text

Victor J. Schoenbach, Ph.D.

with

Wayne D. Rosamond, Ph.D.

Department of Epidemiology
School of Public Health
University of North Carolina at Chapel Hill

Fall 2000 Edition

© 1999, 2000 Victor J. Schoenbach

Unless otherwise indicated, the text and diagrams in this work belong to the copyright owner above. For reprint permission (royalty-free for noncommercial use by nonprofit, accredited, educational organizations), please write to:

Victor J. Schoenbach, Ph.D.

Department of Epidemiology

University of North Carolina

School of Public Health

Chapel Hill, NC 27599-7400 USA

Victor_Schoenbach@unc.edu

Permission to reprint material copyrighted by others and used here by their permission must be obtained directly from them.

August 1999, 2000

Chapel Hill, North Carolina

Preface

Introductory epidemiology courses are often referred to as "methods" courses, and many students come to them hoping to learn the methods that have made epidemiology so important. Certainly methods are an essential aspect of the field, and this text covers the usual complement. But especially for the newcomer, the critical need is to learn how epidemiologists think about health and the factors that affect it, and how epidemiologists approach studying them. Very few methods are unique to epidemiology. "Epidemiologic thinking" is its essence. Therefore, for me the central objective of an introductory course has been to explain the concepts and perspectives of the field.

For nearly 20 years I have had the privilege of teaching the introductory epidemiology course for epidemiology majors at the University of North Carolina School of Public Health and the special pleasure that derives from teaching students who have sought epidemiology out rather than come to learn it only as a school requirement. I have also had the honor of being entrusted by my colleagues with the responsibility for introducing our students to epidemiologic concepts and methods.

Over the years I have written out extensive lecture notes, initially in response to requests from course participants and subsequently to develop my own understanding. Not all course participants have appreciated them, but I have received sufficient positive feedback and expressions of interest from graduates who have gone on to teach their own epidemiology courses that I have decided to recast them as an "evolving text". I use the term "evolving" because I continue to clarify, develop, refine, correct, and, I hope, improve.

Regarding it as an evolving text is also my excuse for the fact that the material is not ready for formal publication. Moreover, unlike a published text, this volume does not claim to be authoritative – nor even thoroughly proofread. As an evolving work, its further development has always taken priority over appearance – and, it must be admitted, occasionally also over accuracy.*

Although the word processing is nearly all my own, the content is certainly not. Besides the extensive development and exposition of epidemiologic concepts and methods from courses and publications by others, I have had the good fortune to study with and learn from outstanding epidemiologists and biostatisticians, among them the late John Cassel, Gerardo Heiss, Barbara Hulka, Michel Ibrahim, Sherman James, Bert Kaplan, David Kleinbaum, Gary Koch, Lawrence Kupper, Hal Morgenstern, Abdel Omran, the late Ralph Patrick, Dana Quade, David Savitz, Carl Shy, the late Cecil Slome, H.A. Tyroler, and Edward Wagner.

---

* Important errata, as I learn about them, are posted on a site on the World Wide Web (http://www.epidemiolog.net/).

My thinking and this text have also greatly benefited from interactions with other colleagues and teachers, co-instructors, teaching assistants, collaborators, associates, research staff, fellows, and students. I must particularly acknowledge the assistance of Charles Poole, who has generously shared his expertise with me through his advanced methods course and frequent consultations. He has even made the ultimate sacrifice – reading this text and sitting through my lectures! The content (errors excepted!) and to some extent the exposition, therefore, represent the knowledge, ideas, examples, and teaching skills of many people, to a much greater extent than the specific attributions, citations and acknowledgements would indicate.

Acknowledgements are of greater interest to authors than to readers, and I ask your forgiveness for including several more. I received my own introduction to epidemiology from the late John Cassel -- intellectual pioneer, inspiring lecturer, and humanist -- and Bert Kaplan -- quintessential scholar, supporter, and friend, whose colleagueship, breadth of knowledge, depth of wisdom, dedication to the ideals of the academy, and personal warmth have enriched the lives of so many. I would also like to express my gratitude to colleagues, staff, secretaries (especially Pat Taylor, Edna Mackinnon Lennon, and Virginia Reid), students, administrators, and family for inspiration, stimulation, feedback, opportunity, advice, guidance, commitment, counseling, assistance, support, affection, and a good deal more.

Enjoy Epidemiology!

Victor J. Schoenbach

Chapel Hill, North Carolina

U.S.A.

August 17, 1999

Postscript: After the 20th anniversary edition of EPID 168 ("Fundamentals of epidemiology"), my teaching responsibilities have changed to its sister course, EPID 160 ("Principles of epidemiology"). EPID 160 serves as the basic introductory course for all students, graduate and undergraduate, who are not majoring in epidemiology. Thus its audience is much more diverse in both interests and preparation. Time will tell if I am able to continue to refine the *Evolving Text*, but if so it will begin to move in the direction of making it more suitable for a general – and international – readership. I have been gratified by the expressions of interest in it in its present form and hope that it will continue to be of use to others.

March 9, 2001.

Table of Contents

Chapter (in Acrobat ®, click on a chapter name to move to that page) **Page***

Preface 1

1. Epidemiology — Definition, functions, and characteristics 3
2. An evolving historical perspective 17
3. Studying populations - basic demography 31
   - Assignment 53
   - Solutions 57
4. The Phenomenon of Disease 59
5. Measuring Disease and Exposure 81
   - Appendix on weighted averages 113
   - Appendix on logarithms 115
   - Assignment 117
   - Solutions 123
6. Standardization of rates and ratios 129
   - Assignment 149
   - Solutions 153
7. Relating risk factors 161
   - Appendix 199
   - Assignment 201
   - Solutions 205
8. Analytic study designs 209
   - Assignment 257
   - Solutions 265
9. Causal inference 269
10. Sources of error 287
    - Appendices 319
    - Assignment 325
    - Solutions 329

11. Multicausality — Confounding 335
Assignment 373
Solutions 377

12. Multicausality — Effect modification 381
Assignment 413
Solutions 417

13. Multicausality — Analysis approaches 423
Assignment (see next chapter)

14. Data analysis and interpretation 451
Assignment 499
Solutions 503

15. Practical aspects of epidemiologic research 507

16. Data management and data analysis 523

17. Epidemiology and public health 551

18. Overview and Conclusion 565

* Note: page numbers do not exactly match the number of the physical page because of unnumbered pages and are not in exact sequence because of chapter revisions or pages inserted from other sources (e.g., published articles for assignments).

Index: although no index is available, the Acrobat® Reader has powerful search capabilities (see the Edit Menu). You can search through all chapters at once by viewing the PDF that contains all of the chapters and assignments in a single document.

Other epidemiology learning materials and resources, including practice examinations, may be found at www.epidemiolog.net

1. Introduction

Definition, characteristics, uses, varieties, and key aspects of epidemiology

What to tell your family and friends

When your family or friends ask what you are studying, and you say "epidemiology", the response is often something like:

> "You're studying *what*?"
>
> "Does that have something to do with skin?"
>
> "Uh-huh. And what *else* are you studying?"

How should you reply? One possibility is to give a formal definition (e.g., "The study of the distribution and determinants of health related states and events in populations, and the application of this study to control health problems" [John M. Last, *Dictionary of Epidemiology*]). Another possible reply is, "Well, *some* epidemiologists study the skin. But epidemiologists study all kinds of diseases and other aspects of health, also. The root word is 'epidemic', rather than 'epidermis'." Another reply could be. "Epidemiology is the study of health and disease in populations. It's a basic science of public health.", though then be prepared to define "public health". And, if you're feeling erudite, you can follow-up with, "'Epidemiology' comes from the Greek *epi* (among, upon), *demos* (people), and *logy* (study)."

Epidemiology in transition?

The above should satisfy your friends, but what about yourself? Particularly if you are entering on the pathway to becoming an epidemiologist, do you know where it will lead you? According to Thomas Kuhn (1970:136-7), textbooks "address themselves to an already articulated body of problems, data, and theory, most often to the particular set of paradigms to which the scientific community is committed at the time they are written.…[They] record the stable *outcome* of past revolutions and thus display the bases of the current normal-scientific tradition". Raj Bhopal's review (1997), however, reports that recent epidemiology texts present a diversity of concepts and information, even in regard to the building blocks of epidemiology. Bhopal sees the fundamental question as "whether epidemiology is primary an applied public health discipline…or primarily a science in which methods and theory dominate over practice and application". He predicts a lively discussion that will sharpen in the 21st century.

Indeed, in the leading commentary in the August 1999 issue of the *American Journal of Public Health*, three of my colleagues including our department chair seek to differentiate between epidemiology (a "science") and public health (a "mission"). They argue that the second half of Last's definition (application and control) describes "the broader enterprise of public health" rather than epidemiology. Epidemiology "contributes to the rationale for public health policies and services and is important for use in their evaluation", but "the delivery of those services or the implementation of those policies" is not "part of epidemiology" (Savitz et al., 1999: 1158-1159). Further, "the product

of research is information, not, as has been argued, 'public health action and implementation' (Atwood et al., 1997: 693)." (Savitz et al.: 1160).

The article by David Savitz, Charles Poole, and William Miller might be regarded in part as a response to the charge made in an article by our previous chair, Carl Shy, that academic epidemiology has "failed to develop the scientific methods and the knowledge base to support the fundamental public health mission of preventing disease and promoting health through organized community efforts" (Shy, 1997). In making this charge, Shy builds on the contention in the Institute of Medicine report on *The Future of Public Health* (Committee for the Study of the Future of Public Health, 1988, which asserted that the U.S. public health system was in "disarray") that schools of public health are too divorced from public health practice. In that vein, in the editorial that precedes the Savitz et al. commentary, the current Director of the Centers for Disease Control and Prevention (CDC) and two of his colleagues assert that, "[Epidemiologists] can make their goal journal publication, public interpretation of findings, or public health interventions", adding that "epidemiology's full value is achieved only when its contributions are placed in the context of public health action, resulting in a healthier populace". (Koplan et al., 1999).

These contrasting positions are not necessarily in conflict. To say that public health action is required to achieve epidemiology's full value does not imply that epidemiology or epidemiologists must launch that public health action, nor does appreciation of epidemiologists' contributions imply that those contributions are epidemiology (as opposed to good works that happen to be done by epidemiologists). Nancy Krieger's challenge:

> "Do epidemiologists and other public health professionals have a responsibility to ask whether the ways we think and work reflect or contribute to social inequality?
> "Proponents of socially responsible science would answer yes. What say you?"
> (Krieger, 1999: 1152)

is more difficult to integrate and is perhaps unfair. But whether we see ourselves first as scientists or first as public health professionals, our work takes place in a societal context, with resources and therefore priorities assigned by political and economic institutions that appear to serve the interests of some people and groups more than of others (Winkelstein, 2000). The research we do and our behavior in our other professional activities inevitably reflect our backgrounds and life experiences, our values and preconceptions, our personal ambitions and responsibilities. In that sense, what is epidemiology and what is not, and who is an epidemiologist and who is not, are determined in part by the custodians of curricula, hiring, research funding, and publication. Thus, you have an opportunity to make epidemiology what you think it should be.

You may, however, need to help develop the methods to enable epidemiologists to do what you think we should. In recent decades there have been great strides in the development and teaching of epidemiologic concepts and methods to study health problems of the individuals in a population, but these concepts and methods are less adequate for understanding population health (Koopman and Lynch, 1999), even in regard to epidemics – the origin of our discipline and its name. Advances in the methods for the study of health and disease in populations – epidemiology's calling card, as it were – may ease some of the apparent conflict between those who see epidemiology first as a scientific enterprise and those who see it foremost as a vehicle for solving major public health

problems (Schwartz and Carpenter, 1999). Even if epidemiologists are willing to study problems that cannot be solved within the prevailing paradigm and the conceptual and instrumental tools that it supplies (Kuhn, 1970), understanding those problems will require effective concepts and methods. Warren Winkelstein (2000) sees the need for a "more expansionist approach" in order to address disease problems arising from pollution, global warming, population growth, poverty, social inequality, civil unrest, and violence. Even without taking the further step of proposing that epidemiology should attempt to reduce these conditions themselves, the challenges for epidemiology are daunting.

Epidemiology functions and areas of application

The perspective in this text is that epidemiology is both a field of research to advance scientific understanding and also of application of knowledge to control disease and advance public health, a (primarily observational) science and a public health profession. Thus, epidemiologists conduct research and also work to control and prevent disease; they are scientists and engineers. Epidemiologic investigation is problem-oriented and tends toward applied research. Although it has a growing body of theory, the field is primarily empirically driven. Partly for these reasons, epidemiologists draw freely from other fields and gravitate towards multidisciplinary approaches.

Milton Terris, a leading exponent of close interrelationships among epidemiology, public health, and policy, has summarized the functions of epidemiology as:

1. Discover the agent, host, and environmental factors that affect health, in order to provide the scientific basis for the prevention of disease and injury and the promotion of health.
2. Determine the relative importance of causes of illness, disability, and death, in order to establish priorities for research and action.
3. Identify those sections of the population which have the greatest risk from specific causes of ill health [and benefit from specific interventions], in order that the indicated action may be directed appropriately. (targeting)
4. Evaluate the effectiveness of preventive and therapeutic health programs and services in improving the health of the population.

   (Milton Terris, The Society for Epidemiologic Research (SER) and the future of epidemiology. *Am J Epidemiol* 1992; 136(8):909-915, p 912)

To these might be added:

5. Studying the natural history of disease from its precursor states through its manifestations and clinical course
6. Conducting surveillance of disease and injury occurrence in populations and of the levels of risk factors – passive (receive reports), active (poll practitioners, conduct surveys)

7. Investigating outbreaks (e.g., hospital-acquired infections, disease clusters, food-borne and water-borne infections) to identify their source and controlling epidemics (e.g., measles, rubella, coronary heart disease, overweight)

Classic and recent examples of epidemiologic investigation

Epidemiology has made significant contributions to the understanding and control of many health-related conditions, and epidemiologists are actively involved in studying many others. Some of the classic investigations and some areas of recent and current attention are listed below:

Scurvy (James Lind) - intervention trial, nutritional deficiency

Scrotal cancer (Percival Pott) - occupational health, carcinogens

Measles (Peter Panum) - incubation period, infectious period

Cholera (John Snow) - waterborne transmission, natural experiment

Puerperal fever (Ignatius Semmelweis) - hygienic prevention

Pellagra (Joseph Goldberger) - "epidemic" disease was not communicable

Rubella and congenital birth defects (Gregg) - prenatal exposure

Retrolental fibroplasia - iatrogenic disease

Lung cancer and smoking - coming of age of chronic disease epidemiology

Fluoride and dental caries - community epidemiology; environmental prevention

Poliomyelitis immunization trial - a massive experiment that demonstrated the effectiveness of the vaccine against this greatly feared virus

Cardiovascular disease - longitudinal community studies; community intervention trials

Breast cancer screening – a large-scale randomized trial of effectiveness of cancer early detection through screening

Reye's syndrome and aspirin - an epidemiologic success involving a rare but devastating disease brought on by a familiar and ubiquitous medicine

Toxic shock syndrome - an epidemiologic success in a "point-source" epidemic resulting from a new product introduction

Estrogens and endometrial cancer - controversies of case-control methodology and bias; pharmacoepidemiology

Psychiatric disorder - challenges in disease classification and assessment

Lead and cognitive development - a crucial role for a biologic marker

Electromagnetic fields - can an exposure be "exonerated"?

Legionnaire's disease - a newly recognized pathogenic bacterium foreshadows the resurgence of infectious diseases as a public health challenge in the U.S.

HIV - a new or newly-recognized virus that has transformed the public health and epidemiology landscape with respect to infectious diseases in general and sexually-transmitted infections specifically

Tuberculosis - reminding epidemiology of its roots; control of a pathogen is very different from its eradication

Injury - epidemiology without disease

Homicide - a behavioral epidemic or an environmental plague?

Varieties of epidemiology

As epidemiology continues to develop and to expand into new areas, the field has diversified into many forms:

Surveillance, "shoe-leather" epidemiology (outbreak investigations), and epidemic control

Microbial epidemiology – biology and ecology of pathogenic microorganisms, their lifecycles, and their interactions with their human and non-human hosts

Descriptive epidemiology – examination of patterns of occurrence of disease and injury and their determinants

"Risk factor" epidemiology – searching for exposure-disease associations that may provide insights into etiology and avenues for prevention

Clinical epidemiology[*] and the evaluation of healthcare – assess accuracy, efficacy, effectiveness, and unintended consequences of methods of prevention, early detection, diagnosis, treatment, and management of health conditions

Molecular epidemiology – investigate disease at the molecular level to precisely characterize pathological processes and exposures, to elucidate mechanisms of pathogenesis, and to identify precursor conditions

---

[*] In David Sackett et al.'s *Clinical Epidemiology*, 2nd ed, it is recounted that when one of the authors (P.T.), then a medical student in England "sought career guidance from a world-renowned London epidemiologist, he was informed that it was 'amoral' to combine epidemiology with clinical practice!"

[**] "Big" in epidemiology might be defined as upwards of $100 million for a study. To put these studies in perspective, the Human Genome Project cost $250 million in public funds, CERN (high energy particle physics research in Switzerland) $638 million/year, the Hubble Space Telescope $3 billion, and the Apollo Program $115 billion. (1999 dollars; data from the National Institutes of Health, the European Space Agency, and NASA, by way of Hannah Fairfield in the *New York Times* (Science Times, 6/27/2000).

Genetic epidemiology – the confluence of molecular biology, population studies, and statistical models with an emphasis on heritable influences on disease susceptibility and expression

Big Epidemiology** – multisite collaborative trials, such as the Hypertension Detection and Follow-up Program (HDFP), Coronary Primary Prevention Trial (CPPT), Multiple Risk Factor Intervention Trial (MRFIT), Women's Health Initiative (WHI)

Entrepreneurial epidemiology – building institutions and careers by winning research funding and facilities

Testimonial epidemiology – giving depositions and testifying in court or in legislative hearings on the state of epidemiologic evidence on a matter of dispute

Social epidemiology – interpersonal and community-level factors influencing health at the population level

Global epidemiology – assessing the effects of human activity on the ecosystem that supports life on Earth.

Characteristics of epidemiology

With so many varieties of epidemiology, it is no wonder that confusion abounds about what is and what is not epidemiology. "Epidemiologic" research tends to:

be observational, rather than experimental;

focus on free-living human populations defined by geography, worksite, institutional affiliation, occupation, migration status, health conditions, exposure history, or other characteristics rather than a group of highly-selected individuals studied in a clinic or laboratory;

deal with etiology and control of disease, rather than with phenomena that are not closely tied to health status;

take a multidisciplinary, empirical approach directed at understanding or solving a problem rather than on advancing theory within a discipline.

However, not all epidemiologic studies have these characteristics.

So how then can you tell if someone is doing epidemiology or not? One wag suggested the following scoring system:

$$\text{score} = \frac{\ln(n^y)k^s d^2}{pc}$$

where:

n = number of subjects

y = number of years of follow-up

k = total direct costs (in $1,000,000)

s = sponsor (NIH=3, other public or foundation=2, corporate=1)

d = principal investigator's degree (EPID PhD=4, MD plus EPID MPH.= 3, MD w/o EPID MPH = 2, other health doctorate = 1)

p = number of first-authored publications that the PI will author

c = percent of the principal investigator's salary that will be covered

The higher the score, the more likely that the study is epidemiology.

Key aspects of epidemiology

A number of other fields – medicine, nursing, dentistry, pharmacy, demography, sociology, health psychology, health education, health policy, nutrition – share many common features and areas of interest with epidemiology (and with each other). Some of the key aspects of epidemiology are:

Epidemiology deals with **populations**, thus involving:

- Rates and proportions
- Averages
- Heterogeneity within
- Dynamics - demography, environment, lifestyle

As other sciences, epidemiology involves **measurement**, entailing the need for:

- Definition of the phenomena
- Spectrum of disease
- Sources of data
- Compromise

Most epidemiologic studies involve **comparison**, introducing considerations of:

- Standards of reference for baseline risk
- Equivalent measurement accuracy
- Adjustment for differences

Epidemiology is fundamentally **multidisciplinary**, since it must consider:

- Statistics, biology, chemistry, physics, psychology, sociology, demography, geography, environmental science, policy analysis, …
- Interpretation - consistency, plausibility, coherence
- Mechanisms - pathophysiology, psychosocial, economic, environmental
- Policy - impact, implications, ramifications, recommendations, controversy

Modes of investigation — descriptive vs. analytic epidemiology

Although the distinction is often difficult to draw, in part because of the greater valuation placed by many on the latter, epidemiologic investigations are sometimes usefully characterized as either **descriptive** or **analytic**.

Descriptive epidemiology

Descriptive epidemiology describes the health conditions and health-related characteristics of populations, typically in terms of **person, place, and time**. This information serves as the foundation for studying populations. It provides essential contextual information with which to develop hypotheses, design studies, and interpret results. Surveillance is a particular type of descriptive epidemiology, to monitor change over time.

Types of descriptive studies:

- Routine analyses of vital statistics (births, deaths), communicable disease reports, other notifiable events (outbreaks, induced abortions)
- Periodic surveys of health status, knowledge, beliefs, attitudes, practices, behaviors, environmental exposures, and health care encounters (e.g., National Center for Health Statistics surveys, Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System)
- Specialized surveys to establish prevalence of a condition, a characteristic, or use of a medical procedure
- Studies comparing information across geographical or political units, or between migrants and persons in their country of origin to look for differences and patterns

Analytic epidemiology

Analytic epidemiology involves the systematic evaluation of suspected relationships, for example, between an exposure and a health outcome. Because of their narrower focus, analytic studies typically provide stronger evidence concerning particular relationships.

Types of analytic studies:

- Case-control studies, comparing people who develop a condition with people who have not
- Follow-up (retrospective, prospective) studies, comparing people with and without a characteristic in relation to a subsequent health-related event
- Intervention trials (clinical, community), in which a treatment or preventive intervention is provided to a group of people and their subsequent experience is compared to that of people not provided the intervention

Analytic studies typically involve the testing of hypotheses, which in turn may arise from

- Case reports
- Case series
- Laboratory studies
- Descriptive epidemiologic studies
- Other analytic studies

The descriptive and analytic classification is more of a continuum than a dichotomy. Many studies have both descriptive and analytic aspects, and data that are collected in one mode may end up being used in the other as well. Whether a particular study is primarily "descriptive" or "analytic" may be a matter of the investigator's "stance" in relationship to the study question and the collection of the data. Since analytic epidemiology is often accorded a higher status than is descriptive epidemiology, with some regarding a study without a hypothesis as "not science", investigators sometimes feel constrained to come up with a hypothesis and present their work as "analytic", even if the hypothesis is contrived or is not the study's real focus.

Sources of data

Since epidemiology studies populations in their ordinary environments, there are many kinds of data that are relevant, and obtaining them can be logistically challenging and expensive. There is accordingly an interest in using data that are already available. Data for political and geographical aggregates are often more readily available than are data on individuals, a distinction referred to as the **level of measurement**. Sources of data for epidemiologic studies include:

Aggregate data

Vital statistics (birth rates, death rates, pregnancy rates, abortion rates, low birth weight)

Demographic, economic, housing, geographical, and other data from the Census and other government data-gathering activities

Summaries of disease and injury reporting systems and registries

Workplace monitoring systems

Environmental monitoring systems (e.g., air pollution measurements)

Production and sales data

Individual-level data

Vital events registration (births, deaths, marriages)

Disease and injury reporting systems and registries

National surveys

Computer data files (e.g., health insurors)

Medical records

Questionnaires - in person, by telephone, mailed

Biological specimens (routinely or specially collected)

Sometimes a distinction is drawn between **primary data** (collected specifically for the study, which is generally advantageous) and **secondary data** (collected for some other purpose, and therefore possibly not as well suited for the question of current interest), though the former is not inevitably superior to the latter. Although data quality is always a paramount, compromises must often be made. Two examples are the use of a **proxy informant** when the person to be interviewed is ill, demented, or deceased and the use of a **proxy variable** when data cannot be obtained for the variable of greatest relevance.

Sources of error

The challenge of data quality in epidemiology is to control the many sources of error in observational studies of human populations. The best understood and most quantifiable is **sampling error**, the distortion that can occur from the "luck of the draw" in small samples from a population. More problematic is error from **selection bias**, where the study participants are not representative of the population of interest.

Selection bias can result from:

Self selection (volunteering)

Nonresponse (refusal)

Loss to follow-up (attrition, migration)

Selective survival

Health care utilization patterns

Systematic errors in detection and diagnosis of health conditions

Choice of an inappropriate comparison group (investigator selection)

Also highly problematic is **information bias**, systematic error due to incorrect definition, measurement, or classification of variables of interest.

Some sources of information bias are:

Recall or reporting bias

False positives or negatives on diagnostic tests

Errors in assignment of cause of death

Errors and omissions in medical records

Observational sciences especially are also greatly concerned with what epidemiologists call **confounding**, error in the interpretation of comparisons between groups that are not truly comparable. Differences in age, gender composition, health status, and risk factors generally must generally be allowed for in making and interpreting comparisons. A major theme in epidemiologic methods is the identification, avoidance, and control of potential sources of error.

Unique contribution of epidemiology

In an earlier era, epidemiology was characterized as "the basic science of public health work and of preventive medicine" (Sheps, 1976:61). Whether or not this claim was ever valid (i.e., whether "the" should be "a" and whether "basic" should be "applied"), epidemiology does have the advantage of a name that ends in "logy" (a factor not to be discounted in this "era of marketing" [George McGovern's apt phrase from the 1980's]) and remains a foundation for the practice of "evidence-based medicine" (definitely a term for the era of marketing). Moreover, epidemiology deals with the "bottom line", with the reality of human health. True, epidemiologic research suffers from many limitations. Indeed, in comparison to laboratory science, epidemiology may seem somewhat crude – akin to sculpting with a hammer but no chisel. But the limitations of epidemiologic research are largely a function of the obstacles epidemiologists must contend with, and both the obstacles and the limitations are inherent in the subject of study – free-living human populations. Laboratory studies provide better control of the confounding influences of genetic, environmental, and measurement variability. But the public health relevance of laboratory findings is often uncertain due to:

Differences between *in vitro* (test tube) and *in vivo* (whole animal) systems

Differences in susceptibility across species

Difficulty of extrapolating across dosages, routes of administration, cofactors, lifespans

Problems in generalizing results from highly controlled settings to free-living populations.

Exquisitely precise knowledge about what happens in cell cultures or experimental animals, while of great value in many respects, cannot tell us enough about human health. Ultimately, public health decisions require data from human populations. If we need to know what happens to people, we must employ epidemiology.

Bibliography

NOTE: In-depth reviews of epidemiologic knowledge in both topical and methodological areas can be found in the periodical *Epidemiologic Reviews*, published by the *American Journal of Epidemiology*. The first issue of the year 2000 (Armenian and Samet, 2000) features essays addressing the current state of epidemiology in a wide range of areas and provides an excellent overview of the field.

Textbook chapters: Charles Hennekens and Julie Buring. *Epidemiology in medicine*, ch. 1-2; Kenneth Rothman. *Modern Epidemiology*, 1st ed., ch. 1; Kenneth Rothman and Sander Greenland, *Modern Epidemiology*, 2nd ed., ch 1. Brian MacMahon and Thomas Pugh. *Epidemiology: principles and methods*. 1 ed., ch. 1-4; Judith Mausner and Shira Kramer. *Epidemiology: an introductory text*., ch. 1-2; Abraham Lilienfeld and David Lilienfeld. *Foundations of epidemiology*. 2 ed, ch. 1, 12; Mervyn Susser. *Causal Thinking in the Health Sciences*; David Kleinbaum, Lawrence Kupper, Hal Morgenstern. *Epidemiologic research*, ch. 2, 3. [Complete citations for these and many other textbooks are available at www.epidemiolog.net/]

Armenian, Haroutune K., Jonathan M. Samet (eds). Epidemiology in the year 2000 and beyond. *Epidemiologic Reviews* 2000; 22(1):1-185

Bhopal RS. Which book? A comparative review of 25 introductory epidemiology textbooks. *J Epidemiol Community Health* 1997;51:612-622.

Bhopal, Raj. Paradigms in epidemiology textbooks: in the footsteps of Thomas Kuhn. *Am J Public Health* 1999; 89:1162-1165.

Duffy J. *A history of public health in NYC*. NY, Russell Sage, 1974. Especially chapter 3, Launching the NYC Health Department, 48-69.

Koopman JS, Lynch JW. Individual causal models and population systems models in epidemiology. *Am J Public Health* 1999; 89:117-1174.

Koplan, Jeffrey P.; Stephen B. Thacker, Nicole A. Lezin. Epidemiology in the 21st century: calculation, communication, and intervention. *Am J Public Health* 1999; 89:1153-1155.

Krieger N. Questioning epidemiology: objectivity, advocacy, and socially responsible science. *Am J Public Health* 1999; 89:1151-1153.

Kuhn, Thomas S. *The structure of scientific revolutions*, 2nd ed, Chicago, University of Chicago, 1970.

Marmot, Michael. Facts, opinions and affaires du coeur. *Am J Epidemiol* 1976; 103:519-526.

McGavran EG. What is public health? *Canadian Journal of Public Health* 1953 (December), 47-61.

Pattner, W.I. The public health movement. In: *From poor law to welfare state*. NY, The Free Press, Macmillan, 1974, 116-133.

Rosner D. Health care for the "truly needy", Nineteenth Century origins of the concept. *Milbank Memorial Quarterly* 1982; 60(3):355-385.

Savitz DA, Poole C, Miller WC. Reassessing the role of epidemiology in public health. *Am J Public Health* 1999; 89:1158-1161.

Schwartz S, Carpenter KM. The right answer for the wrong question: consequences of type III error for public health research. *Am J Public Health* 1999; 89:1175-1180.

Sheps, Ceil G. *Higher education for public health*. New York, Prodist for the Milbank Memorial Fund, 1976.

Shy, Carl M. The failure of academic epidemiology: witness for the prosecution. *Am J Epidemiol* 1997; 145:479-484.

Stallones, Reuel A. To advance epidemology. *Ann Rev Public Health* 1980; 1:69-82.

Terris, Milton. The epidemiologic tradition. *Public Health Reports* 1979;94(3):203-209.

Winkelstein, Jr., Warren. Interface of epidemiology and history: a commentary on past, present, and future. *Epidemiologic Reviews* 2000; 22:2-6.

Dimensions in the training of an epidemiologist

I. Epidemiologic perspective

1. Public health aspects: -- History of epidemiology, epidemiology as a public health science, clinical and public policy implications.
2. Scientific aspects: -- Problem conceptualization, philosophy of inference, study designs, interpretation of data, concepts of bias and multicausality.

II. Measurement and analysis: Measures of disease frequency and extent, study designs and strategies, control of sources of error, statistical inference, data analysis and interpretation.

III. Weighing epidemiologic evidence: Critical reading and synthesizing of information.

IV. Proposal development: Specification of research hypotheses, study populations, measurement tools, analysis strategies; human subjects protection; "grantsmanship".

V. Study design and execution: Protocol development, subject recruitment, instrumentation, data collection, quality control, reporting and communications collaboration and working with oversight bodies, presentation of findings.

VI. Data management: Manipulation and analysis of data using computers and statistical software packages.

VII. Substantive knowledge: General background in health-related sciences and multidisciplinary understanding of specific areas of research.

VIII. Epidemiologist roles: Development of skills for teaching, consultation, review of proposals and manuscripts, participation in professional meetings, leadership of multidisciplinary research teams, and continuing professional development.

(Used for a number of years by the UNC Department of Epidemiology as an outline of areas of required competencies)

2. An evolving historical perspective

The evolution of epidemiology into a science of the distribution of disease in populations and evaluation of interventions for disease prevention and therapy.

Propositions from studying history of epidemiology

1. Life has not always been the way it is in the developed countries today.
2. Scientific understanding of disease and the factors that affect it is largely a product of the last 150 years, with very rapid advances in the last half-century..
3. Epidemiologic studies have not always been like _____ (insert the name of your favorite epidemiologic study).
4. There are many histories of epidemiology
   - History of health and disease
   - History of ideas and concepts
   - History of methods
   - History of knowledge gained through these concepts and methods
   - History of teachers and students
   - History of organizations and actions

A brief history of public health

Community attempts to prevent and limit the spread of disease go back to antiquity. For example, religious traditions against eating pork and shellfish reflect the special hazards of eating those foods when inadequately preserved or prepared. As often happens in public health, even without an understanding of the underlying etiology, effective preventive measures can be taken.

Successes in prevention reinforce the concept that disease can be prevented through human action other than prayers and sacrifices to the gods, which in turn encourages additional attempts at prevention. By the 1600's, the practices of isolation and quarantine had begun to be employed to prevent the spread of certain diseases; by the 1800's these practices had become common in the American colonies. Methods of smallpox inoculation also began to be used and apparently mitigated some epidemics, even before Edward Jenner's introduction of a safe vaccine based on cowpox virus.

Acknowledgement: The following material draws heavily on lectures at the UNC Department of Epidemiology by Drs. Abraham Lilienfeld (1984) and Joellen Schildkraut (1989, 1990, 1991).

With the 19th century came two dramatic advances in the effectiveness of public health – "the great sanitary awakening" (Winslow, quoted in *The Future of Public Health* [FPH]: 58) and the advent of bacteriology and the germ theory. Those of us who see all progress in the field of health in terms of laboratory discoveries and medicines have not had the experience of living in a 19th century city. In New York City, piles of garbage two-three feet high were accompanied by epidemic smallpox and typhus. The crowding, poverty, filth, and lack of basic sanitation in the working class districts of the growing cities provided efficient breeding grounds for communicable diseases. Diseases that formerly arrived from outside to cause epidemics in basically healthy populations now became permanent residents. Quarantine and isolation, which were somewhat effective against individual cases and illness brought by travelers, were inadequate against mass endemic disease.

Moreover, industrialization and urbanization brought people of different classes geographically closer. No longer able to escape to their country estates, well-to-do families also fell prey to the highly contagious diseases that incubated among the working class. The shared vulnerability and the succession of reports of conditions in the working class supported the view that while poverty might still reflect individual weakness and moral defects, society nevertheless had to take actions to improve conditions.

In England, the Poor Law Commission led by Edwin Chadwick studied the English health of the working class. Their famous – and controversial – *General Report on the Sanitary Conditions of the Labouring Population of Great Britain* presented a "damning and fully documented indictment of the appalling conditions" (Chave, in FPH: 59-60). The studies revealed that the *average age at death* for laborers was 16 years. For tradesmen it was 22 years; for the gentry, 36 years. In London more than half of the working class died before their fifth birthday (Winslow, in FPH).

A comparable document in the United States was Lemuel Shattuck's 1850 *Report of the Massachusetts Sanitary Commission*. Unlike Chadwick's report, however, Shattuck's report went largely ignored due to the political turmoil in the United States. After the Civil War, though, many of its recommendations were adopted, and it is now regarded as one of the most influential American public health documents (FPH: 61).

Though controversial in many ways, sanitary reforms fit reasonably well with the moral views of the time. Much of the scientific rationale for the reforms – the relatively nonspecific model by which filth and putrid matter gave off emanations (miasma) that gave rise to disease – has only modest correspondence to modern biological understanding. Nevertheless, many of the reforms did reduce the transmission of disease and were therefore effective.

But the advance in understanding of infectious disease that constituted the arrival of the bacteriologic era at the end of the century dramatically increased the effectiveness of public health action. In one dramatic example, mosquito control brought the number of yellow fever deaths in Havana from 305 to 6 in a single year (Winslow, in FPH: 65). Cholera, typhoid fever, and tuberculosis, the great scourges of humanity, rapidly came under control in the industrialized countries.

	Time line for the history of public health and epidemiology.
Antiquity	Concepts of health closely tied to religion (e.g., Old Testament)
	Greek writers draw links to environmental factors (e.g., Hippocrates)
	Romans associate plumbism with wine from lead-glazed pottery
1334	Petrarch introduces the concept of comparison and indeed of a clinical trial
1603	John Graunt – Bills of Mortality and the "law of mortality". The first life table, giving the probability of dying at each age.
1700	Bernadino Ramazzini – "father of occupational epidemiology"; also breast cancer in nuns
1706-1777	Francois Bossier de Lacroix (known as Sauvages) – systematic classification of diseases (*Nosologia Methodica*)
1747	James Lind – scurvy experiment
1775	Percival Pott – scrotum cancer findings
1798	Edward Jenner – cowpox vaccination against smallpox
1787-1872	Pierre Charles Alexandre Louis (1787-1872) – the "Father of Epidemiology", *La methode numerique*
	LaPlace, Poisson – the birth of statistics
1834	William Farr, William Guy, William Budd (all students of Louis) – founded the Statistical Society of London
1847	Ignaz Semmelweiss (Vienna) – discovers transmission and prevention of puerperal fever
1849	John Snow – waterborne transmission of cholera
1850	Epidemiological Society of London established
1851	John Grove – *On the nature of epidemics* (presented the germ theory)
	Oliver Wendell Holmes and George Shattuck, Jr. (and Shattuck's student, Edward Jarvis) – founded the American Statistical Society
1870	Beginning of the era of bacteriology
1887	The Hygienic Laboratory, forerunner of the National Institutes of Health, is created within the Marine Hospital Service in Staten Island, NY
1900	Yule – notion of spurious (i.e., nonsubstantive) correlations, confounding ("Simpson's paradox")
1914-1918	Joseph Goldberger studies pellagra
1920	Split between U.S. organized medicine and physicians interested in public health (the latter were interested in national health insurance; public health concern vs. individual concern)
1937	Austin Bradford Hill, *Principles of Medical Statistics*
1942	Office of Malaria Control in War Areas (in US; became Communicable Disease Center (CDC) in 1946, Center for Disease Control in 1970, Centers for Disease Control in 1980, and Centers for Disease Control and Prevention in 1992)
1948	World Health Organization (WHO)
1948	John Ryle becomes first chairman of social medicine at Oxford. Observed that physicians have curiously little concern with prevention.
1950's-1970's	Epidemiology successes – fluoride, tobacco, blood pressure and stroke, CHD risk factors, toxic shock syndrome, Legionnaire's disease, Reye's syndrome, endometrial cancer and exogenous estrogens
1975	Lalonde Report (Canada)
1979	*Healthy People U.S.* and *Health Objectives for the Nation*

1988	U.S. Institute of Medicine *Report of the Committee for the Study of the Future of Public Health* – Public health system is in "disarray" – AIDS, injuries, teen pregnancies, Alzheimer's disease

Rise of epidemiology

Epidemiology was at the core of many of the studies that led to the above advances and to subsequent ones. But until well into the 20th century, epidemiology was not a distinct profession and/or practice, so it is not meaningful to say when its contributions began. The studies that led to the Chadwick and Shattuck reports drew on concepts that had arisen during earlier centuries, including the use of quantitative reasoning, the idea of comparing groups or populations, the collection of vital statistics, and methods of analysis (e.g., the life table).

The birth of modern epidemiology occurred during the 19th century. According to David Morens (*Epidemiology Monitor*, February 1999: 4), epidemic investigations prior to the middle of that century were mostly descriptive, rather than etiologic in orientation. Peter Panum, however, investigated the 1846 measles outbreak on the Faroe Islands "much the way an Epidemic Intelligence Service Officer at CDC would today". The classic investigations on the transmission of cholera (John Snow), typhoid fever (William Budd), and puerperal fever (Ignaz Semmelweis) led to understanding and the ability to reduce the spread of major infections. John Grove presented the germ theory in his 1851 treatise *On the nature of epidemics.*

Pierre Charles Alexandre Louis (1787-1872), sometimes called the "Father of Epidemiology", systematized the application of numerical thinking ("*la methode numerique*") and championed its cause. Using quantitative reasoning, he demonstrated that bloodletting was not efficacious therapy, and wrote books on tuberculosis and typhoid. Louis' influence was widespread, primarily through his students. (An interesting historical observation is that Louis was of lower class background; absent the French Revolution, he would probably not have had the opportunity to contribute to science and medicine.)

Many of Louis' students became leading exponents of and contributors to epidemiology. William Farr pioneered the use of statistics in epidemiology and introduced the concepts of the death rate, dose-response, herd immunity, and cohort effect. He also showed that prevalence equals incidence times duration and the need for large numbers to demonstrate associations. He and two other students of Louis (William Guy and William Budd) founded the Statistical Society of London. William Guy studied tuberculosis in relation to occupation and, I believe, conceptualized the odds ratio – the method for estimating relative risk from case-control data. Two other of Louis' students, Oliver Wendell Holmes and George Shattuck, Jr. (and Shattuck's student, Edward Jarvis) founded the American Statistical Society (see genealogy table in Lilienfeld and Lilienfeld, 2nd ed., Fig. 2-1).

Epidemiology continued to grow and develop, particularly in Britain and America. In addition to the continuing challenges from urban crowding and large-scale immigration, the revolution in bacteriology had great applicability for military forces, for which infection and disease were major threats to effectiveness. Thus, 20th century combat brought epidemiologists into the war effort. The Hygienic Laboratory (the forerunner of the National Institutes of Health, originally established

as a one-room bacteriology laboratory in an attic of the Marine Hospital Service in Staten Island, NY) provided laboratory support for the U.S. military during the Spanish-American War (Winkelstein, 2000). The U.S. Army Medical Corps and its British counterpart played major roles in preserving the health of the troops in several wars.

The relationship of epidemiology to war has been a reciprocal one. The U.S. Centers for Disease Control and Prevention (CDC) was born as the World War II Office of Malaria Control in War Areas, becoming the Communicable Disease Center in 1946, the Center for Disease Control in 1970, the Centers for Disease Control in 1980, and receiving its present name in 1992. The CDC's Epidemic Intelligence Service was established in response to concern about importation of exotic diseases from Asia, a concern arising during the Korean War. In the second half of the 20th century, epidemiology flourished, with the creation of departments of epidemiology in many universities and corporations, dramatic expansion of research (and funding for biomedical research in general), broadening of methodological and technological capabilities, growth of professional societies and journals, and coverage of epidemiology in the mass media.

Threads in the fabric of the development of epidemiology

Quantitative reasoning
Comparative studies – comparison of groups or populations
Vital statistics system
Hygienic and public health movement
Improvements in diagnosis and classification
Statistics
Computers
Personal computers
User-friendly statistical software
Biotechnology revolution
Genomics

The importance of context

Public health advocates often accuse medicine of being reactive, since physicians treat disease after it occurs whereas public health professionals work to prevent disease. Interestingly, though, advances in public health knowledge and practice occur typically as reactions to public health problems. A century and a half ago, for example, cholera epidemics in London stimulated the public health movement and the development of the London Epidemiological Society. During the past two decades, the emergence and re-emergence of major infectious pathogens (HIV, TB) have stimulated the resurgence of infectious disease epidemiology, which as recently as the 1970's seemed to be on the road to extinction, as well as to an enormous expansion in other types of research directed at infectious disease.

Wars are also a very important factor in public health, devastating to public health and public health programs in populations that suffer attack and engines of advances in public health knowledge in countries whose homeland remains undamaged. Improved treatment of wounds (Britain) and the purification, testing, and manufacture of penicillin (Britain and the U.S.) are only two of the many advances stimulated by military exigencies. Apart from military motives, the growth of government is responsible for public health advances for other reasons when there are supportive attitudes about what government should do. For example, the French Revolution and the growth of populist thinking in Europe were strong stimuli to interest in public health.

Scientific progress is fundamental to public health advances, of course, since regardless of what people think that government *should* do, what it *can* do is constrained by available knowledge and technology. What government can do is also constrained by attitudes and beliefs about what is proper. Former Surgeon General Koop has related how, during a 1940's radio program to talk about his studies of childhood cancer, he was told that he could not say the word "cancer" (it was to be referred to as "that dread disease"). Progress in preventing HIV and sexually transmitted diseases has had to contend with legal and extra-legal restrictions on open discussion about sex and particularly about anal sex.

These are only a few of the myriad influences on the evolution of public health and epidemiology. Further examples of these influences, most of which affect each other as well as public health, are:

Changing demography, economics, transportation, commerce, technology, organizations, politics, wars –
The entire health care delivery system has been transformed through the rise of managed care organizations.

Changing diseases and afflictions through the centuries –
Hunger, violence, infections, malnutrition, reproductive disorders, chronic diseases, environmental and occupational diseases, violence and injury, health care and pharmaceuticals, mental health, aging – different disease patterns dominate at different times, as the conditions of life change

Developing scientific knowledge and technology changes understanding of disease and approaches to studying it –
Introduction of Pap smear in 1940s led to knowledge of natural history of cervical cancer. Development of coronary angiography enabled visualizing of atherosclerosis during life as well as coronary artery spasm. Consider the impact of the development of microscopy, the stethoscope, electrocardiograms, culture techniques, biochemistry, cytology, computers, angiography, radioimmunoassay, DNA probes, …

Expanding social and political consciousness –
Hygienic movement, Marxism, social democracy, health promotion movement, minority health. Increased demand for (and on) epidemiology and public health (e.g., the Lalonde Report).

Expanding social organization and investment in public health resources increases the opportunities for epidemiologic research and application –

- Hospitals
- Vital statistics systems

- Health surveys
- Research funding
- Disease registries
- Insurance systems
- Record systems, computerized databases

The challenge of hindsight

In order to grasp the significance of the evolution of ideas, we need to put ourselves in the mindset of the time and appreciate the imagination (and deviance) necessary to see things in a new way. Many of the problems faced by past investigators seem so manageable compared to the ones we face today. But how did those problems look without the benefit of the knowledge and concepts that we take for granted.

Induction and latency

Consider the example of the incubation period. In infectious diseases, there is commonly an incubation period, typically on the order of 1-14 days. Until this phenomenon became known and accepted, it must have been difficult to make the connection between the onset of an illness and an exposure some two weeks earlier. Panum helped to document this phenomenon, and his studies of measles onset and previous exposure to cases are a classic of careful description and inference. With chronic diseases, the "incubation period" is much longer. Pellagra develops over a period of several months. Atherosclerotic heart disease and cancer can take 5, 10, 20, or even 30 years. Lengthy separation of cause and effect is certainly much more formidable than the 2 weeks involved in measles, but is it more formidable in terms of the level of knowledge then and now?

Rarity of disease

Rarity of a disease is in some respects an advantage for studying it and in some respects an obstacle. Epidemics are easy to study in the sense that each occurrence represents a form of natural experiment. They provide contrasts between the before and the after (e.g., arrival of a ship to the Faroe Islands, arrival of a person with typhoid fever in a previously unaffected village). With an endemic disease, on the other hand, there is no obvious contrast to stimulate perception of new events or new modes of living that could have introduced the disease. On the other hand, very rare diseases are difficult to study because of the difficulty of assembling enough cases.

Thoroughness of methods

Some famous investigators are recognized as such for advances in the methodology of their studies – advances in rigor, exquisite thoroughness, and painstaking attention to detail – before such methods were in common use. We now take it for granted, and grant proposal reviews enforce, that an investigator will conduct a systematic review of existing evidence, make use of vital statistics data, formulate precise definitions of disease and other variables, collect data in an even-handed manner, employ checks of reliability and validity of the data, and analyze the data with due attention to

alternative explanations of the findings. But each of these and other desirable methodologic practices had to be introduced at a time when it was not common practice. A common theme in the "classics" is that each investigation involved careful, systematic and detailed observation – "shoe leather" epidemiology. Not all of the practice of epidemiology is as glorious as the celebrated insights.

Disease prevention

The classic studies also gave rise to health promotion/disease prevention recommendations involving sanitary practices, personal hygiene, and diet – even before the identification of the actual etiologic or preventive agent. But is there a lesson in the observation that the dietary changes recommended by Goldberger for prevention of pellagra – increased intake of meat and dairy products – is in some respects the reverse of current recommendations for the prevention of cancer and CHD? It is also interesting to contrast these diseases and the interventions they recommended with those for contemporary epidemics (CHD, lung cancer, motor vehicle injuries, handgun fatalities). Do you suppose the public reacts differently to being told to eat *less* meat than it did to being told to eat *more* meat?

Insight based on but not constrained by knowledge

Enduring recognition over time comes from distinctive accomplishment, from achievement beyond the expected. One mark of distinction is the attainment of insight that builds on existing knowledge but is not unduly constrained by it. Scientific advances generally build on knowledge that has been successively accumulated by many people over many years. But such knowledge is understood in terms of existing paradigms (see Thomas Kuhn, *The structure of scientific revolutions*). If the existing paradigm or theoretical structure that governs the interpretation of observations is inadequate to the problem at hand, then progress demands a new or modified paradigm.

Almost by definition, a great step forward in thinking occurs in advance of general understanding. Avogadro's theory that the number of molecules in a gas is a function of its volume took 50 years to become accepted. X-rays were originally regarded as an elaborate hoax (Kuhn, 1970). In a number of the epidemiologic classics, the prevailing theories were misleading. A key contribution was the discarding of certain beliefs of the time, and the investigator had to contend with active opposition to his investigations.

According to David Morens (*Epidemiology Monitor*, February 1999: 4), when Panum's 1847 work on measles appeared in French several years later, an unsigned review of his work in the *British and Foreign Medico-Chirurgical Review* observed "'It is seldom, indeed, that an opportunity like that here described is afforded to a prudent and able man of science, who, like our author, rejecting all previously conceived opinions, diligently investigates the truth for himself.'" Joseph Goldberger, in his studies of pellagra about 60 years later also had to depart from the accepted wisdom of the time. Not long before he began his work, a 1914 commission had concluded that pellagra was an infectious and/or hereditary disease. Goldberger's careful study of all the facts enabled him to deduce that pellagra was not, in fact, a communicable disease. This study took him three months. It then took him several years, including some outlandish (heroic?) experiments in order to convince

his scientific peers of the correctness of his deductions. In Goldberger's case, others had known the pertinent facts, but their import had not been grasped.

William Farr fought the idea that cholera was spread by germs because in his data high altitude was associated with cholera, consistent with theories about atmospheric pressure and miasmas. Lind's discoveries were not adopted by the British Navy for a full 40 years, and Percival Pott's discovery about how to prevent scrotal cancer, though quickly adopted in Denmark, was not adopted in England for nearly a century. The classic papers on lung cancer and tobacco smoke, published in the *Journal of the American Medical Association* by Wynder and Graham and Doll and Hill, were almost rejected by the editor because of the lack of existing knowledge supporting the association. Despite numerous studies yielding similar findings, eminent statisticians (R.A. Fisher, Berkson) remained highly skeptical for many years.

> "Truth is the daughter of Time and not of authority." Leonardo, Bacon, Baglivi
>
> "It is the customary fate of new truths to begin as heresies and to end as superstitions." Huxley

The study of history broadens our vision and suggests that for us to rise above the common wisdom of our time we may have to accept the discomfort that comes with deviating from the conventional. For example, if an epidemiologist were to suggest that psychiatric disorders are spread by transmission of thoughts, this suggestion would be ridiculed. Was the suggestion that water was a vehicle of transmission of cholera and typhoid similarly regarded in the last century? What about the transmission of measles virus through air? Can we achieve the acuity of hindsight without the wait?

Conceptual and philosophic basis for epidemiologic advances – changing paradigms

Humors in the body
Miasma (17th century)
Contagium vivum
Concept of specificity of disease and causal agent
Multicausality
Molecular and genetic
Biotechnology

Bibliography

NOTE: The EPID 168 web site (www.sph.unc.edu/courses/epid168/) has links to two web sites devoted to the life, times, and studies of John Snow and to sites on the history of medicine at the National Library of Medicine and the University of Alabama at Birmingham School of Medicine.

Brockington, C. Fraser. The history of public health. In: W. Hobson (ed), *Theory and practice of public health.* NY, Oxford, 1979, 1-8.

Committee for the Study of the Future of Public Health. *The future of public health*. Washington, DC, National Academy Press, 1988.

Comstock, George W. Tuberculosis – a bridge to chronic disease epidemiology. *Am J Epidemiol* 1986; 124:1-16.

Dawber, Thomas R.: The Framingham Study. Cambridge, Mass.: Harvard, 1980. (Chapters 1-5 provide an interesting and easy-to-read introduction to this landmark investigation, requiring no prior background in epidemiology or cardiovascular disease. These chapters provide a good introduction to basic epidemiologic concepts.)

Dubos, Rene. *Man adapting.* New Haven, CT, Yale, 1965.

Dubos, Rene. *The mirage of health*. NY, Anchor, 1961

Epidemiology Monitor. The future of epidemiology. *Epidemiology Monitor* 1999 (February); 21(2)

Elmore, Joann G. and Alvan R. Feinstein. Joseph Goldberger: an unsung hero of American clinical epidemiology. *Ann. Intern. Med*. 1994;121:372-375.

Fee, Elizabeth. Improving the people's health: some Hopkins' contributions. *Am J Epidemiol* 1991;134:1014-1022.

Goldberger, J.: *Goldberger on pellagra.* M. Terris, ed. Baton Rouge, Louisiana State University Press, 1964.

Goldberger, J, Wheeler GA, Sydenstricker E: A study of the diet of nonpellagrous and of pellagrous households. *JAMA* 71:944, 1918.

Koplan, Jeffrey P.; Stephen B. Thacker, Nicole A. Lezin. Epidemiology in the 21st century: calculation, communication, and intervention. *Am J Public Health* 1999; 89:1153-1155.

Krause, Richard M. The origin of plagues: old and new. *Science* 1992;257:1073-1078.

Kuhn, Thomas S. *The structure of scientific revolutions*, 2nd ed, Chicago, University of Chicago, 1970.

Larson, Elaine. Innovations in health care: antisepsis as a case study. *Am J Public Health* 1989; 79:92-99.

Lilienfeld and Lilienfeld. *Foundations of Epidemiology.* Chapter 2.

Lilienfeld, Abraham M. Ceteris Paribus: the evolution of the clinical trial. *Bulletin of the History of Medicine* 1982 (Spring); 56:1-18.

Lilienfeld, A.M. (ed) *Times, Places, and Persons.* Baltimore: The Johns Hopkins University Press, 1978.

Lind, J.: *A treatise of the scurvy.* Edinburgh, Sands, Murray, and Cochran for A. Millar, 1753.

McNeill, William H. *Plagues and peoples.* Garden City, NY, 1976

Pott, P.: Chirurgical observations related to the cataract, the polypus of the nose, the cancer of the scrotum, the different kinds of ruptures and the mortification of the toes and feet. London, England, Hawes, Clarke, and Collins, 1775. In: National Cancer Institute Monograph No. 10, 1962, pp. 7-13.

Rosen, George. *A history of public health.* Baltimore, Johns Hopkins Univ, 1993.

Rosenberg, Charles E. *The Cholera Years.* Chicago, University of Chicago.

Sartwell, Philip E. and Frances Stark. American Journal of Epidemiology: its evolution since 1965. *Am J Epidemiol* 1991 (Nov 15); 134(10):1041-1046.

Silverman, W.A.: *Retrolental Fibroplasia, A Modern Parable.* New York: Grune & Stratton, 1980.

Snow, John. *On the mode of communication of cholera*. In: Snow on Cholera. New York, The Commonwealth Fund, 1936.

Susser, Mervyn. Epidemiology in the United States after World War II: the evolution of technique. *Epidemiologic Reviews* 1985; 7:147-177.

Terris, Milton. The Society for Epidemiologic Research (SER) and the future of epidemiology. *Am J Epidemiol* 1992;136:909-915.

Vandenbroucke JP, Rooda HME, Beukers H. Who made John Snow a hero? *Am J Epidemiol* 1991 (May 15); 133(10):967-973? See also letters in *Am J Epidemiol* 1992 (Feb 15);135(4):450.

Wain, Harry. *A history of preventive medicine*. Springfield, IL, 1970.

Wilcocks, Charles. *Medical advance, public health and social evolution.* NY, Pergamon, 1965

Winkelstein, Jr., Warren. Interface of epidemiology and history: a commentary on past, present, and future. *Epidemiologic Reviews* 2000; 22:2-6.

3. Studying populations – basic demography

Some basic concepts and techniques from demography - population growth, population characteristics, measures of mortality and fertility, life tables, cohort effects.

The "demi" in epidemiology

Since the primary subject matter of epidemiology is people (except for veterinary epidemiologists, who apply the same concepts and methods to studying other animal populations), a logical place to begin the study of epidemiology is with some basic concepts of demography.

*Population growth – an epidemic of homo sapiens**

For its first few million years, the species that we refer to as *homo sapiens* numbered probably fewer than 10 million, due to high mortality. In about 8000 B.C., with the beginning of agriculture, significant population growth began, bringing world population to about 500 million over a 6000-year period. At that point (1650 AD), growth accelerated sharply, so that world population doubled in 150 years (1 billion in 1800), doubled again in 130 years (1930), and doubled yet again in 45 years (4 billion in 1975). Every decade the world's population increases by about 1 billion, mostly in the developing countries. The population will reach 6 billion in early 1999. It is projected to reach 9.5 billion by 2030 and 12.6 billion by 2100.

World Population in mid-1997 (millions)

Region	Population
Asia	3,552
Africa	743
Europe	729
Latin America & Caribbean	490
North America	298
Oceania (Australia, NZ, and Pacific)	29
World	5,840

(does not add due to rounding)

* *Note about sources*: Much of the following has been drawn from publications by the Population Reference Bureau (PRB), especially "Population: A lively introduction" and "The future of world population" (see bibliography). This table comes from their 1997 World population data sheet. The PRB web site (www.prb.org) has a wealth of data and links to sources of information on population- and health-related topics.

In 1997, 86 million more people lived on planet Earth than the previous year, for an estimated annual world population growth rate of 1.47%. At that rate, world population would double in 47 years. The world population growth rate is the difference between the birth rate of 24 per 1,000 people and the death rate of 9.

Over the time, differing growth rates can dramatically alter the age, geographic, racial, and affluence distribution of the world's population. In 1950, two thirds of the world's population lived in what is usually referred to as the developing world. The proportion was three-quarters in 1990 and is projected to grow to 85% by 2025 and 90% by 2100. Thus, whatever improvements in health take place in the industrialized world, world demographic and health indicators will be primarily influenced by the situation in the developing world.

The Demographic Transition

A fundamental model developed to describe population dynamics is the Demographic Transition model. The model posits four stages in the evolution of the population in a society.

1. High fertility, high mortality (pre-industrial)
2. High fertility, declining mortality (industrializing)
3. Declining fertility, low mortality
4. Low fertility, low mortality (stable population)

The first stage (pre-industrial) prevailed throughout the world prior to the past few centuries. Rapid population growth takes place in Stages 2 and 3, because high birth rates, necessary for population survival in Stage 1, are embedded in the cultural, religious, economic, and political fabric of pre-modern societies. As economic and public health advances decrease mortality rates, rapid population growth occurs until the society adjusts to the new realities and fertility decline.

The Demographic Transition Model was constructed from the European experience, in which the decline in death rates was gradual. It remains to be seen how this model will play out in the developing world of today, in which the decline in death rates has occurred much more rapidly and in which social change takes place against a backdrop of and in interaction with the post-industrial world of electronic communications, multi-national production and marketing, and international travel. There is some evidence that the model will also apply to the developing world of today. But the timetable for completion of the demographic transition in the developing world will determine the ultimate size of the world's population.

Demographic balancing equation

If birth and death are the two most fundamental demographic processes, migration is probably the third. The size of the world's population is (at least at present) completely determined by birth and death rates, but the population in any particular region or locale is also determined by net migration. These three processes are expressed in the demographic balancing equation—the increase (or decrease) in a population as the algebraic sum of births, deaths, immigration, and emigration. The following table gives the equation for the world and for the U.S. for 1991.

The demographic balancing equation for the United States (from McFalls, 1991) (numbers in thousands)

		Starting population	+	(Births – Deaths)	+	(Immigration – Emigration)	=	Ending population
		Starting population	+	(Natural increase)	+	(Net migration)	=	Ending population
World	=	5,245,071	+	(142,959 – 50,418)				
	=	5,245,071	+	92,541			=	5,337,612
U.S.	=	248,168	+	(4,179 – 2,162)	+	(853 – 160)		
	=	248,168	+	2,107	+	693	=	250,878

In recent decades, on a world basis, the migration has perhaps had its greatest impact on urbanization. In the forty years from 1950 to 1990, the urban population in the countries of the Third World increased over five-fold, from 286 million to about 1.5 billion. About 40 percent of this growth resulted from rural to urban migration. The U.N. predicts that by the year 2000 there will be 19 Third World cities with populations over 10 million. In contrast to Tokyo, Los Angeles, New York, London, and other glamorous metropolises, overcrowded urban areas in poor countries are characterized by inadequate housing, sanitation, transportation, employment opportunities, and other essentials of healthy living, the ingredients for misery and the spread of microorganisms.

Population age structure and the population pyramid

For every 10 people in the world:

3 are younger than 15 years of age

4 live in an urban area

6 live in Asia (2 in China, 1 in India)

8 live in developing countries

An important dynamic in population growth is the reciprocal relationship between the rate of natural increase (births - deaths) and the age structure of the population. The latter is one of the strongest influences on the growth rate of a population, since both fertility and mortality vary greatly by age. A younger population has a higher rate of natural increase; a high rate of natural increase in turn lowers the median age of a population.

In Africa, which has the highest birth (40/1,000) and growth (2.6%) rates, only 56% of the population are older than 15 years. In contrast, in Europe, where average birth rates have been close to replacement level for many years, four-fifths of the population (81%) are older than 15 years. In fact, Europe as a whole experienced overall negative growth in 1997, due to birth and death rates of 10 and 14, respectively, in Eastern Europe (including Russia). Since nearly all (96%)

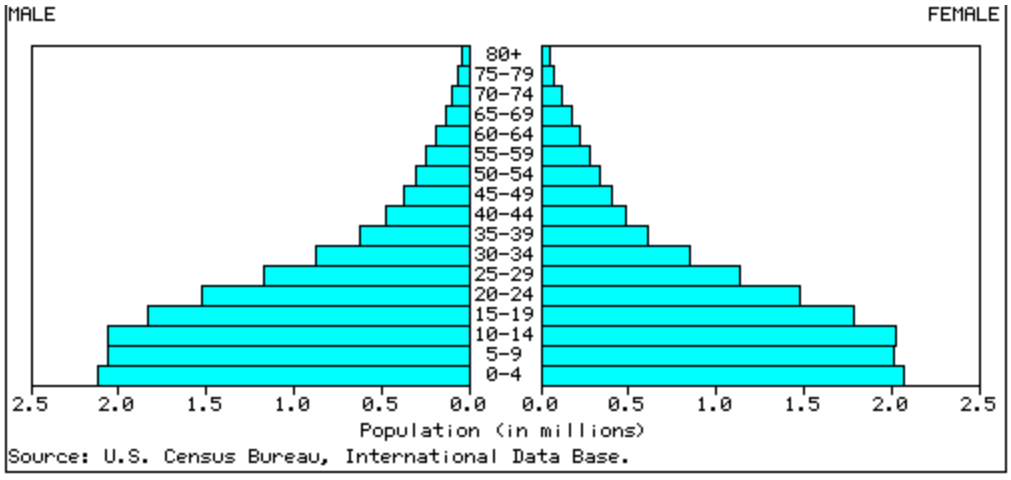
of the increase in the world's population takes place in the developing world, the developing countries are becoming younger while the wealthier countries are becoming older.

Nevertheless, fertility control is increasing in the developing world. As it does, the age structure of the population shifts upwards, since the larger birth cohorts of previous years are followed by relatively smaller birth cohorts. The average age of the world's population, around 28 years, is projected to increase to 31-35 years, so that the proportion of persons 60 years and older will grow from about 9% to 13-17% (Lutz, 1994). This proportion will range from as low as 5% in sub-Saharan Africa to much as 30% in Western Europe. In China, where fertility has been successfully regulated for decades, the proportion of the population age 60 and older will rise to about 20% in 2030 (Lutz, 1994).

The population pyramid

Demographers display the age structure of a population by constructing a graph in which the population size in each age band is depicted by a horizontal bar that extends from a centerline to the left for one gender and to the right for the other, with the age bands arranged from lowest (at the horizontal axis) to highest. A population pyramid for a population that is growing rapidly, e.g. Kenya, resembles a pillar that is very broad at the base (age 0-1 years) and tapers continuously to a point at the top. In contrast, the population pyramid for a zero-growth population, e.g. Denmark, resembles a bowling pin, with a broader bottom and middle, and narrower base and top.

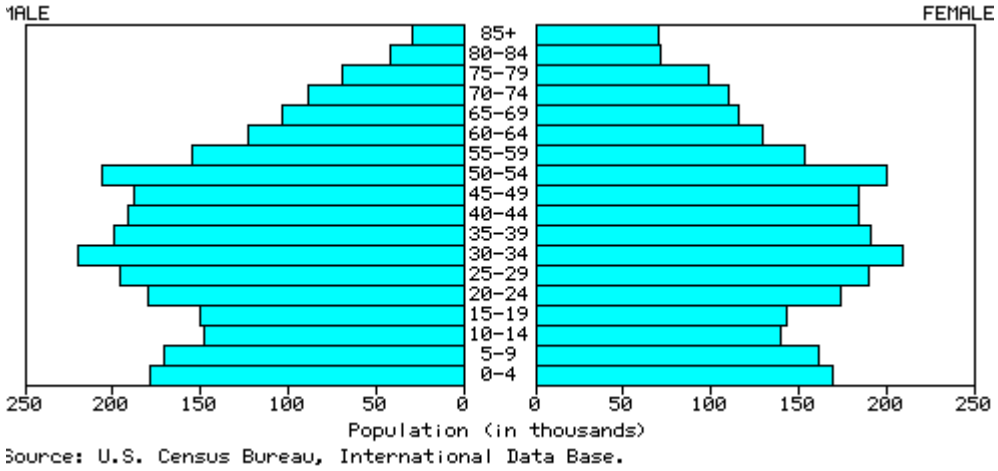
Kenya, 1998



The population pyramid for a country shows the pattern of birth and death rates over the past decades, since apart from immigration and emigration, the maximum size of any age group is set by the birth cohort that it began as, and its actual size shows its subsequent mortality experience. For example, the 1989 population pyramid for Germany shows the deficit of older males resulting from losses in World Wars I and II and narrowings corresponding to the markedly lower wartime birth

rates. Similarly, bulges in the reproductive years often produce bulges at the bottom, since more women of reproductive age usually translates into more births.

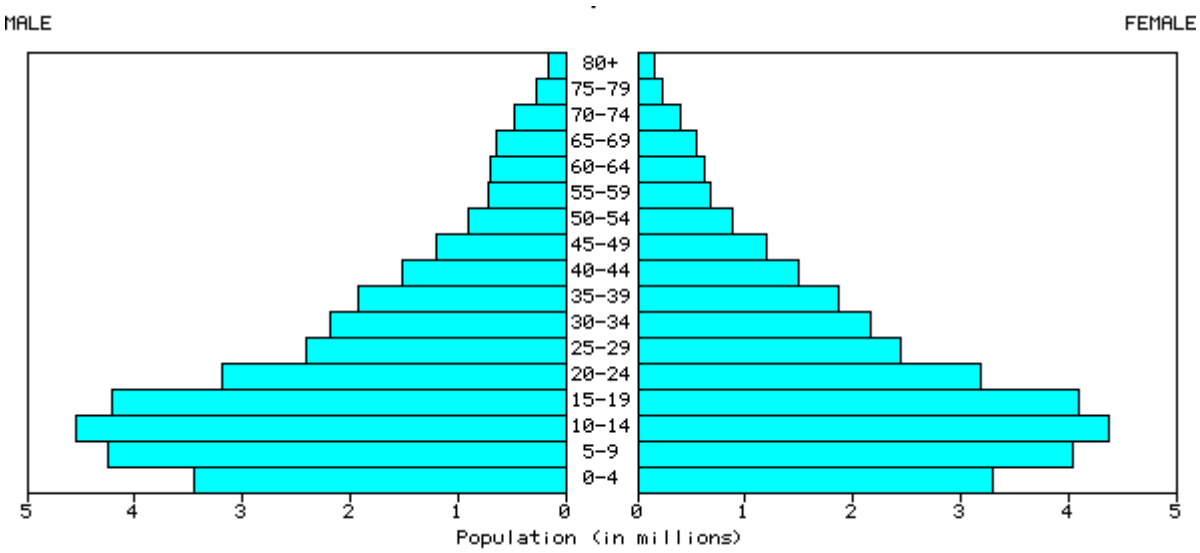
Denmark, 1998 (note change in scale)



Using the pyramid it is easy to see how a growing population becomes younger and the transition to lower fertility makes it older. Widespread family planning makes new birth cohorts smaller, so that the pyramid consists of a broad middle (persons born before the adoption of family planning) being pushed upward by a narrower base. Initially this age distribution makes life easier for adults, especially women, since effort and resources for childrearing and support are proportionally lower. However, when the adults who first adopted family planning reach retirement age, there are fewer younger people available to support them. Unless productivity and savings have risen sufficiently, the society will be hard pressed to support its elderly members—an issue of concern in affluent societies today.

The population pyramid for Iran has a number of distinctive features. Iran embraced family planning in the 1960's, one of the first developing countries to do so. The Islamic revolution of 1979, however, regarded the family planning program as "pro-West" and dismantled the program. Moreover, with the war with Iraq made population growth seem advantageous. When the war ended and reconstruction became the priority, the government reversed its policy and inaugurated a new family planning program with an extensive information campaign and, in 1993, powerful economic disincentives for having more than three children. These measures reduced the total fertility rate (see below) from 5.2 children in 1989 to 2.6 children in 1997. (This account is taken from Farzaneh Roudi, *Population Today*, July/August1999). The jump in the birth rate following the revolution can be seen in the large size of the 15-19 year-old band (born 1979-1983) compared to the next older one; the subsequent curtailment of births shows up as a relatively small number of children 4 years old and younger. (Note: these population pyramids come from the U.S. Bureau for the Census International Database and were downloaded from the Population Bureau web site.)

Iran, 1998



Source: U.S. Census Bureau, International Data Base.

Influence of population age composition

Since rates of most diseases and medical conditions, injuries, and health-related phenomena such as violence vary greatly by age, a population's age structure affects much more than its growth rate. As the 76 million "Boomers" in the post-World War II baby boom cohort to which President Bill Clinton belongs have moved up through the population pyramid, as a pig which has been swallowed by a python, they expanded school and university enrollments, created an employment boom first in obstetrics, pediatrics, construction, urban planning, and diaper services, and subsequently increased demand for baby clothes, toys, appliances, teachers, school buildings, faculty, managers, automobile dealers, managers, health professionals, and investment counselors.

But in their wake, the Boomers have faced the contraction of many of those employment opportunities as their larger numbers and the smaller job-creating needs of the following generation increased competition at every stage. On the horizon are substantial increases in the need for geriatricians and retirement facilities, providing more employment opportunity for the generations that follow the Boomers but also a heavier burden for taxes and elder-care. A baby "boomlet" is also moving up the pyramid, as the Boomers' children create an "echo" of the baby boom.

The baby boom is a key contributor to the projected shortfalls in funding for Social Security, Medicare, and pensions in the coming decades. The following projections were made several years ago but are still relevant:

When the baby boom cohort retires

	1995	2030
Retired population (%)	12	20
Workers per retired person	3.4	2.0
Combined Social Security and Medicare tax rate per worker (including employer's share)	15%	28%

(Source: Who will pay for your retirement? The looming crisis. Center for Economic Development, NY, NY. Summarized in TIAA-CREF quarterly newsletter the Participant, November 1995: 3-5.)

The U.S. is the fasted growing industrialized country, with a 1% growth rate (about 30% of which is due to immigration). Providing for the needs of senior citizens will be even more difficult in Europe, where most countries are now close to zero population growth and already 14% of the population are age 65 years or older. It has been projected that in 100 years there will be only half as many Europeans as today, which for many raises concerns about economic health, military strength, and cultural identity.

Sex composition

Another fundamental demographic characteristic of a population is its sex ratio (generally expressed as the number of males per 100 females). A strongly unbalanced sex ratio affects the availability of marriage partners, family stability, and many aspects of the social, psychological, and economic structure of a society.

Sex ratios are affected by events such as major wars and large-scale migration, by cultural pressures that favor one sex, usually males, by unequal mortality rates in adulthood, and by changes in the birth rate. Because of higher male mortality rates, the sex ratio in the U.S. at birth falls from about 106 at birth, to about 100 by ages 25-29, and to 39 for ages 85 and above. Migration in search of employment is a frequent cause of a sex ratio far from 100. For example, oil field employment in the United Arab Emirates has brought that country's sex ratio as high as 218.

Although changes in birth rates do not alter sex ratios themselves, if women usually marry older men, a marked increase or decrease in the birth rate will produce an unbalanced sex ratio for potential mates. Girls born at and after a marked increase in the birth rate will encounter a deficit of mates in the cohort born before birth rates increased; boys born before a marked decrease will encounter a deficit of younger women. The substantial declines in birth rates in Eastern Europe following the collapse of Communism may lead to a difficult situation for men born before the collapse. In the U.S., casualties from urban poverty and the War on Drugs have created a deficit of marriageable men, particularly African American men. Because of assortive mating and the legacy of American apartheid, the effects of the deficit are concentrated in African American communities,

with many African American forced to choose between raising a family by themselves or remaining childless.

Women's status in society is a key factor in relation to the sex ratio and fertility in general. For example, women's opportunities for education and employment are strongly and reciprocally related to the birth rate. In China, where a "one-child" policy for urban families was adopted as a dramatic step toward curbing growth in its huge population, the sex ratio at birth is now 114 (a normal ratio is 105 boys for 100 girls). The approximately 12% shortfall of girls arises from families' desire for a mail offspring and is believed to be due to a combination of sex-selective abortion, abandonment, infanticide, and underreporting (Nancy E. Riley, China's "Missing girls": prospects and policy. Population Today. February 1996;24:4-5).

Racial, ethnic, and religious composition

Race (a classification generally based on physical characteristics) and ethnicity (generally defined in relation to cultural characteristics), though very difficult to define scientifically, have been and continue to be very strong, even dominant factors, in many aspects of many societies. Thus, the racial, ethnic, and religious composition of a population is linked with many other population characteristics, as a function of the beliefs, values, and practices of the various groups and of the way societies regard and treat them. While Americans are most conscious of racial and ethnic issues in relation to African Americans, Latinos, and Native Americans/American Indians, and Asian conflicts related to race, ethnicity, and religion are a major phenomena throughout the world and throughout history, as the following VERY selective list recalls:

Balkans - Serbs, Croats, and Muslims (Bosnia), Serbs and Albanians (Kosovo)

Northern Ireland - Catholics and Protestants

Rwanda - Hutu's and Tutsi's

Middle East/Northern Africa - Jews, Christians, and Muslims

Iran's massacre of Bahai's

Kurds in northern Iran and Turkey

Indonesia - massacres of ethnic Chinese

East Timor

India/Pakistan - Hindus and Muslims

Europe - Christians and Jews (centuries of persecution climaxing though not ending with the Nazi's systematic extermination of over 6 million Jews, gypsies, and other peoples)

Germany - Catholics and Protestants (The Hundred Years War)

Americas - Europeans, white Americans, African Americans, and Native Americans/American Indians

The pervasiveness, strength, viciousness, and persistence of human reactions to differences in physical features, practices, beliefs, language, and other characteristics have had and will have powerful effects on public health.

Demographic concepts, measures, and techniques

The discussion above uses many demographic terms, concepts, and measures. We now give precise definitions.

The (crude) **birth rate** is the number of births during a stated period divided by population size.

The (crude) **death rate** is the number of deaths during a stated period divided by population size.

Population-based rates are usually expressed per 100, 1000, 10,000, or per 100,000 to reduce the need for decimal fractions. For example, 2,312,132 deaths were registered in the United States in 1995, yielding a (crude) death rate was 880 per 100,000 population. This rate represented a slight increase over the preceding year's rate of 874 (Source: Anderson et al., *Report of final mortality statistics*, 1995. Monthly Vital Statistics Report 45(11) suppl 2, June 12 1997, National Center for Health Statistics (CDC), http://www.cdc.gov/nchswww/data/mv4511s2.pdf). Birth rates are generally expressed per 1,000 per year. For example, the lowest birth rates in the world are about 10, in several European countries; the highest are about 50, in several African countries.

When the numerator (deaths or births) in a given calculation is small, data for several years may be averaged, so that the result is more precise (less susceptible to influence by random variability). For example, taking the average number of births over three years and dividing by the average population size during those years yields a 3-year average birth rate. The average population size may be the average of the estimated population size for the years in the interval or simply the estimated population for the middle of the period (e.g., the middle of the year for which the rate is being computed). Where the population is growing steadily (or declining steadily), the mid-year population provides a better estimate than the January 1st or December 31st population size, so the mid-year population is also used for rates computed for a single year. Typical birth and death rate formulas are:

$$\text{Birth rate} = \frac{\text{Births during year}}{\text{Mid-year population}} \times 1{,}000$$

$$\text{Death rate} = \frac{\text{Deaths during year}}{\text{Mid-year population}} \times 1{,}000$$

$$\text{5-year average death rate} = \frac{\text{Deaths during the year period}}{\text{Population estimate for the middle of the third year}} \times 1{,}000$$

Fertility and fecundity

An obvious limitation of the birth rate is that its denominator includes the total population even though many members (e.g., young children) cannot themselves contribute to births - and only women give birth. Thus, a general fertility rate is defined by including in the denominator only women of reproductive age:

$$\text{General fertility rate} = \frac{\text{Births during year}}{\text{Women of reproductive age (mid-year estimate)}} \times 1{,}000$$

Note that fertility, as distinct from fecundity, refers to actual births. Fecundity refers to the biological ability to have children (the opposite of sterility).

Disaggregating by age

A key consideration in interpreting overall birth, death, fertility, and almost any other rates is that they are strongly influenced by the population's age and sex composition structure. That fact does not make these "crude" overall rates any less real or true or useful. But failure to take into account population composition can result in confusion in comparing crude rates across populations with very different composition.

For example, the death rate in Western Europe (10) is higher than in North Africa (8). In other words, deaths are numerically more prominent in Western Europe than in North Africa. It would be a serious error, though, to interpret these rates as indicating that conditions of life and/or health care services are worse in Western Europe than in North Africa. The reason is that Western Europe would be expected to have a higher (crude) death rate because its population is, on the average, older (15% age 65 or above) than the population of North Africa (4% age 65 and above).

To enable comparisons that take into account age structure, sex composition, and other population characteristics, demographers (and epidemiologists) use specific rates (i.e., rates computed for a specific age and/or other subgroup - demographers call these "refined" rates). These specific rates can then be averaged, with some appropriate weighting, to obtain a single overall rate for comparative or descriptive purposes is desired, then an average of the specific rates is used, with some appropriate weighting. Such weighted averages are called adjusted or standardized rates (the two terms are largely synonymous). The United States age-adjusted death rate for 1995 was 503.9 per 100,000, slightly lower than the 507.4 age-adjusted death rate for 1994 (NCHS data in Anderson et al., 1997, see above). The reason that the age-adjusted death rate declined from 1994 to 1995 while the crude death rate increased is that the latter reflects the aging of the U.S. population, whereas the former is adjusted to the age distribution of a "standard" population (in this case, the U.S. population for 1940).

Total fertility rate (TFR)

Standardization of rates and ratios is the topic for a later in the course. But there is another important technique that is used to summarize age-specific rates. For fertility, the technique yields the **total fertility rate** (TFR) -- the average number of children a woman is expected to have during her reproductive life. The average number of children born to women who have passed their fecund years can, of course, be obtained simply by averaging the number of live births. In contrast, the TFR provides a projection into the future.

The TFR summarizes the fertility rate at each age by projecting the fertility experience of a cohort of women as they pass through each age band of their fecund years. For example, suppose that in a certain population in 1996 the average annual fertility rate for women age 15-19 was 110 per 1000 women), 180 for women age 20-29, and 80 for women 30 years and older. The TFR is simply the sum of the annual fertility rate for each single year of age during the fecund years. So 1000 women who begin their reproductive career at age 15 and end it at age 45 would be expected to bear:

Calculation of total fertility rate (TFR)
For 1000 women from age 15 through age 45 years

Age	Births	
15	110	
16	110	
17	110	(average annual fertility
18	110	from ages 15-19 = 110/1000)
19	110	
20	180	
21	180	
22	180	(average annual fertility
. . .		from ages 20-29 = 180/1000)
29	180	
30	80	
31	80	(average annual fertility
. . .		from ages 30-45 = 80/1000)
44	80	
45	80	
	3,630	

or about 3.6 children born to each woman.

(This TFR could also be calculated more compactly as
110 x 5 + 180 x 10 + 80 x 16 = 3,630)

Note that the TFR is a hypothetical measure based on the assumption that the age-specific fertility rates do not change until the cohort has aged beyond them. The TFR is a projection, not a prediction—essentially, a technique for summarizing a set of age-specific rates into an intuitively meaningful number.

Life expectancy

The technique, of using current data for people across a range of ages to project what will happen to a person or population who will be passing through those ages, is also the basis for a widely-cited summary measure, life expectancy. Life expectancy is the average number of years still to be lived by a group of people at birth or at some specified age. Although it pretends to foretell the future, life expectancy is essentially a way of summarizing of a set of age-specific death rates. It thus provides a convenient indicator of the level of public health in a population and also a basis for setting life insurance premiums and annuity payments.

In order to understand life expectancy and TFR's, it is important to appreciate the difference between these demographic summary measures and actual predictions. A prediction involves judgment about what will happen in the future. Life expectancy and TFR's are simply ways of presenting the current experience of a population. Thus, my prediction is most of us will live beyond our life expectancy!

The explanation for this apparent paradox is that life expectancy is a representation of age-specific death rates as they are at the present time. If age-specific death rates do not change during the rest of our lives, then our life expectancy today will be an excellent estimate of the average number of years we will live. However, how likely are today's age-specific death rates to remain constant? First, we can anticipate improvements in knowledge about health, medical care technology, and conditions of living to bring about reductions in death rates. Second, today's death rates for 40-90 year-olds represent the experience of people who were born during about 1900-1960.

Today's over-forties Americans lived through some or all of the Great Depression of the 1930s, two world wars, the Korean War, the Vietnam War, atmospheric nuclear bomb testing, unrestrained DDT use, pre-vaccine levels of mumps, polio, measles, rubella, chicken pox, pre-antibiotic levels of mycobacterium tuberculosis, syphilis, and other now-treatable diseases, varying levels of exposure to noxious environmental and workplace substances, a system of legally-enforced apartheid in much of the nation, limited availability of family planning, and lower general knowledge about health promotive practices, to list just a smattering of the events and conditions that may have affected subsequent health and mortality. Although changes in living conditions are not always for the better (death rates in Russia and some other countries of the former Soviet Union have worsened considerably since the breakup), the United States, Western Europe, Japan, and many countries in the developing world can expect that tomorrow's elderly will be healthier and longer-lived than the elderly of the previous generation. For these reasons life expectancy, computed from today's age-specific death rates, probably underestimates the average length of life remaining to those of us alive today.

Since it is a summary of a set of age-specific mortality rates, life expectancy can be computed from any particular age forward. Life expectancy at birth summarizes mortality rates across all ages. Life expectancy from age 65 summarizes mortality rates following the conventional age of retirement. Accordingly, life expectancy at birth can be greatly influenced by changes in infant mortality and child survival. The reason is that reductions in early life mortality typically add many more years of life than reductions in mortality rates for the elderly. The importance of knowing the age from which life expectancy is being computed is illustrated by the following excerpt from a column

prepared by the Social Security Administration and distributed by Knight Ridder / Tribune News Service (*Chapel Hill Herald*, June 28, 1998: 7):

> Q. *I heard that the Social Security retirement age is increasing. Is this true and if so, why?*
>
> A. Yes, it's true. When Social Security was just getting started back in 1935, the average American's life expectancy was just under age 60. Today it's more than 25 percent longer at just over 76. That means workers have more time for retirement, and more time to collect Social Security. And that's why Social Security's retirement age is gradually changing ... to keep pace with increases in longevity. A worker retiring today still needs to be age 65 to collect full benefits, but by 2027, workers will have to be age 67 for full retirement benefits.

It is certainly the case that longevity today is much greater than when the Social Security system was begun, so that it is now expected to provide support over a much larger fraction of a person's life. However, the life expectancies cited are life expectancies from birth. Although children who die obviously do not collect retirement benefits, neither do they make contributions to Social Security based on their earnings. For Social Security issues, the relevant change in life expectancy is that from age 62 or 65, when workers become eligible to receive Social Security retirement benefits. Every year's increase in life expectancy beyond retirement means an additional year of Social Security benefits. This life expectancy (now 15.7 and 18.9 years, respectively, for U.S. males and females age 65 years) has also increased greatly since 1935.

Life expectancy computation and the current life table

Life expectancy is computed by constructing a demographic life-table. A demographic life table depicts the mortality experience of a cohort (a defined group of people) over time, either as it occurs, has occurred, or would be expected to occur. Imagine a cohort of 100,000 newborns growing up and growing old. Eventually all will die, some as infants or children, but most as elderly persons. The demographic life table applies age-specific risks of death to the surviving members of the cohort as they pass through each age band. Thus, the demographic life table (also called a current life table) is a technique for showing the implications on cohort survival of a set of age-specific death rates.

Excerpt from the U.S. 1993 abridged life table (total population)

Age interval (years)	Risk Of Death	Number still alive	Deaths
x-x+n	$_nQ_x$	l_x	$_nD_x$
(A)	(B)	(C)	(D)
<= 1 yr	.00835	100,000	835
1-5	.00177	99,165	176
5-10	.00106	98,989	105
10-15	.00126	98,884	125
15-20	.00431	98,759	426
20-25	.00545	98,333	536
25-30	.00612	97,797	599
30-35	.00797	97,198	775
35-40	.01031	96,423	994
40-45	.01343	95,429	1,282
45-50	.01842	94,147	1,734
50-55	.02808	92,413	2,595
55-60	.04421	89,818	3,971
60-65	.06875	85,847	5,902
65-70	.10148	79,945	8,113
70-75	.14838	71,832	10,658
75-80	.21698	61,174	13,274
80-85	.32300	47,900	15,472
>= 85 yr	1.00000	32,428	32,428

(Source: National Center for Health Statistics)

(The algebraic symbols beneath the column headings show traditional life table notation; "x" refers to the age at the start of an interval, "n" to the number of years of the interval.)

For example, here are the first four columns of the U.S. 1993 abridged life table, from the National Center for Health Statistics world wide web site ("abridged" means that ages are grouped rather than being listed for each individual year). The table begins with a cohort of 100,000 life births (first line of column C). For each age interval (column A), the cohort members who enter the interval (column C) are subjected to the risk of dying during that age interval (column B), producing the number of deaths shown in column D and leaving the number of survivors shown in the next line of column B. Thus, in their first year of life, the 100,000 live newborns experience are subjected to a risk of death of 0.00835 (835/100,000), so that 835 die (B x C) and 99,165 survive (B - D) to enter

age interval 1-5 years. Between ages one and five, the 99,165 babies who attained age one year are subjected to a five-year risk of death of 0.00177 (177/100,000), so that 176 die (0.0017 x 99,165) and 98,989 (99,165 - 176) attain age six.

Notice that the age-specific risks of death (proportion dying, column B) increase from their lowest value at age 5-10 years, at first gradually, then increasingly steeply until during the age interval 80-85 nearly one-third of cohort survivors are expected to die. Correspondingly, the numbers in column D (deaths) also increase gradually, then more steeply—but not quite as steeply as do the risks in column B. The reason is that the actual number of deaths depends also on the number of people at risk of death (survivors, column C) which drops gradually at first, then more and more rapidly as the risks increase. Notice also the very high risk of death for infants: the 0.0085 means that 835 of 100,000 infants—nearly 1% --die during just one year. In contrast, only 177 of the surviving infants die during the following four years.

Death risks versus death rates

An important technical issue to consider at this point is that the risks in column B are not the same as the age-specific death rates discussed above, though the latter are the basis for deriving the risks in column B. There are two reasons. First, all but the first two of the values in column B show the risk for a five-year interval. Second, an (annual) death rate is an average value over an interval, based on the average population at risk for the interval, typically estimated by the mid-year population (which is why such death rates are called "central death rates"). In contrast, the risks in column B apply not to the average population or mid-year population but to the population at the start of the interval, which in a life-table is always greater than the average population size during the interval.

Assume that the death rate during an age interval remains fixed, so that the cohort experiences deaths during each month of the interval. Cohort members who die in the first months of the interval are obviously no longer at risk of dying later during the interval. A decreasing population with fixed death rates means that the number of deaths in each month of the interval also decreases. The calculation of the risk for the interval takes into account the fact that the cohort shrinks during the interval. At young ages, when age-specific death rates are small, the shrinkage is slight so the one-year risk is very close to the annual death rate and the five-year risk is very close to five times the average annual death rate. But at older ages, substantial shrinkage occurs and the risk is therefore less than the number of years times the average annual death rate.

To illustrate:

During infancy, the cohort loses 835 members, so that it shrinks from 100,000 to 99,165. The average or mid-year population, then, is approximately (.5)(100,000 + 99,165) or, equivalently, 100,000 -.5(835) = 99,582.5. This number is very close to 100,000, so it is easy to see why the death rate during the first year (835 deaths divided by 99,582.5 = 0.00839) is almost identical to the first-year risk (0.00835). Similarly, during the next four years (ages 1-5), the average annual death rate during the interval is approximately 0.000444 (176 deaths/4 years, divided by 99,077, the average population during the interval). Multiplying this rate by four years gives 0.00178, nearly identical to the four year risk (0.00177).

At the other end of the life table, cohort size loses 15,472 members, declining from 47,900 at age 80 to 32,428 at age 85. The average annual death rate is 0.07704 (15,472 / 5 years divided by the average size of the cohort, 40,164). Multiplying this rate by five years gives 0.38522, which is considerably greater than the five-year risk in column B (0.32300). We can come much closer to this five-year risk if we treat the five-year interval like a miniature life-table by dividing up the five-year interval into single years and applying the average annual death rate (0.07704) to each year of the cohort:

Age	Annual Death Rate	Proportion surviving that year	Cumulative Proportion surviving	Cumulative Proportion dying
80-81	0.07704	0.92296	0.92296	0.07704
81-82	0.07704	0.92296	0.85186	0.14814
82-83	0.07704	0.92296	0.78623	0.21377
83-84	0.07704	0.92296	0.72566	0.27434
84-85	0.07704	0.92296	0.66975	0.33025

The cumulative 5-year risk calculated from the cumulative proportion dying comes very close to the value figure in column B of the table (0.32300). If we divide each year into 12 months, or 52 weeks, or 365.25 days, the life-table-type calculation comes even closer. (Using calculus, it can be shown that in the limit, as the number of units becomes infinite and their size approaches zero, the life-table computation of the 5-year = 1 - exp(-5 x 0.07704) = 0.3197.)

Deriving life expectancies

Now we present the rest of the NCHS (abridged) U.S. 1993 life table, by including its three right-most columns.

Column E shows the sum of the number of years lived by all members of the cohort during each age interval. During a five-year interval, most cohort members will live for five years, but those who die during the interval will live fewer years. During the lowest risk five years (ages 5-10), nearly all of the 98,989 cohort members who enter the interval (column C) will live 5 years, for a total number of years of life of 494,945, which is just slightly above the value in column E. Between ages 80 and 85, however, only about two-thirds of the entering cohort live all five years, so the number in column E (201,029) is much lower than five times column C (239,500). However, if we use the average population size (40,164) to estimate years of life lived during ages 80-85, we obtain 5 x 40,164 = 200,820, which is very close to the number in Column E. (The numbers in column E also can be explained in terms of the concept of a “stationary population”.)

The next column (F) gives the sum of the number of years of life during the specific age interval and the remaining intervals. For example, the 395,851 total years of life remaining for the cohort members who attain age 80 are the sum of the 201,029 years to be lived during 80-85 plus the 194,822 years left for those who survive to age 85. The 669,345 years for cohort members reaching age 75 are the sum of the 273,494 years to be lived during the age 75-80 interval plus the 395,851 years remaining for members who reach age 80.

U.S. 1993 abridged life table (total population)
(Source: National Center for Health Statistics)

Age Interval (years)	Risk of death	Number still alive	Deaths	Years lived	Years remaining	Life expectancy
x-x+n	$_nQ_x$	l_x	$_nD_x$	$_nL_x$	T_x	
(A)	(B)	(C)	(D)	(E)	(F)	(G)
<= 1 yr	.00835	100,000	835	99,290	7,553,897	75.5
1-5	.00177	99,165	176	396,248	7,454,607	75.2
5-10	.00106	98,989	105	494,659	7,058,359	71.3
10-15	.00126	98,884	125	494,177	6,563,700	66.4
15-20	.00431	98,759	426	492,829	6,069,523	61.5
20-25	.00545	98,333	536	490,352	5,576,694	56.7
25-30	.00612	97,797	599	487,486	5,086,342	52.0
30-35	.00797	97,198	775	484,098	4,598,856	47.3
35-40	.01031	96,423	994	479,771	4,114,758	42.7
40-45	.01343	95,429	1,282	474,168	3,634,987	38.1
45-50	.01842	94,147	1,734	466,717	3,160,819	33.6
50-55	.02808	92,413	2,595	455,985	2,694,102	29.2
55-60	.04421	89,818	3,971	439,733	2,238,117	24.9
60-65	.06875	85,847	5,902	415,279	1,798,384	20.9
65-70	.10148	79,945	8,113	380,318	1,383,105	17.3
70-75	.14838	71,832	10,658	333,442	1,002,787	14.0
75-80	.21698	61,174	13,274	273,494	669,345	10.9
80-85	.32300	47,900	15,472	201,029	395,851	8.3
>= 85	1.00000	32,428	32,428	194,822	194,822	6.0

Life expectancy, then, the average number of years of life remaining after a given age, is the total years of life left (column F) divided by the number of cohort members who have attained that age (column C). Since the cohort numbers 100,000 at birth, life expectancy at birth is simply 7,553,897 / 100,000 = 75.5. The 89,818 cohort members who attain age 55 years have a total of 2,238,117 total years of life remaining, or an average of 24.9 years.

An advantage of surviving is that the average age the cohort will expect to attain keeps rising also. Fifty-year-olds have an average life expectancy of 29.2, for an expected age at death of 79.2; 70-year-olds have an average life expectancy of 14.0, for an expected age at death of 84 years. The reason, of course, is that cohort members who live shorter lives bring down the average; when they drop out the average is reduced by less than the number of years of the interval.

Cohort life tables

Because the current life table uses risks derived from current (or recent) death rates at each age, the life expectancies are simply a technique for summarizing them more meaningfully than if we took a simple average of age-specific death rates. Of course, in actual fact, age-specific death rates are likely

to change, hopefully to decline. If they do, then by the time a cohort of newborns reach age 20, they will experience not the 1993 death rates for 20-year-olds but those in effect in 2013. Similarly, they will experience the death rates for 30-year-olds in effect in 2023, for 40-year-olds in 2033, and so forth.

The cohort life table is constructed to take account of changing death rates. Of course, if such a life table is to be based on observed death rates, it can apply only to a cohort born sufficiently in the past. If, for example, we create a cohort life table for persons born in 1880, then we can use the observed death rates for the appropriate age for each year or interval beginning in 1880. Average years of life remaining at each age of a life table constructed from historical death rates summarizes the actual mortality experience of past birth cohorts. In epidemiology, cohort life tables are used much more often than current life tables, because the life table technique is often useful for analyzing data collected during the follow-up of a cohort (some authors call these follow-up life tables).

The cohort in a current or cohort life table loses members only to death, so that everyone who survives an interval is included in the next one. The cohorts studied by epidemiologists, on the other hand, can lose members who become lost to follow-up so that their vital status cannot be determined. Moreover, epidemiologists usually study outcomes other than all-cause mortality, so epidemiologic cohorts lose members who migrate or withdraw from the study or who become ineligible to have the outcome of interest (e.g., due to such reasons as death from another cause, surgical removal of an organ prior to the development of the disease of interest, or discontinuance of a medication being studied). In addition, the members of an epidemiologic cohort may not enter the cohort at the same calendar time or age.

A follow-up life table provides a way of representing and analyzing the experience of an epidemiologic cohort. In one common type of follow-up life table, people being studied are entered into the cohort on the basis of an event, such as employment, illness onset, surgery, attaining age 18, or sexual debut, and are then followed forward in time. Their time in the cohort (and in the life table) is computed with respect to their enrollment event. At each time interval following initiation of follow-up, the number of outcomes observed is analyzed in relation to the cohort members whose status is observed for all or part of the interval. Where the precise time of follow-up for each cohort member is unknown, then some intermediate number is used, in analogy to the use of the mid-year population for a central death rate.

Cohort effects

The life table and the TFR are both based on the concept of a cohort proceeding through time, and both employ the assumption that age-specific rates remain constant. In actuality, of course, age-specific rates do change over secular time, and populations are composed of many cohorts, not only one. Since age, secular time, and cohort are fundamentally tied to one another - as time advances, cohorts age - it can be difficult to ascertain whether an association with one of these aspects of time reflects the influence of that aspect or of another.

When we look at a single age-specific rate for a given year, we have no indication of the extent to which that rate reflects the influences of chronological age, calendar time-associated changes in the social and physical environment, or characteristics of the cohort that happens to be passing through that age during that year. Even if we look at a given age interval across a span of calendar years or at multiple ages in a given year, there is no way for us to know whether what appear to be changes associated with aging or the passage of time are really reflections of the characteristics of different cohorts (i.e., characteristics acquired due to environmental experiences at a formative period of life, such as exposure to lead in infancy or to radiation in adolescence).

Attempts to disentangle the interwoven effects of age, secular time, and cohort are referred to as "age-period-cohort" analyses. The most straightforward approach involves assembling data from more than one period and from a broad range of ages, and then examining the data in relation to age, period, and cohort. For example:

Age-period-cohort analysis of mean serum cholesterol (mg/dL, hypothetical data)

60-69	200^{A}	210^{B}	235^{C}	240^{D}	$\underline{230}^{E}$
50-59	205^{B}	230^{C}	235^{D}	$\underline{225}^{E}$	215^{F}
40-49	240^{C}	230^{D}	$\underline{220}^{E}$	210^{F}	200^{G}
30-39	225^{D}	$\underline{215}^{E}$	205^{F}	195^{G}	185^{H}
20-29	$\underline{210}^{E}$	200^{F}	190^{G}	180^{H}	170^{I}
	1950-59	1960-69	1970-79	1980-89	1990-96

Birth cohorts:

A - 1890-1899	D - 1920-1929	G - 1950-1959
B - 1900-1909	E - 1930-1939 (underlined)	H - 1960-1969
C - 1910-1919	F - 1940-1949	I - 1970-1979

From the columns (calendar decades), it appears that serum cholesterol increases by 15 mg/dL per decade of age. If we had only one calendar decade of data, this observation is all that we can make, leading us to overstate the relationship between age and cholesterol. With the full data, we can follow the birth cohorts longitudinally, which reveals that for a given cohort cholesterol rises by 5 mg/dL per decade of age, but that also each new cohort has 10 mg/dL lower average cholesterol than the previous one.

This observation can be labeled a "cohort effect" and has the capability to confuse interpretation of cross-sectional (one point in time) data. (The reason that the 15 mg/dL increase does not continue at the older ages in the earlier decades is that I decided to precede the secular decline in cholesterol with a secular rise, so that the earliest cohorts had lower cholesterol levels than the ones that came afterwards.)

Thought question: Professors typically comment that with each entering class (i.e., cohort), students seem to be younger. Is this an effect of age, secular time, or cohort? (See bottom of page for the answer.)

Bibliography

Colton, Theodore. in Statistics in Medicine, pp. 237-250.

De Vita, Carol J. The United States at mid-decade. *Population Bulletin*, Vol. 50, No. 4 (Washington, D.C.: Population Reference Bureau, March 1996).

Falkenmark, Malin and Carl Widstrand. Population and water resources: a delicate balance. *Population Bulletin*, Vol. 47, No. 3 (Washington, D.C.: Population Reference Bureau, November 1992).

Lutz, Wolfgang. The future of world population. *Population Bulletin*, Vol. 49, No. 1, June 1994.

McFalls, Joseph A., Jr. Population: a lively introduction. *Population Bulletin*, Vol. 46, No. 2 (Washington, D.C.: Population Reference Bureau, October 1991)

McMichael, Anthony J. Planetary overload. Global environmental change and the health of the human species. NY, Cambridge, 1993.

Mosley, W. Henry and Peter Cowley. The challenge of world health. Population Bulletin, Vol. 46, No. 4 (Washington, D.C.: Population Reference Bureau, December 1991).

Petersen, William. *Population* 2nd ed. Macmillan, London,1969

Remington, Richard D. and M. Anthony Schork. Statistics with applications to the biological and health sciences. Englewood Cliffs, NJ, Prentice-Hall, 1970.

Roudi, Farzaneh. Iran's revolutionary approach to family planning. *Population Today*, July/August 1999; 27(7): 4-5. (Washington, D.C.: Population Reference Bureau)

Web sites: Data and publications are now widely available over the World Wide Web. A list of useful sites (e.g., www.ameristat.org, www.cdc.gov and www.cdc.gov/nchswww, www.who.int) is available at www.sph.unc.edu/courses/epid168/.

Answer: *Age* - the aging of the professors!

Descriptive studies and surveillance - Assignment

Reading: John C. Bailar and Elaine M. Smith. Progress against cancer? *N Engl J Med* 1986; 314:1226-32.

1. During the 20 years from 1962 to 1982, the number of Americans dying from cancer increased 56%, from 278,562 to 433,795. If a news reporter asked you how these numbers could be regarded as anything other than an indication of clear defeat in the "War against Cancer" declared by President Nixon in 1971, what issues about interpretion of these numbers would be important to explain to the reporter?

2. Assuming perfect information, which measure — mortality, incidence, or survival — is the best index of possible progress against cancer? Why?

3. What are the limitations of using national mortality statistics to assess changes in cancer rates?

4. If breast cancer was the leading cause of cancer death in women in 1982, why are the breast cancer mortality rates in Figure 2 so far below those for lung and colon/rectum cancer?

5. Prostate cancer mortality rates in Figure 2 have remained stable despite continual increases among nonwhite men. What are possible reasons why the overall rates have remained stable in spite of this increase?

6. For which change in site-specific cancer mortality in Figure 2 would epidemiology most like to claim credit? Who or what probably deserves credit? Explain.

7. What are some of the limitations of available incidence data for assessing progress against cancer?

8. What are some of the limitations of using case survival data for assessing progress against cancer?

The following questions pertain to the Standardization topic, which has not been covered yet. But see what you can do with them from information in the article or from your own knowledge.

9. Why has the dramatic decline in age-adjusted cancer mortality in Americans under age 30 had so little impact on total cancer mortality?

10. Why do the authors elect to use a direct, rather than an indirect, adjustment procedure for mortality and incidence rates?

11. Figure 5 projects age-adjusted cancer mortality to the year 2000. Would direct or indirect adjustment have been more appropriate for this figure? If the NCI goal is achieved, will crude cancer mortality fall more or less sharply than the projection in Figure 5?

12. Has the War on Cancer been lost? Should resources be shifted from research on cures to research on prevention? Why?

Descriptive studies and surveillance - Assignment solutions

1. The absolute number of cancer deaths is a function of the number of Americans and their personal characteristics, such as age and gender. During the period from 1962 to 1982, the U.S. population has increased in size, and the population age distribution has shifted so that both median age and the proportion of Americans above age 65 have increased. These changes would greatly increase the number of cancer deaths independently of any advance or retreat in the "War on Cancer". Conversely, the decline in the population sex ratio (males:females) would lower the number of cancer deaths (since the death rate for men is greater than that for women). Therefore rates and proportions, which express numbers of deaths in relation to population size, are more informative than are raw numbers. Similarly, adjustment for characteristics (e.g., age and sex) that are regarded as irrelevant to the question at hand provide a better comparison.

2. The choice of a measure(s) depends upon the study question or objective as well as on the availability of data. Mortality measures reflect both incidence and survival and are also more widely available (from vital statistics data), so mortality is the best single indicator for a "bottom line" index. On the other hand, nonfatal disease entails suffering, disability, and costs, so incidence is in some ways a better (and more demanding) measure of disease burden than is mortality, especially when the disease does not lead rapidly and inevitably to death. Of course, progress against cancer can take many forms, including reduced incidence, detection and simpler treatment of presymptomatic or precancerous lesions, improved survival, less pain and suffering, and improved quality of life among survivors. A thorough examination would involve all these dimensions.

3. Cancer mortality statistics necessarily depend upon the classification and coding of cause of death. Death may occur from or in the presence of multiple pathologic processes (e.g., cancer, heart disease, lung disease), in which case a decision must be made in selecting the "underlying cause" that determines how the death is tabulated in vital statistics reports. All of these factors can differ from place to place and can change over time, as diagnostic methods, access to care, and understanding of disease improve. So various factors besides the incidence of a disease and the effectiveness of treatment for it can complicate comparisons of mortality rates.

4. The breast cancer (and prostate cancer) mortality rates shown in Figure 2 are based on the entire population (p 1227, col 2), even though primarily (only?) women (and only men) contribute deaths to the respective numerators.

5. For overall prostate cancer mortality rates to remain stable in spite of increases among nonwhite men, prostate cancer mortality rates among white men must have declined.

6. Epidemiology would presumably most like to claim credit for the decline in stomach cancer, because of its steepness and because the decline reflects lower incidence, i.e., prevention. But the decline is probably a result of improvements in socioeconomic status, nutrition, and transport, storage, and preservation of foodstuffs, which did not come about as the result of findings or recommendations from epidemiology. In fact, the decline began before chronic disease epidemiology had really got underway.

7. Incidence data are available for only a (nonrandom) portion of the U.S. population (SEER data cover only 10%) and go back only about 25 years. There are too few data to estimate stable annual rates for nonwhites. Furthermore, the clinical importance of lesions found through sensitive screening procedures is sometimes uncertain. If lesions have the microscopic appearance of cancer they will be reported, yet in some cases they may not behave as cancer or may progress so slowly that they will not influence the life or health of the patient (as appears to be the case for the majority of prostate cancers).

8. Survival rates have as their denominator cases of a disease. Any problems in defining a case and classifying individuals as cases can confound survival rates. In particular, "overdiagnosis" (classifying as cancer lesions that do not or at least do not yet exhibit malignant behavior) will spuriously inflate survival rates. Also, earlier detection of truly malignant lesions, by advancing the time of detection ("lead time") from the time when symptoms occur, will increase the time between detection and death (survival time) regardless of an effect of treatment.

9. Because such deaths account for a very small portion of total cancer mortality, their influence on total cancer mortality is minor.

10. Direct adjustment uses weighted averages obtained from a single set of weights, so the adjusted rates are comparable to one another. Indirect adjustment uses weights from each separate group to compute its adjusted rate, so technically speaking, these rates can be compared only to the standard population.

11. Direct adjustment is appropriate because the figure compares mortality rates for different years (which would be problematic unless all are adjusted using the same weights) and the numbers of deaths are adequate to satisfy direct adjustment's need for stable estimates of age-specific rates. Since the population of the U.S. is aging, declines in age-specific mortality rates will be partly offset by a greater proportion of the population in age groups with higher mortality rates. Therefore, the actual (crude) death rate for cancer will not decline as sharply as will the age-adjusted death rate (assuming we achieve the NCI goal).

12. Perhaps not lost, but certainly not won. On the other hand, people affected by cancer (their own or a loved one's) are generally much more interested in and grateful for new treatments than are people who are never affected by cancers grateful for preventive measures. This is the paradox of public health and a major challenge to shifting the allocation of resources towards prevention.

Bailar and Smith assert: "By making deliberate choices among these measures, one can convey any impression from overwhelming success against cancer to disaster." (page 1231). Or as stated in the "evolving text", the choice of a measure depends upon the objective of the measure (!).

4. The Phenomenon of Disease

Concepts in defining, classifying, detecting, and tracking disease and other health states. The concept of natural history – the spectrum of development and manifestations of pathological conditions in individuals and populations.

Definition and classification of disease

Although the public health profession is sometimes inclined to refer to the health care system as a "disease care system", others have observed that public health also tends to be preoccupied with disease. One problem with these charges is that both "health" and "disease" are elusive concepts.

Defining health and disease

Rene Dubos (*Man Adapting,* p348) derided dictionaries and encyclopedias of the mid-20th century for defining "disease as any departure from the state of health and health as a state of normalcy free from disease or pain". In their use of the terms "normal" and "pathological", contemporary definitions (see table) have not entirely avoided an element of circularity.

Rejecting the possibility of defining health and disease in the abstract, Dubos sees the criteria for health as conditioned by the social norms, history, aspirations, values, and the environment. Thus diseases that are very widespread come to be considered as "normal" or an inevitable part of life. Dubos observes that in a certain South American tribe, pinta (dyschromic spirochetosis) was so common that the Indians regarded those *without* it as being ill. Japanese physicians have regarded chronic bronchitis and asthma as unavoidable complaints, and in the mid-19th century U.S., Lemuel Shattuck wrote that tuberculosis created little alarm because of its constant presence (Dubos, 251). As for the idealistic vision of health embodied in the WHO Constitution:

> ". . . positive health . . . is only a mirage, because man in the real world must face the physical, biological, and social forces of his environment, which are forever changing, usually in an unpredictable manner, and frequently with dangerous consequences for him as a person and for the human species in general." (*Man Adapting,* 349)

Clearly, general definitions of health and disease involve biological, sociological, political, and philosophical considerations. Such definitions also have important implications, since they delimit appropriate arenas for epidemiology and public health. But even with a consensus on a general definition, we will still face major challenges in recognizing and classifying the myriad diversity of health-related phenomena encountered in epidemiology and other health (disease) sciences.

Some definitions of disease and health

Dorland's Illustrated Medical Dictionary (28th ed., Phila, Saunders, 1994):

Disease – "any deviation from or interruption of the normal structure or function of any part, organ, or system (or combination thereof) of the body that is manifested by a characteristic set of symptoms and signs . . .".

Health – "a state of optimal physical, mental, and social well-being, and not merely the absence of disease and infirmity."

Stedman's Medical Dictionary (26th ed., Baltimore, Williams & Wilkins, 1995):

Disease –

1. An interruption, cessation, or disorder of body functions, systems, or organs;
2. A morbid entity characterized usually by at least two of these criteria: recognized etiologic agent(s), identifiable group of signs and symptoms, or consistent anatomical alterations.
3. Literally dis-ease, the opposite of ease, when something is wrong with a bodily function."

Health

1. The state of the organism when it functions optimally without evidence of disease or abnormality.
2. A state of dynamic balance in which an individual's or a group's capacity to cope with all the circumstances of living is at an optimum level.
3. A state characterized by anatomical, physiological, and psychological integrity; ability to perform personally valued family, work, and community roles; ability to deal with physical, biological, and psychological and social stress; a feeling of well-being; freedom from the risk of disease and untimely death."

Taber's Cyclopedic Medical Dictionary (17th ed. Phila., FA Davis, 1993. Ed. Clayton L. Thomas):

Disease – "Literally the lack of ease; a pathological condition of the body that presents a group of clinical signs and symptoms and laboratory findings peculiar to it and that sets the condition apart as an abnormal entity differing from other normal or pathological body states. The concept of disease may include the condition of illness or suffering not necessarily arising from pathological changes in the body. There is a major distinction between disease and illness in that the former is usually tangible and may even be measured, whereas illness is highly individual and personal, as with pain, suffering, and distress." [Examples given include: hypertension is a disease but not an illness; hysteria or mental illness are illnesses but have no evidence of disease as measured by pathological changes in the body.]

Classification is the foundation

As stated in an early (1957) edition of the *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death* (ICD):

> "Classification is fundamental to the quantitative study of any phenomenon. It is recognized as the basis of all scientific generalization and is therefore an essential element in statistical methodology. Uniform definitions and uniform systems of classification are prerequisites in the advancement of scientific knowledge. In the study of illness and death, therefore, a standard classification of disease and injury for statistical purposes is essential." (Introduction, pp. vii-ix)

The eminent Australian statistician, Sir George H. Knibbs, credited Francois Bossier de Lacroix (1706-1777), better known as Sauvages, with the first attempt to classify diseases systematically, in his *Nosologia Methodica*.

A classification is not merely a set of names to be applied to phenomena, although a **nomenclature** – a list or catalog of approved terms for describing and recording observations – is essential. As explained in the ICD:

> "Any morbid condition that can be specifically described will need a specific designation in a nomenclature. . . This complete specificity of a nomenclature prevents it from serving satisfactorily as a statistical classification [which focuses on groups of cases and aims to bring together those cases that have considerable affinity]. . . . A statistical classification of disease must be confined to a limited number of categories which will encompass the entire range of morbid conditions. The categories should be chosen so that they will facilitate the statistical study of disease phenomena.
>
> "Before a statistical classification can be put into actual use, it is necessary that a decision be reached as to the inclusions for each category. . . . If medical nomenclature were uniform and standard, such a task would be simple and quite direct. Actually the doctors who practise and who will be making entries in medical records or writing medical certificates of death were educated at different medical schools and over a period of more than fifty years. As a result, the medical entries on sickness records, hospital records, and death certificates are certain to be of mixed terminology which cannot be modernized or standardized by the wave of any magician's wand. All these terms, good and bad, must be provided for as inclusions in a statistical classification."

There is not necessarily a "correct" classification system. In classifying disease conditions, choices and compromises must be made among classifications based on etiology, anatomical site, age, and circumstance of onset, as well as on the quality of information available on medical reports. There may also need to be adjustments to meet varied requirements of vital statistics offices, hospitals, armed forces medical services, social insurance organizations, sickness surveys, and numerous other

agencies. The suitability of a particular system depends in part on the use to be made of the statistics compiled and in part on the information available in deriving and applying the system.

Defining and measuring the phenomena

Perhaps the first and most important issue in planning or interpreting an epidemiologic study is the definition and measurement of the disease and/or phenomena under study. How satisfactorily this issue can be addressed depends on the nature of the phenomena, the extent of knowledge about it, and the capability of available technology. The specific circumstances can range from the report of a case or series of cases that do not fit the characteristics of any known disorder to a disease that has been thoroughly studied and for which highly accurate and specific diagnostic procedures are available.

In the former category would fall the investigation of the condition that now bears the label chronic fatigue syndrome, where a vague collection of nonspecific symptoms was proposed to constitute a previously unrecognized disease entity, which still awaits a consensus regarding its existence. In situations such as these, a first task is formulating at least a provisional case definition in order to proceed with the investigation. In the latter category would fall rabies, where a specific, highly virulent organism has been identified and produces characteristic manifestations. Psychiatric disorders would fall somewhere in between. The nub of the problem is that the clarity with which features of the condition – etiologic factors, co-factors, natural history, response to treatment – can be linked to it depends on how effective are definition and measurement at excluding other entities whose different features will become mixed with those truly characteristic of the condition.

Consider an example. Although angina pectoris had been described in the 18th century (by William Heberden), and some 19th century physicians recognized an association between this symptom and coronary artery sclerosis found at autopsy, the syndrome of acute myocardial infarction (MI) was not recognized until the 20th century. According to W. Bruce Fye [The delayed diagnosis of myocardial infarction: it took half a century. *Circulation* 1985; 72:262-271)] the delay was due to the belief until 1912 that MI was invariably fatal and also to (1) the inconstant relationship of symptoms to pathological findings, (2) excessive reliance on auscultation as an indicator of cardiac disease, (3) failure to routinely examine coronary arteries or myocardium at autopsy, (4) tardiness of clinicians to incorporate new pathophysiologic discoveries into medical practice, (5) willingness to accept theories of disease not supported by scientific evidence, (6) pre-occupation with the new field of bacteriology, and (7) the lack of diagnostic techniques with which to objectively identify coronary artery obstruction or its consequences during life. (This list of reasons fits very well into Thomas Kuhn's description of the process of paradigm shifts – see citation in chapters 1 and 2.)

Classification criteria and disease definition

Since no two entities are completely identical, we (often unconsciously) group them together or differentiate between them according to what we believe to be important for our purposes. Even conditions with different etiologies may nevertheless have the same prognosis or the same response to treatment. Decisions about how far to subdivide categories of what appears to be a single entity depend, therefore, on the difference it may make, the level of knowledge, and our conceptual model.

As we gain more sophisticated understanding of the pathophysiological and biochemical mechanisms of disease conditions – to which the dramatic advances in molecular biology have contributed greatly – opportunities to differentiate among conditions now treated as a single entity and questions about whether to do so are becoming more frequent. For example, a mutation in the p53 gene is present in about 50% of cancers. Should cancers be classified according to whether or not an aberrant p53 gene is present? Is this aspect more important than the anatomical site or the histologic type? If two cancers of the same site and histologic type have mutations at different loci of p53, should they be classified apart?

There are two broad approaches to defining a disease entity. These two approaches are manifestational criteria and causal criteria [see discussion in MacMahon and Pugh].

Manifestional criteria

Manifestational criteria refer to symptoms, signs, behavior, laboratory findings, onset, course, prognosis, response to treatment, and other manifestations of the condition. Defining a disease in terms of manifestational criteria relies on the proposition that diseases have a characteristic set of manifestations (syndromes). Most chronic and psychiatric diseases are defined by manifestational criteria (examples: diabetes mellitus, schizophrenia, cancers, coronary heart disease).

Causal criteria

Causal criteria refer to the etiology of the condition, which, of course, must have been identified in order to employ them. Causal criteria are most readily available when the condition is simply defined as the consequences of a given agent or process (e.g., birth trauma, lead poisoning). The other group of conditions where causal criteria are available consists mostly of infectious diseases for which the pathogen is known, e.g., measles. Through the use of causal criteria, diverse manifestations recognized as arising from the same etiologic agent (e.g., the various presentations of infection with *Treponema pallidum* [syphilis] or with *Borrelia burgdorferi* [Lyme disease]) can be classified as the same disease entity. Similarly, conditions that have a similar presentation (e.g., gonorrhea, chlamydia) can be differentiated.

Manifestational versus causal criteria

The rationale for defining diseases based on manifestational criteria is borne largely of necessity – until we know the etiology, what else can we do? – and partly of the expectation that conditions with similar manifestations are likely to have the same or at least related etiology. Although this expectation has often been fulfilled, it is by no means a certainty. Simply because two conditions have identical manifestations (to the extent that we are currently able to and knowledgeable enough to measure these) does not ensure that they are the same entity in all other relevant respects, notably etiology. For example, even if major depressive disorder could be diagnosed with 100% agreement among psychiatric experts, the possibility would still exist that the diagnosis embraces multiple disease entities with very different etiologies. Similarly, an etiologic process that leads to major depressive disorder may be expressed with different manifestations depending upon circumstances and host characteristics.

Replacement of manifestional criteria by causal criteria

Nevertheless, the process seems to work. The evolution of the definition and detection of a disease, with the replacement of definitions based on manifestational criteria with definitions based on causal criteria, is well illustrated by HIV/AIDS. In 1981, clinicians in San Francisco reported seeing young American men with Kaposi's sarcoma, a tumor previously seen only in elderly Mediterranean males. Physicians in Europe found similar tumors in people from Africa. Shortly afterward, the Centers for Disease Control (CDC) noted that requests for pentamidine, a rarely prescribed antibiotic used for treating pneumocystis carinii pneumonia (PCP – an opportunistic infection generally seen only in medically immunosuppressed patients), had increased sharply from California. Investigation revealed that PCP was occurring in apparently otherwise healthy young men.

A first step in investigating a new, or at least different, disease is to formulate a case definition that can serve as the basis for identifying cases and conducting surveillance. The Acquired Immunodeficiency Syndrome (AIDS) was defined by the CDC in terms of manifestational criteria as a basis for instituting surveillance (reporting and tracking) of this new syndrome. The operational definition grouped diverse manifestations – Kaposi's sarcoma outside its usual subpopulation, PCP and other opportunistic infections in people with no known basis for immunodeficiency – into a single entity on the basis of similar epidemiologic observations (similar population affected, similar geographical distribution) and their sharing of a particular type of immunity deficit (elevated ratio of T-suppressor to T-helper lymphocytes).

After several years human immunodeficiency virus (HIV, previously called human lymphotrophic virus type III) was discovered and demonstrated to be the causal agent for AIDS, so that AIDS could then be defined by causal criteria. However, because of the long latency between infection and the development of AIDS, manifestational criteria are still a part of the definition of AIDS, though not of HIV infection itself. The original CDC reporting definition was modified in 1985 (*Morbidity and Mortality Weekly Report* [*MMWR*] 1985;34:373-5) and again in 1987 (*MMWR* 1987:36 [suppl. no. 1S]:1S-15S) to incorporate (1) a broader range of AIDS-indicator diseases and conditions and (2) HIV diagnostic tests. The proportions of AIDS cases that meet only the newer definitions vary by gender, race, and risk category.

In parallel with the institution of U.S. reporting definitions there has been an evolution in the international disease classification for AIDS. An original interim ICD classification for AIDS was issued on October 1, 1986, with the expectation that periodic revisions would be required. The first revision (January 1, 1988) characterized the causal agent and the change in terminology from human T-cell lymphotropic virus-III (HTLV-III) to HIV (Centers for Disease Control. Human immunodeficiency virus (HIV) infection codes and new codes for Kaposi's sarcoma: official authorized addenda ICD-9-CM (Revision 2) effective October 1,1991. *MMWR* 1991; 40(RR-9):1-19). The 1991 revision dealt only with morbidity reporting and provided for more detail about manifestations of HIV infection. All manifestations of HIV infection were to be coded, but a hierarchical classification was made for the stage of HIV infection. Distinctions were made between conditions occurring with HIV infection (e.g., 042.0: HIV with toxoplasmosis) and those occurring due to HIV infection (e.g., 042.1: HIV causing tuberculosis).

To recapitulate the above discussion, where we are fortunate, the classification based on manifestational criteria will closely correspond with that based on causal criteria but this is by no means assured because:

1. A single causal agent may have polymorphous effects (e.g., cigarette smoking is a causal factor for diverse diseases, herpes zoster causes chicken pox and shingles);
2. Multiple etiologic pathways may lead to identical (or at least apparently identical) manifestations, so that a (manifestationally-defined) disease entity may include subgroups with differing etiologies;
3. Multicausation necessitates a degree of arbitrariness in assigning a single or necessary cause to a disease category. For example, nutritional status and genetic constitution are contributing factors for tuberculosis. Had medical knowledge developed differently, tuberculosis might be known as a nutritional disorder with the bacillus as a contributory factor.
4. Often, not all persons with the causal agent (e.g., hepatitis A) develop the disease.

In actual epidemiologic practice, most disease definitions are based on manifestational criteria and proceed on the general assumption that the greater the similarity of the manifestations, the more likely the illness represents a unitary disease entity. The objective is to form classifications that will be useful in terms of studying the natural history of the disease and its etiology and also for treatment and prevention. There are numerous contemporary (e.g., Gulf War syndrome, chronic fatigue syndrome) as well as historical examples of this basic approach.

Disease classification systems

As diseases are defined they are organized into a classification. The primary disease classification system in use is the International Classification of Disease (ICD), now published by the World Health Organization. Introduced in 1900 for the purposes of classifying causes of death, the ICD was expanded to cover illness and injury in 1948. In the United States, the National Center for Health Statistics publishes an adapted version of the ICD to incorporate syndromes and illnesses not listed in the WHO edition. The American Psychiatric Association performs a similar function for classification of mental disorders, with its Diagnostic and Statistics Manual (DSM) (see below).

Disease classification systems do not necessarily provide the kind of information needed for public health research and policymaking. Diseases and deaths related to tobacco use, for example, cannot be identified from ICD codes, though there has been a movement toward having tobacco use appear as a cause on the death certificate. In the injury area, injuries are classified according to the nature of the injury (e.g., laceration, puncture, burn) rather than the nature of the force that caused it (e.g., gunshot, fire, automobile crash, fall). Injury prevention researchers advocate the use of E (External) codes to permit tabulation by the external cause of the injury.

Classification systems, of course, must be periodically revised to conform to new knowledge and re-conceptualizations. Revisions typically include changes in:

1. usage of diagnostic terms (e.g., for heart disease);

2. disease definitions;
3. organization of categories based on new perceptions about similarities among conditions (e.g., joint occurrence of hypertension and CHD);
4. coding rule (e.g., priorities for selecting an underlying cause of death when multiple diseases are present).

Such changes come at a price, in the form of discontinuities in disease rates over time and confusion for the unwary. For example, prior to 1986, carcinoid tumors were reportable to the National Cancer Institute's Surveillance, Epidemiology, and End Results program (SEER) only if they were specifically described as "malignant gastric carcinoid." In 1986, any tumor described as "gastric carcinoid" was considered malignant and therefore was reportable to SEER. This change produced a substantial rise in the rate of gastric carcinoid tumors in 1986.

Similar problems in comparing rates over time, across geographical area, or among different health care providers can arise from differences or changes in "diagnostic custom" or terminology preferences (see Sorlie and Gold, 1987). In addition, the Diagnosis Related Group (DRG) system introduced in the United States to control the costs of federal reimbursement to hospitals for health care has undoubtedly influenced discharge diagnoses in favor of those with higher reimbursement opportunity

Conceptual questions in classifying diseases

Even without the complicating factors of diagnostic custom or changes in classification systems, by its very nature classification poses difficult conceptual questions whose resolutions underlie the disease definitions we employ. Some examples:

1. What constitutes "similarity"?
   Examples: adult versus juvenile onset diabetes; melanoma in the retina versus in the skin; pneumonia of viral, bacterial, or chemical origin; cancers with different genetic "signatures".
2. What is the appropriate cutpoint on a continuum?
   Examples: blood pressure and hypertension; plasma glucose and diabetes; alcohol consumption and alcoholism; fetal death and gestational age.
3. How should ambiguous situations be handled?
   Examples: hypertension controlled with drugs; subclinical infection; alcoholism, schizophrenia or depressive disorder in remission.

As perceptions and understanding changes, so do the answers to the questions. For example, in moving from DSM-III to DSM-IV, the American Psychiatric Association removed the distinction between "organic" and "inorganic" psychiatric disorders, added categories for premenstrual syndrome and gender identity problems, and introduced a V-code (nonpathological descriptor) for religious or spiritual problems ("Psychiatrists set to approve DSM-IV", *JAMA* 7/7/93, 270(1):13-15).

Classifying cause of death

Since mortality data are generally the most widely available, epidemiologists encounter the above problems most often in evaluating the accuracy of cause-specific mortality rates. Cause-specific mortality rates are tabulated using the "underlying cause of death", and until recently this was the only cause available in electronic form. The underlying cause of death is defined as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances or violence which preceded the fatal injury" (*Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death*, Geneva, WHO, 1977: 609-701, quoted in Kircher and Anderson, *JAMA* 1987:349). According to Kircher and Anderson, most physicians confuse cause and mechanism. For example, congestive heart failure, cardiorespiratory arrest, asphyxia, renal failure are mechanisms – the means by which the cause exerts its lethal effect.

The following are additional operational problems in assigning a cause of death (see Israel *et al*. 1986):

1. Many conditions can coexist without a direct etiologic chain. When a combination of causes is forced into a single cause, the choice may be arbitrary, even if systematic, and the true circumstances obscured.
2. There is confusion about disease terms; e.g., it is often unclear whether "metastatic" disease refers to a primary or secondary tumor.
3. There is confusion among certifiers about the meaning of classification terms (e.g., "underlying", "immediate", and "contributory" causes). [Confusion is perhaps to be expected, given the complexity of concept and circumstances. According to the ICD, "The words 'due to (or as a consequence of)' . . . include not only etiological or pathological sequences, but also sequences where there is no such direct causation but where an antecedent condition is believed to have prepared the way for the direct cause by damage to tissues or impairment of function even after a long interval." (*Manual of the international statistical classification of diseases, injuries, and causes of death, based on the recommendations of the Ninth Revision Conference*, 1975. Geneva: WHO, 1977:700, quoted in *MMWR* 1991 (26 July);40:3)]
4. Death certificates are often completed late at night or in haste, sometimes to speed funeral arrangements, by a sleepless physician who has never seen the deceased before and for whom care of the living is understandably a higher priority. Partly for these reasons death certificate information is often sloppy or incomplete. Amended certificates with more complete information can be but are rarely filed, and unlikely diagnoses are rarely queried.

Both mortality statistics and case ascertainment for epidemiologic studies can readily be affected by such problems and circumstances (see Percy, *et al*. 1981). Epidemiologic studies for which cause of death is important often have a copy of each death certificate reviewed by a trained nosologist, an expert in classifying diseases, to confirm or correct questionable cause of death information. If resources are available, medical records may be obtained to validate a sample of the death certificates and/or to resolve questions.

To illustrate the challenge of classifying cause of death, consider the following case example from Kircher and Anderson:

A 65-year-old woman was first seen by her physician five years before her death when she complained of dyspnea and substernal chest pain precipitated by exertion. The electrocardiogram on a standardized exercise test demonstrated depression in the ST segments of 1.5 mV. The patient's symptoms were alleviated by a hydrochlorothiazide-trimterene combination (Dyazide) and sublingual nitroglycerine until nine months before death, when the frequency and severity of angina increased. Propranolol hydrochloride was prescribed.

One month before death and ten days after the onset of a flulike illness, the patient developed chills, fever, and pleuritic pain. An x-ray film of the chest revealed patchy consolidation of both lungs. The leukocyte count was 20 x 109/L (20,000/mm^3). Blood cultures were positive for pneumococci. Seventy-two hours after penicillin G potassium therapy was initiated, the symptoms subsided.

One month after the episode of pneumonia, the patient sustained a myocardial infarction. Five days after the patient's admission to the hospital, death occurred suddenly. An autopsy revealed severe coronary atherosclerosis, left anterior descending coronary artery thrombosis, acute myocardial infarction, left ventricular myocardial rupture, hemopericardium, and cardiac tamponade.

In this case, the immediate cause of death was rupture of the myocardium. The rupture was due to an acute myocardial infarction occurring five days before death. The underlying cause of death – the condition setting off the chain of events leading to the death – was chronic ischemic heart disease. The deceased had had this condition for at least five years before her death. Influenza and pneumonococcal pneumonia should also be shown as other significant conditions that contributed to death.

Instructions for coding cause of death on death certificates can be found on the web page for the National Vital Statistics System of the National Center for Health Statistics, CDC (http://www.cdc.gov/nchs/about/major/dvs/handbk.htm). As of August 2000, the web page included links for a tutorial by the National Association of Medical Examiners and various handbooks.

Psychiatric disorders – a special challenge

The challenges of classification of physical disease are formidable, but psychiatric disorders present an even greater challenge due to the difficulty of finding satisfactory answers to the most basic of questions, "what is a case?" (John Cassel, Psychiatric epidemiology. In: S. Arieti (ed), *American handbook of psychiatry*. 2nd ed. NY, Basic Books, 1974, vol. 2, 401-410; Kendell, *Arch Gen Psychiatry* 1988; 45:374-376). Despite a (modest) increase in resources and effort aimed at unraveling the etiology of these disorders, causal relationships have been very difficult to demonstrate. A key reason for the lack of progress may be problems with the definition of mental disorders (Jacob KS. The quest for the etiology of mental disorders. *J Clin Epidemiol* 1994;47:97-99).

Laboratory and other objectively measurable physiological signs have been a tremendous asset for defining and classifying diseases. Accordingly the need to rely almost exclusively on symptoms, behavioral observation, response to treatment, course, and outcome – manifestations that are more difficult to measure with reliability and precision – has put psychiatric nosology at a great disadvantage compared with physical illness. Although recent progress in psychiatric nosology, reflected in DSM III, III-R, and IV is believed to have improved reliability of diagnoses, the resulting diagnostic classifications are probably heterogeneous with respect to etiology. Subclassification based on biological and molecular variables, to take advantage of the significant advances in biology and biotechnology, and on refined measures of environmental and psychological variables may reveal etiologic associations that are masked by the current reliance on syndrome-based definitions (Jacob). On the other hand, if a disorder represents a "final common pathway", as has been argued with respect to unipolar major depressive disorder, then diverse etiologies could conceivably result in a biologically cohesive phenomenon.

Measuring accuracy in classification and detection

In general, any deviation between the (often-unknown) truly relevant biological entity and the result of the system used to define and detect or quantify it can be regarded as measurement error. Later in the course we will take up the concept of information bias, which deals with the effects of measurement error on study findings. Here, though, we will present the basic measures used in epidemiology to quantify accuracy of detection and classification methods. These measures can be applied to the detection of any entity, of course, whether it is a disorder, an exposure, or any characteristic. Besides their use in epidemiology in general, these measures are important for the selection and interpretation of diagnostic tests used in clinical practice.

If a condition or characteristic can be present or absent, then the accuracy of our system of detection and labeling can be assessed by its ability to detect the condition in those who have it as well as by its ability to correctly classify people in whom the condition is absent. Note that for a rare condition, overall accuracy [(a+d)/n in the table below] primarily reflects the correct identification of noncases, thus giving little information about the correct identification of cases. Also, overall accuracy ignores the fact that different kinds of errors have different implications.

Epidemiologists therefore employ separate, complementary measures for the correct classification of cases and of noncases. The basic measures are:

Sensitivity – the proportion of persons who have the condition who are correctly identified as **cases**.

Specificity – the proportion of people who do not have the condition who are correctly classified as **noncases**.

The definitions of these two measures of validity are illustrated in the following table.

Classification contingency table

		True status +	True status –		
Classified status	+	a	b	(a + b)	(Positive tests)
	–	c	d	(c + d)	(Negative tests)
Total		a + c (Cases)	b + d (Noncases)		

In this table:

Sensitivity (accuracy in classification of **cases**) = a / (a + c)

Specificity (accuracy in classification of **noncases)** = d / (b + d)

Sometimes the following terms are used to refer to the four cells of the above table:

a = True positive, TP – people with the disease who test positive

b = False positive, FP – people without the disease who test positive

c = False negative, FN – people with the disease who test negative

d = True negative, TN – people without the disease who test negative

However, these terms are somewhat ambiguous (note that "positive" and "negative" refer to the result of the test and not necessarily to the true condition). The relative costs (financial and human) of false negatives and false positives are key factors in choosing between sensitivity and specificity when a choice must be made. The more urgent is detection of the condition, the greater the need for sensitivity. Thus, a condition that has severe consequences if left untreated and which can be readily treated if detected early implies the need for a test with high sensitivity so that cases are not missed. A condition for which an expensive, invasive, and painful diagnostic workup will follow the results of a positive test implies the need for a test with high specificity, to avoid false positive tests.

Criterion of positivity and the receiver operating characteristic

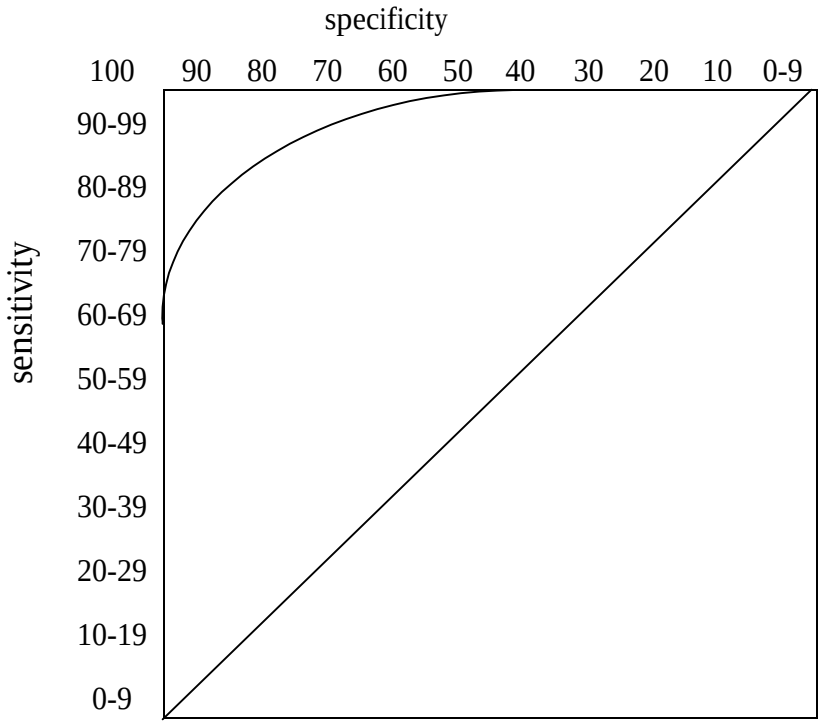
Often, however, the differences between cases and noncases are subtle in relation to the available classification system or detection methods. In such cases we can increase one type of accuracy only by trading it off against the other by where we set the "criterion of positivity", the cutpoint point used to classify test results as "normal" or "abnormal".

Screening for abnormal values of physiologic parameters is a typical situation. If we are attempting to classify people as diabetic or not based on their fasting blood glucose level, then we can set our cutpoint low in order to be sure of not missing diabetics (i.e., high sensitivity for detecting cases) but in doing so we will also include more people whose blood glucose falls at the upper part of the

distribution but are not diabetic (i.e., low specificity). If instead we choose a high cutpoint in order to avoid diagnosing diabetes when it is not present, then we are likely to miss diabetics (low sensitivity) whose blood glucose falls in the lower part of the distribution. Tests that involve rendering a judgment based on an image or specimen (e.g., mammography and cytology) involve similar, though less quantifiable, tradeoffs. As we shall see, in addition to the relative consequences of false negative and false positive tests, the decision of where to set the criterion of positivity should also take into account the prevalence of the condition in the population to be tested.

One useful technique for comparing the performance of alternative tests without first having to select a criterion for positivity and also for selecting a cutpoint is to graph the **receiver/response operating characteristic (ROC)** for each test (the concept and terminology come from engineering). The ROC shows the values of sensitivity and specificity associated with each possible cutpoint, so that its graph provides a complete picture of the performance of the test. For example, the sample ROC curve in the figure indicates that at 80% sensitivity, the test is about 95% specific. At 95% sensitivity, the specificity is only about 74%. If high sensitivity (e.g., 98%) is essential, the specificity will be only 60%.

Sample ROC curve



An ROC curve that consisted of a straight line from the lower left-hand corner to the upper right-hand corner would signify a test that was no better than chance. The closer the curve comes to the upper left-hand corner, the more accurate the test (higher sensitivity and higher specificity).

Predictive value

Sensitivity and specificity are, in principle, characteristics of the test itself. In practice, all sorts of factors can influence the degree of sensitivity and specificity that are achieved in a particular setting (e.g., calibration of the instruments, level of training of the reader, quality control, severity of the condition being detected, expectation of positivity). However, for any particular sensitivity and specificity, the yield of a test (accurate and inaccurate positive test results) will be determined by how widespread the condition is in the population being tested. The typical difficulty is that, since the number of people without the condition is usually much larger than the number with the condition, even a very good test can easily yield more false positives than true ones.

The concept of **predictive value** is used to assess the performance of a test in relation to a given frequency of the condition being sought. The **positive predictive value** (PPV) is defined as the proportion of people with the condition among all those who received a positive test result. Similarly, the **negative predictive value** is the proportion of people without the condition among all those who received a negative test result. Using the same table as before:

Classification contingency table

		True status +	True status –		
Classified status	+	a	b	(a + b)	(Positive tests)
	–	c	d	(c + d)	(Negative tests)
Total		a + c	b + d		
		(Cases)	(Noncases)		

Positive predictive value (PPV) = a / (a + b)

Negative predictive value (NPV) = d / (c + d)

Predictive value is an essential measure for assessing the effectiveness of a detection procedure. Also, since predictive value can be regarded as the probability that a given test result has correctly classified a patient, this concept is also fundamental for interpreting a clinical measurement or diagnostic test as well as the presence of signs or symptoms. The PPV provides an estimate of the probability that someone with a positive result in fact has the condition; the NPV provides an estimate that someone with a negative result does not in fact have the condition. (For a full discussion of the use of predictive value and related concepts in diagnostic interpretation, see a clinical epidemiology text, such as that by Sackett *et al.*)

In a clinical encounter prompted by symptoms, there is often a substantial probability that the patient has the condition, so both sensitivity and specificity are important in determining the proportion of cases and noncases among those who receive positive tests. However, in a screening

program in the general population, the specificity will typically dominate. Even with perfect sensitivity, the number of true cases cannot exceed the populatiion size multiplied by the prevalence, which is usually small. The number of false positives equals the false positive rate (1–specificity) multiplied by the number of nocases, which for a rare disease is almost the same as the population size. So unless the prevalence is greater than the false positive rate, the majority of test positives will not have the disease. For example, if only 1% of the population has the condition, then even if the specificity is 95% (false positive rate of 5%) the group who receive positive tests will consist primarily of noncases:

Cases detected (assume 100% sensitivity):

100% sensitivity x 1% with the condition = 1% of population

False positives:

95% specificity x 99% without the condition = 94.05% of population correctly classified, leaving 5.95% incorrectly labeled positive

Total positives:

1% + 5.95% = 6.95% of population

Proportion of positives who are cases (PPV) = 1% / 6.95% = 14%

In a population of 10,000 people, the above numbers become 100 (1%) cases, all of whom are detected, and 9,900 noncases, 595 of whom receive positive tests, for a total of 695 people receiving positive tests, 100 of whom have the condition. We will take up some of the these issues at the end of our discussion of natural history of disease.

Natural history of disease

Diseases and other phenomena of interest in epidemiology are processes. For example, the process by which bronchogenic lung cancer arises involves a progression over many years of the development of abnormal cells in the bronchial epithelium. Several grades of abnormality (metaplasia, mild dysplasia, moderate dysplasia, severe displasia) have been described. For the most part these abnormalities have the potential to disappear spontaneously or regress. However, in one or a number of cells the abnormality progresses to carcinoma *in situ* and then to invasive carcinoma. Depending upon the focus of investigation, the process can extend to very early stages. If our focus is on primary prevention, we might consider the process of smoking onset, usually in adolescence, the development of nicotine addiction and the smoking habit, and repeated attempts to quit. We might also consider the effects of tobacco marketing on smoking onset and maintenance, and the effects of legislation, litigation, competition, and investment opportunities on tobacco industry practices.

Thus, defining, observing, and measuring health and disease requires an appreciation of the concept of natural history – the evolution of a pathophysiologic process. "Natural" refers to the process in the absence of intervention. Natural history encompasses the entire sequence of events and developments from the occurrence of the first pathologic change (or even earlier) to the resolution of the disease or death. The natural history of a disease may be described through a **staging**

classification. Staging can aid in defining uniform study groups for research studies, determining treatment regimens, predicting prognosis, and in providing intermediate end-points for clinical trials.

Natural history therefore includes a **presymptomatic** period and a **postmorbid** period. Of particular interest for epidemiologists is the former, the period of time before clinical manifestations of the disease occur and, for infectious diseases, the period of time between infection and infectiousness. For non-infectious diseases, the term **induction period** refers to the "period of time from causal action until disease initiation" (Rothman and Greenland, p14). The induction period may be followed by a **latent period** (also called **latency**), which is the "time interval between disease occurrence and detection" (Rothman and Greenland, p15). This distinction, though not made by all authors, is important for diseases that can be detected through screening tests, since the latent period represents the stage of the disease natural history when early detection is possible.

The distinction is also important for designing epidemiologic studies. Since the time of disease detection may be advanced through the application of screening and diagnostic tests, the number of cases detected can change with technology. Also, the collection of historical exposure data should be guided by a concept of when such exposure would have been biologically relevant. For a factor believed to contribute to the initiation of a disease, exposure must occur before that point. For a factor believed to contribute to promotion or progression of the condition, exposure can take place following initiation.

For infectious diseases, there are two markers of epidemiologic importance: disease detection and the onset of infectiousness. **Incubation period** refers to the "time from infection to development of symptomatic disease" (Halloran, p530). This term is sometimes applied to non-infectious diseases, but often without a precise meaning. The incubation period thus covers both the induction and latent periods as these are defined for non-infectious diseases. In contrast, the term latent period has a different meaning for infectious diseases, where it denotes "the time interval from infection to development of infectiousness" (Halloran, p530). Since an infected person may be infectious before developing symptoms, while symptomatic, or after disappearance of symptoms, there is no absolute relationship of incubation and latent periods for infectious disease. Relevant concepts are **inapparent or silent infection** (asymptomatic, either infectious or non-infectious) and **carrier** (post-symptomatic but still infectious) (Halloran, pp530-531).

Infectious disease	
Incubation	"time from infection to development of symptomatic disease" (Halloran, p530)
Latency	"the time interval from infection to development of infectiousness" (Halloran, p530)
Non-infectious disease	
Induction	"period of time from causal action until disease initiation" (Rothman and Greenland, p14)
Latency	"time interval between disease occurrence and detection" (Rothman and Greenland, p15)

Acute versus chronic diseases

Historically, disease natural histories have been classified into two broad categories: acute and chronic. **Acute diseases** (typically infections) have short natural histories. **Chronic diseases** (e.g., cancer, coronary heart disease, emphysema, diabetes) have long natural histories. So great has been the dichotomy of acute/infectious disease versus chronic/noninfectious disease that many epidemiologists and even departments of epidemiology are frequently regarded as one or the other.

In 1973 in the first Wade Hampton Frost Lecture, Abraham Lilienfeld regretted the concept of "Two Epidemiologies" and sought to emphasize the aspects in common between infectious and noninfectious epidemiology (see *Am J Epidemiol* 1973; 97:135-147). Others (e.g., Elizabeth Barrett-Connor, Infectious and chronic disease epidemiology: separate and unequal? *Am J Epidemiol* 1979; 109:245) have also criticized the dichotomy both in terms of its validity and its effect on epidemiologic investigation.

The growth of evidence for viral etiologies for various cancers (notably T-cell leukemias and cervical cancer) as well as other chronic diseases (e.g., juvenile onset diabetes mellitus and possibly multiple sclerosis) and for the central roles of immune system functions in chronic diseases demonstrates the importance of building bridges between the two epidemiologies. Also problematic for the identities acute = infectious and chronic = noninfectious are slow viruses, such as HIV. HIV may or may not produce a brief, flu-like syndrome within a week or so after infection. During the several weeks or months the host antibody response develops, and the virus enters a prolonged subclinical state during which the virus appears to remain quiescent. Many years may elapse until a decline in CD4 lymphocytes occurs and results in (chronic) immune deficiency.

Knowledge of the pathophysiology of early HIV infection is the basis for the Serologic Testing Algorithm for Recent HIV Seroconversion (STARHS, Janssen *et al*., 1998). The STARHS technique uses an assay whose sensitivity has been deliberately reduced. Specimens found to be HIV-positive in a sensitive assay are retested with the "de-tuned assay". Failure to detect antibody with the less sensitive assay most likely signifies that the infection was recently-acquired and the antibody response has not fully developed. Thus, the technique makes it possible to establish what proportion of HIV infections in a population occurred recently, indicating the level of continuing transmission.

Spectrum of disease

Diseases typically involve a spectrum of pathologic changes, some of which are considered disease states and some pre-disease states. The spectrum of disease concept has been studied, at the cellular and molecular level, for both coronary artery disease and cancer. Seeing more of the full spectrum or sequence can make us less certain at what point the "disease" has actually occurred.

Coronary artery disease:

Coronary artery disease pathogenesis is now understood in considerable detail (e.g., see Fuster *et al. N Engl J Med*, Jan 23, 1992;326(4):242 and Herman A. Tyroler, Coronary heart disease in the 21st century. *Epidemiology Reviews* 2000;22:7-13). "Spontaneous" atherosclerosis is initiated by chronic minimal (Type I) injury to the arterial endothelium, caused mainly by a disturbance in the pattern of blood flow in certain parts of the arterial tree. This chronic injury can also be potentiated by various factors, including hypercholesterolemia, infection, and tobacco smoke constituents.

Type I injury leads to accumulation of lipids and monocytes (macrophages). The release of toxic products by macrophages leads to Type II damage, which is characterized by adhesion of platelets. Growth factors released by macrophages, platelets, and the endothelium lead to the migration and proliferation of smooth-muscle cells, contributing to the formation of a "fibrointimal lesion" or a "lipid lesion". Disruption of a lipid lesion leads to Type III damage, with thrombus formation.

Small thrombi can contribute to the growth of the atherosclerotic plaque. Large thrombi can contribute to acute coronary syndromes such as unstable angina, myocardial infarction, and sudden ischemic death. Autopsy studies have revealed early, microscopic lesions in infants, though they regress. In adolescents, fatty streaks containing smooth-muscle cells with lipid droplets are observed. At this age fatty streaks are not surrounded by a fibrotic cap, which develops on some lesions in the 20's. Progression to clinically manifest, enlarging atherosclerotic plaques, such as those causing exertional angina, may be slow (probably in response to Type I and Type II injury) or rapid (in response to Type III injury). At what point in this process is "coronary artery disease" present?

Cancer:

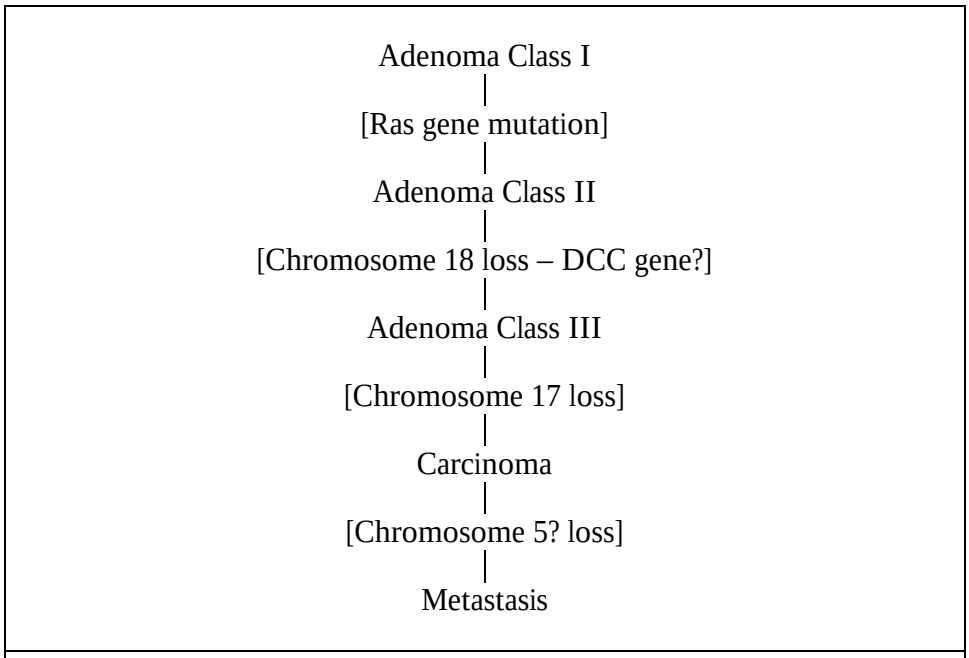
Cancer is also a multistage process, involving tumor initiation, promotion, conversion, and progression. Shields and Harris (Molecular epidemiology and the genetics of environmental cancer. *JAMA* August 7, 1991;266(5):681-687) describe the process as follows: "Tumor initiation involves

the direct effects of carcinogenic agents on DNA, mutations, and altered gene expression. The attendant defects are involved in tumor promotion, whereby cells have selective reproductive and clonal expansion capabilities through altered growth, resistance to cytotoxicity, and dysregulation of terminal differentiation. Tumor promotion further involves an 'initiated' cellular clone that may also be affected by growth factors that control signal transduction. During this process, progressive phenotypic changes and genomic instability occur (aneuploidy, mutations, or gene amplification). These genetic changes enhance the probability of intiated cells transforming into a malignant neoplasm, the odds of which are increased during repeated rounds of cell replication. During tumor progression, angiogenesis allows for a tumor to grow beyond 1 or 2 mm in size. Ultimately, tumor cells can disseminate through vessels, invading distant tissues and establishing metastatic colonies." (681-682) When did "cancer" begin?

One of the cancers where understanding of natural history process has progressed to the identification of specific gene mutations is colon cancer. The process begins with initiation, e.g. chemical or radiation genetic damage to cell. It is now believed that alteration of a gene on chromosome 5 induces a transformation from normal colonic epithelium to a hyperproliferative epithelium. Initiated cells may then go through a series of distinct stages. The transformation process is enhanced by "promoters", which may be harmless in the absence of initiation. Stages that have so far been identified and the accompanying genetic alterations are shown in the accompanying figure. The progression from normal epithelium to cancer takes about ten years.

Stage of the cancer at diagnosis is influenced by various factors including screening and largely determines the outcome of therapy. Basic stages are: localized (in tissue of origin), regional spread (direct extension to adjacent tissues through tumor growth), and metastatic spread (tumor sheds cells that form new tumors in distant areas). Symptoms of various kinds develop according to the particular tissues and organs affected, and even the particular type of tumor cell (e.g., tumors in nonendocrine tissues can sometimes produce hormones).

Thus, the natural history of a disease can involve many, complex processes and developments long before the appearance of a clinical syndrome and even before the existence of the "disease" can be detected with the most sophisticated clinical tests. Moreover, particularly since some of the early stages are spontaneously reversible, it is not always clear even theoretically when the "disease" itself is present.



(DCC = "deleted in colorectal cancer" gene) Correction of any one of chromosomes 5, 17, or 18 reduces malignancy, but only chromosome 18 restoration increases responsiveness to growth-inhibitory effects of transforming growth factor beta (TGF-beta).

Sources: Eric J. Stanbridge, Identifying tumor suppressor genes in human colorectal cancer. *Science* 5 Jan 1990; 247:12-13; Colorectal cancer: new evidence for the adenoma/carcinoma sequence. Editorial. *The Lancet* 25 July 1992; 340:210-211)

Understanding the natural history of diseases and other conditions of interest is fundamental for prevention and treatment, as well as for research. The effectiveness of programs for early detection and treatment of cancer, for example, depends upon the existence of an extended period where the cancer or a premalignant lesion is asymptomatic yet detectable and where treatment is more effective than after symptoms appear. In order to evaluate the efficacy of therapeutic interventions, knowledge of the natural history in the absence of treatment is crucial. These concepts will be illustrated by considering cancer screening procedures.

Natural history and screening

Population screening is defined as the application of a test to asymptomatic people to detect occult disease or a precursor state (*Screening in Chronic Disease*, Alan Morrison, 1985). The immediate objective is to classify them as being likely or unlikely of having the disease under investigation. The goal is to reduce mortality and morbidity on the basis of evidence that earlier treatment improves

patient outcomes. The design and evaluation of population screening programs depend crucially on the natural history of the disease in question.

For a screening program to be successful it must be directed at a suitable disease and employ a good test. Diseases for which screening may be appropriate are typically cancers of various sites (e.g., breast, cervix, colon, prostate), infectious diseases with long latency periods such as HIV and syphilis, and physiologic derangements or metabolic disorders such as hypertension, hypercholesterolemia, phenylketonuria, etc. What these conditions have in common is that they have serious consequences which can be alleviated if treatment is instituted early enough. The natural histories of these conditions involve a period of time when the condition or an important precursor condition (e.g., dysplasia) is present but during which there are no symptoms that will lead to detection.

Earlier in this topic we defined the latent period as the time between disease initiation and its detection. Cole and Morrison (1980) and Morrison (1985) refer to the total latent period as the **total pre-clinical phase** (TPCP). However, only a portion of the TPCP is relevant for screening – the period when the condition can be detected with the screening test. Cole and Morrison refer to this portion as the **detectable pre-clinical phase** (DPCP). The preclinical phases end when the patient seeks medical attention because of diagnostic symptoms. The DPCP is that part of the TPCP that begins when the screening test can detect the disease. Thus, the DPCP can be advanced if the screening test can be improved. The preclinical phase can be shortened by teaching people to observe and act promptly on early or subtle symptoms.

For a condition to be a suitable one for population screening, it must have a prolonged DPCP, thus providing ample time for advancing the date of disease detection and treatment. For a screening test to be suitable, it must be inexpensive, suitable for mass use, and without risk. It must have good sensitivity, so that the condition is not missed too often, which may give clients false reassurance. Moreover, the relevant sensitivity is for detecting the DPCP, rather than clinical disease, since it is the detection of the DCPC that provides the advantage from screening. The test must have excellent specificity, to avoid an excessive number of false positive tests. Importantly, the test must be able to maintain these attributes when administered and interpreted in volume in routine practice.

A major stumbling block in recommending population screening is the need to balance any benefit from early detection of cases against the expense, inconvenience, anxiety, and risk from the medical workups (e.g., colonoscopy, biopsy) that will be needed to follow-up positive tests on people who do not in fact have the condition. As demonstrated earlier, even a highly accurate test can produce more false positives than true ones when applied in a population where condition is very rare. Low positive predictive value (high proportion of false positives) has been a principal argument against HIV screening among applicants for marriage licenses, screening mammograms for women under age 50 years, and prostate cancer screening with prostate specific antigen (PSA).

(Although the test itself may be the same, it is important to distinguish between the use of a test for screening and its use for diagnosis. Since in the latter context the test has been motivated by the presence of signs or symptoms and history, the prevalence of the condition among the test recipients is much greater, so that a positive test has a much higher positive predictive value. The

term case-finding is sometimes used to refer to the application of the test to asymptomatic patients in a primary care setting. Case-finding likely assures effective follow-up for people receiving a positive test, though possible issues related to economic and personal costs of false positives remain.)

Criteria for early detection of disease through screening

Criteria to be met before screening for a given disease:

1. Natural history of disease must be understood
2. Effective treatment is available
3. A test is available by which the disease can be recognized in its pre-clinical phase
4. The application of screening makes better use of limited resources than competing medical activities

Evaluation of screening programs

Early outcomes for evaluating a screening program are stage of the disease and case fatality. If the screening is effective, the stage distribution for cases should be shifted towards earlier stages and a greater proportion of patients should survive for any given time period. Late outcomes are reduced morbidity and mortality. However, these outcome measures can all be affected by features of disease definition and natural history. Three potential pitfalls are lead time, length bias, and overdiagnosis.

Lead time is the amount of time by which screening advances the detection of the disease (i.e. the time between detection by a screening test and detection without a screening test). Even if the interval between the (unknown) biologic onset of the disease and death is unchanged, earlier detection will lengthen the interval between diagnosis and death so that survival appears lengthened. Lead time bias results when a screening program creates the appearance of delaying morbidity and mortality but in reality does not alter the natural history.

Length bias results if tumors are heterogeneous in respect to their aggressiveness, with slower growing tumors having a more favorable prognosis (or at least longer time to death). Slower growing tumors are more likely to be detected by screening, since they are present and asymptomatic longer (i.e., they have a longer DPCP) than are rapidly growing, aggressive tumors. So tumors detected by screening will overrepresent slow growing, hence survivable, tumors than will cancers detected because of appearance of symptoms (the latter cases are called "interval cases" because they are detected during the interval between screens).

Overdiagnosis results from the detection, by the screening test, of nonmalignant lesions that are judged to be malignant or to have malignancy potential. Prior to the use of the screening test, such lesions would not be detected, so their true prognosis may be unknown. If persons with these apparently very early lesions are counted as having the disease, yet such lesions would not in any event progress to clinically-significant tumors, the survival experience of cases

detected by screening will appear better. Overdiagnosis is a particular concern in evaluating the efficacy of prostate cancer screening.

Randomized trials, in which mortality is compared between a group offered screening and a group not offered screening (the classic study of this type is the Health Insurance Plan [HIP] trial of breast cancer screening) provide protection against these biases. But because they must usually be very large and of long duration, such trials are often difficult and very costly. The National Cancer Institute is currently conducting a very large (74,000 men and 74,000 women) and lengthy randomized trial to evaluate the effectiveness of screening for prostate, lung, colorectal, and ovarian cancers.

Both natural history and screening considerations come into play in such questions as the interpretation of secular changes in incidence and mortality. According to the NCI SEER (Surveillance, Epidemiology and End Results) Program, newly diagnosed cases of breast cancer increased between 1950 and 1979 at an annual rate of 1%, and between 1980 and 1984 at an annual rate of 3% (Breast cancer incidence is on the rise – but why? *JNCI* June 20, 1990; 82(12):998-1000). There has also been a "dramatic" upsurge in *in situ* breast cancer diagnosed since 1983. Breast cancer mortality overall was stable in the 1970s and began to fluctuate in the mid-1980s. Are the observed changes due to increased use of mammography? In support of that interpretation is the fact that among white women age 50 and older, localized disease has increased (i.e., a shift in the stage distribution) during the 1980s. There has also been a rapid increase in sales and installation of new mammography units during the 1980s, and the number of mammograms has risen dramatically. Or, could the observed changes be due to changes in risk factors (e.g., oral contraceptives, alcohol consumption, diet)? The observation of a striking increase in estrogen-receptor positive cancers suggests some biological change has occurred.

Another cancer where issues of natural history and early detection are of great importance is cancer of the prostate. The substantial (e.g., around 30% in men age 50 years and older) prevalence of previously undetected prostate cancer found at autopsy has demonstrated that many more men die with prostate cancer than from prostate cancer. Although "indolent" prostate cancers have the pathological features of cancer, if their growth is so slow that they will never become clinically manifest, should they be considered as the same disease as cancers of clinical importance? In addition, the lengthy natural history of most prostate cancers raises the concerns of lead time bias, length bias, and overdiagnosis for any observational approach to evaluating the efficacy of screening for early prostate cancer. In addition, there are major questions about the effectiveness of both existing modes of treatment and existing modes of early detection. Prostate cancer incidence **doubled** from 90 per 100,000 in 1985 to 185 per 100,000 in 1992, undoubtedly as a result of the dissemination of prostatic-specific antigen (PSA) screening. Meanwhile, prostate cancer mortality has decreased, though more modestly. These trends are consistent with the claim that screening with PSA reduces mortality, though the issue remains controversial for a number of reasons.

Bibliography

Lilienfeld and Lilienfeld – *Foundations of epidemiology*, Chapters 3-7; MacMahon & Pugh – *Epidemiology principles and methods*, Chapters 4, 7-10; Mausner & Kramer – *Epidemiology: an introductory text*, Chapter 1-2, 6. Kelsey, Thompson & Evans – *Methods in observational epidemiology*, pp. 23-31 and 46-53.

Chorba, Terence L.; Ruth L. Berkelman, Susan K. Safford, Norma P. Gibbs, Harry F. Hull. Mandatory reporting of infectious diseases by clinicians. *MMWR* 1990 (June 20); 39(9):1-6.

Cole, Phillip; Alan S. Morrison. Basic issues in population screening for cancer. *JNCI* 1980; 64:1263-1272.

Dubos, Rene. *Man adapting.* New Haven, CT, Yale, 1965.

Goodman, Richard A.; Ruth L. Berkelman. Physicians, vital statistics, and disease reporting. Editorial. *JAMA* 1987; 258:379-380.

Halloran, M. Elizabeth. Concepts of infectious disease epidemiology. In: Rothman and Greenland, *Modern Epidemiology* 2ed, Philadelphia, Lippincott-Raven, 1998, ch 27.

Israel, Robert A.; Harry M. Rosenberg, Lester R. Curtin. Analytical potential for multiple cause-of-death data. *Am J Epidemiol* 1986; 124:161-179. See also: Comstock, George W.; Robert E. Markush. Further comments on problems in death certification. 180-181.

Jacob KS. The question for the etiology of mental disorders. *J Clin Epidemiol* 1994;47:97-99.

Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA* 1998; 280:42-48.

Kircher, Tobias; Robert E. Anderson. Cause of death: proper completion of the death certificate. *JAMA* 1987;258:349-352.

Kircher T, Nelson J, Burdo H. The autopsy as a measure of accuracy of the death certificate. *N Engl J Med* 1985; 313:1263-9.

Lindahl, B.I.B; E. Glattre, R. Lahti, G. Magnusson, and J. Mosbech. The WHO principles for registering causes of death: suggestions for improvement. *J Clin Epidemiol* 1990; 43:467-474.

Mirowsky, John and Catherine E. Ross. Psychiatric diagnosis as reified measurement. *J Health and Social Behavior* 1989 (March): 30:11-25 plus comments by Gerald L. Klerman (26-32); Marvin Swartz, Bernard Carroll, and Dan Blazer (33-34); Dan L. Tweed and Linda K. George (35-37); and rejoinder by Mirowsky and Ross (38-40).

Morrison, Alan S. *Screening in chronic disease*. NY, Oxford, 1985. (His chapter on screening in Rothman and Greenland is an excellent succinct presentation of concepts related to screening programs and their evaluation.)

National Vital Statistics System, National Center for Health Statistics, CDC. Various handbooks on completing death certificates. http://www.cdc.gov/nchs/about/major/dvs/handbk.htm

Percy, Constance; Calum Muir. The international comparability of cancer mortality data. *Am J Epidemiol* 1989; 129:934-46.

Percy, Constance; Edward Stanek III, Lynn Gloeckler. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *Am J Public Health* 1981; 71:242-250.

Rothman, Kenneth J. Induction and latent periods. *Am J Epidemiol* 1981; 114:253-9.

Sorlie, Paul D., Ellen B. Gold. The effect of physician terminology preference on coronary heart disease mortality: an artifact uncovered by the 9th Revision ICD. *Am J Public Health* 1987; 77:148-152.

Stallones, Reuel A. The rise and fall of ischemic heart disease. Scientific American 1980; 243:53-59.

Stanbridge, Eric J. Identifying tumor suppressor genes in human colorectal cancer. *Science* 5 Jan 1990; 247:12-13)

World Health Organization. *Manual of the international statistical classification of diseases, injuries, and causes of death*. Based on recommendations of the Seventh Revision Conference, 1955. Vol I. World Health Organization, Geneva, 1957.

5. Measuring Disease and Exposure

Descriptive statistics; measuring occurrence and extent of disease; prevalence, incidence (as a proportion and as a rate), and survivorship; weighted averages, exponents, and logarithms.

> "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of Science, whatever the matter may be."
>
> Lord Kelvin (quoted in Kenneth Rothman, *Modern Perspectives in Epidemiology*, 1 ed. Boston, Little Brown, 1986, pg 23)

At the beginning of this text we noted four key aspects of epidemiology: its multidisciplinary nature, and its concern with populations, measurement, and comparison. As all empirical scientistis, epidemiologists devote a great deal of attention to issues of measurement – the application of numbers to phenomena. Every object of study – a disease, an exposure, an event, a condition – must be defined and measured. Since epidemiology deals with populations, epidemiologists need methods to describe and summarize across populations. This chapter discusses various aspects of measurement, including the definition, computation, and interpretation of key measures of health events and states in populations. The next chapter deals with comparisons between these measures.

Numeracy: applying numbers to phenomena

Numeracy is the concept of summarizing phenomena quantitatively. Faced with an infinitely detailed and complex reality, the researcher attempts to identify and quantify the meaningful aspects. Two of the innumerable examples of this process in epidemiology are:

Atherosclerosis score: David Freedman, an epidemiologist who received his doctoral degree from UNC, conducted his dissertation research on the relationship of atherosclerosis in patients undergoing coronary angiography to plasma levels of homocysteine. A basic question he had to address was how to measure atherosclerosis in coronary angiograms. Should he classify patients as having a clinically significant obstruction, count the number of obstructions, or attempt to score the extent of atherosclerosis? An atherosclerosis score would capture the most information and could provide a better representation of the phenomenon as it might be affected by homocysteine levels. But should an atherosclerosis score measure surface area of involvement or extent of narrowing? How should it treat lesions distal to an occlusion, which have no effect on blood flow? These and other decisions would need to depend upon his conceptual model of how homocysteine would affect the endothelium. For example, would homocysteine be involved primarily in causing initial damage, in which case the total surface area involved would be relevant, or would it be involved in the progression of atherosclerosis, in which case the extent of narrowing would

be relevant. Compromises might be forced by limitations in what measurements could be made from the angiograms.

Measuring smoking cessation: at first glance, smoking cessation, in a study of the effects of smoking cessation or of the effectiveness of a smoking cessation program, would seem to be straightforward to define and measure. Even here, though, various questions arise. The health benefits from cessation may require abstinence for an extended period (e.g., years). However, biochemical validation techniques, considered necessary when participants would have a reason to exaggerate their quitting success, can detect smoking during a limited period of time (e.g., about seven days for salivary cotinine). Should cessation be defined as no tobacco use for 7 days, to facilitate validation, or for at least a year, when the relapse rate is much lower?

Conceptual models underlie measures

In general, how we apply numbers and what type of measures we construct depend upon:

1. the purpose of the measure
2. the nature of the data available to us.
3. our conceptualization of the phenomenon

These three factors will pervade the types of measures to be covered.

Ideally we would like to watch phenomena unfold over time. In practice we must often take a few measurements and infer the rest of the process. Conceptual models pervade both the process of applying numbers to phenomena and the process of statistically analyzing the resulting data in order to identify patterns and relationships. Not being able to record all aspects of phenomena of interest, we must identify those aspects that are biologically, psychologically, or otherwise epidemiologically important. These aspects are embodied in operational definitions and classifications. The method by which we apply numbers and analyze them must preserve the important features while not overburdening us with superfluous information. This basic concept holds for data on individuals (the usual unit of observation in epidemiology) and on populations. Although we employ mathematical and statistical models as frameworks for organizing the resulting numbers, for estimating key measures and parameters, and for examining relationships, conceptual models guide all of these actions.

Levels of measurement

One area where objectives, availability of data, and conceptual models come to bear is the **level of measurement** for a specific phenomenon or construct. Consider the construct of educational attainment, a variable that is ubiquitous in epidemiologic research. We can (1) classify people as being or not being high school graduates; (2) classify them into multiple categories (less than high school, high school graduate, GED, trade school, technical school, college, professional degree, graduate degree); (3) record the highest grade in school they have completed; or (4) record their scores on standardized tests, which we may need to administer.

The first alternative listed illustrates the most basic "measurement" we can make: a **dichotomous** (two category) classification. People can be classified as "cases" or "noncases", "exposed" or "unexposed", male or female, etc. Communities can be classified as having a mandatory seat-belt law or not, as having a needle exchange program or not, etc.

Potentially more informative is a **polytomous** (more than two categories) classification, such as country of origin, religious preference, ABO blood group, or tumor histology (e.g., squamous cell, oat cell, adenocarcinoma). A polytomous classification can be **nominal** – naming categories but not rank ordering them, as is the case for the four examples just given – or **ordinal**, where the values or categories can be rank-ordered along some dimension. For example, we might classify patients as "non-cases", "possible cases" "definite cases" or injuries as minimal, moderate, severe, and fatal.

The values of the different levels of a nominal variable provide no information beyond identifying that level, and so they can be interchanged without constraint. We can code squamous cell "1", oat cell "2", and adenocarcinoma "3"; or instead, squamous cell "2" and oat cell "1" or even "5"). The numbers simply serve as names. The values of the different levels of an ordinal variable signify the ranking of the levels. The values can be changed, but generally not interchanged. We can use "1", "2", "3", respectively, for non-case, possible case, and definite case, or we can use "1" "3" "8", but we can not use "1" "3" "2", since this coding would not preserve the ordering.

When the values themselves, or at least the size of the intervals between them, convey information, then the phenomenon has been measured at the **interval** level. Temperature measured on the Fahrenheit scale is an interval scale, since although we can say that 80°F is twice 40°F, the ratio is not meaningful in terms of the underlying phenomenon. Psychological scales are often regarded as being interval scales. What differentiates an interval scale from most of the measures we use in physical sciences is the absence of a fixed **zero point**. Since only the intervals convey meaning, the scale can be shifted up or down without changing its meaning. An interval scale with values "1", "1.5", "2", "3", "4" could just as well be coded "24", "24.5", "25", "26", "27".

A **ratio** scale, however, has a non-arbitrary zero point, so that both intervals and ratios have meaning. Most physical measurements (height, blood pressure) are ratio scales. The values of an ratio scale can be multiplied or divided by a constant, as in a change of units, since comparisons of intervals and ratios are not distorted. If value B is twice value A before multiplication, it will still be twice value A afterwards. A ratio scale with values "1", "1.5", "2", "3", "4" can be transformed to "2", "3", "4", "6", "8" (with appropriate substitution of units), but not as "2", "2.5", "3", "4", "5", since only intervals but not ratios will be preserved.

One type of ratio scale is a **count**, such as birth order or parity. A count is a **discrete** variable, because its possible values can be enumerated. A **continuous** variable, as defined in mathematics, can take on any value within the possible range, and an infinitude of values between any two values. Measurements in epidemiology are no where nearly as precise as in the physical sciences, but many measurements used in epidemiology have a large enough number of possible values to be treated as if they were continuous (e.g., height, weight, or blood pressure).

Whether continuous or discrete, however, both interval and ratio scales generally imply a linear relationship between the numerical values and the construct being measured. Thus, if we measure

educational attainment by the number of years of school completed, we are implying that the increase from 10th grade to 11th grade is the same as the increase from 11th grade to 12th grade, even though the latter usually conveys a high school diploma. We are also implying that completing 12th grade with three advance-placement or honors classes in a high-achievement school is the same as completing 12th grade with remedial courses in a low-achievement school, or as completing 12th grade but reading at only a 9th grade level, or completing 12th grade but without taking any mathematics beyond elementary algebra, etc., not to mention ignoring the educational aspects of travel, speaking multiple languages, or having learned a trade. Even chronological age may not be an interval or ratio scale when certain ages have special meaning (e.g., 16 years, 18 years, 21 years, 40 years, 65 years). Many measures that appear to be interval or ratio scales may not really behave as such, due to **threshold effects** (differences among low values have no real significance), **saturation effects** (differences among high values have no real significance), and other **nonlinearities**.

Absolute and relative measures — the importance of a denominator

While the absolute values of age, educational attainment, blood pressure, and cigarettes/day are meaningful, other measures are expressed as concentrations (e.g., 20 μg of lead per deciliter of blood, 500 T-cells per cubic centimeter of blood, 1.3 persons per room, 392 persons/square kilometer) or **relative** to some other dimension (e.g., body mass index [weight/height^2], percent of calories from fat, ratio of total cholesterol to HDL cholesterol). Most population-level measures are not meaningful unless they are relative to the size and characteristics of a population and/or to expected values, even if only implicitly. Other than a report of cases of small pox, since the disease has now been eradicated world wide, how else can we assess whether a number of cases represents an outbreak or even an epidemic? For this reason epidemiologists often refer disparagingly to absolute numbers of cases or deaths as "numerator data". Exceptions illustrate the general principle. A handful of cases of angiosarcoma of the liver in one manufacturing plant led to an investigation that uncovered this hazard from vinyl chloride. A handful of cases of adenocarcinoma of the vagina in teenage women in one hospital led to the identification of the effect of diethylstilbesterol (DES) on this disease. A handful of cases of acquired immunodeficiency syndrome (AIDS) alerted public health to the start of this pandemic. Since these were very rare or previously unobserved conditions, an expectation was already defined.

Types of ratios

As illustrated with several of the above examples, we express a quantity relative to another by forming a ratio, which is simply the quotient of two numbers, a numerator divided by a denominator. Ratios are ubiquitous in epidemiology, since they enable the number of cases to be expressed relative to their source population.

Two special classes of ratios in epidemiology are proportions and rates. **Proportions** are ratios in which the numerator is "contained in" or "part of" the denominator. The statement that 12% of the population is age 65 or above expresses a proportion, since people age 65 and above are a fractional component of the population. Because the numerator is a fractional component of the denominator, a proportion can range only between 0 and 1, inclusive. Proportions are often expressed as percentages, but any **scaling factor** can be used to yield a number that is easier to express. For example, the proportion 0.00055 would often be expressed as 5.5 per 10,000 or 55 per

100,000. Note that the ratio of abortions to live births, although of the same order of magnitude, is *not* a proportion, since the numerator is not contained in the denominator.

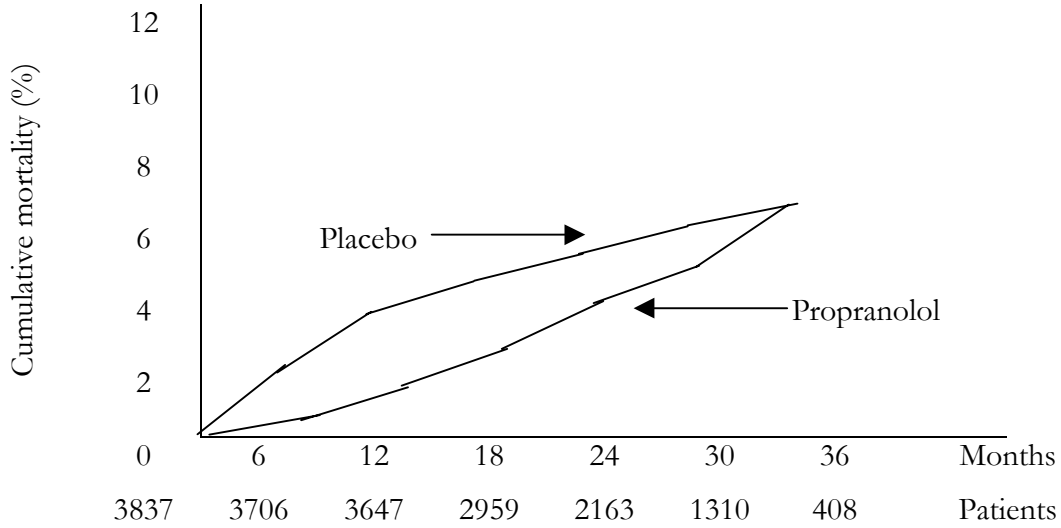
Although many types of ratios (including proportions) are frequently referred to as "rates", in its precise usage a **rate** is the ratio of a change in one quantity to a change in another quantity, with the denominator quantity often being time (Elandt-Johnson, 1975). A classic example of a rate is velocity, which is a change in location divided by a change in time. Birth rates, death rates, and disease rates are examples if we consider events — births, deaths, newly diagnosed cases — as representing a "change" in a "quantity". Rates can be absolute or relative, according to whether the numerator is itself a ratio that expresses the change relative to some denominator. Most rates in epidemiology are relative rates, since as discussed above the number of cases or events must generally be related to the size of the source population.

"Capturing the phenomenon"

All measures, of course, are summaries or indicators of a complex reality. The question always is, "does the measure capture what is important about the phenomenon given our objective?". This principle applies at both the individual level (for example, when can a person's constantly-varying blood pressure and heart rate be meaningfully represented by single numbers) and population level.

For example, although the proportion of a group of patients who survive for 5 years is a measure of treatment effectiveness, if the proportion is low then when deaths occur is especially important. The statement that the "five-year survival rate following coronary bypass surgery was 60%" does not tell us whether the 40% who died did so during the procede, soon afterward, gradually during the period, or not until at least three years following surgery. When the **time-to-occurrence** of an event is important, then survivorship analysis is employed, such as in the following figure similar to that reported from the Beta-blocker Heart Attack Trial (BHAT), a double-blinded, randomized trial of propranolol to treat patients experiencing an acute myocardial infarctions.

Life table cumulative mortality in the Beta Blocker Heart Attack Trial (BHAT)

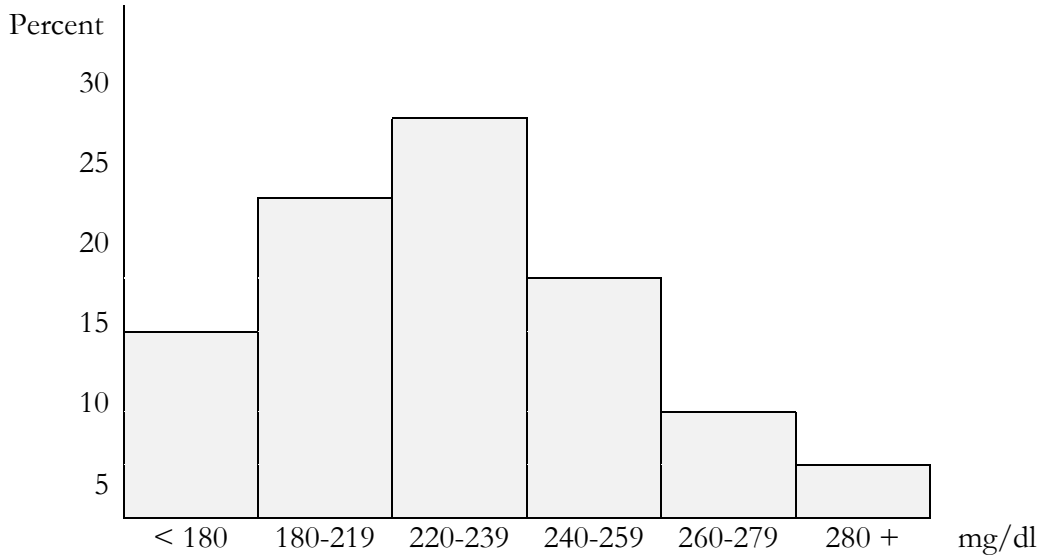


[Source: *JAMA*, March 26, 1982; 247:1707]

Distributions – the fuller picture

More generally, when the object of study involves not merely "presence" or "occurrence" but rather a **polytomous** or **measurement** variable, we should examine the full distribution, e.g.

Serum cholesterol levels - Distribution



Although distributions are informative, they are cumbersome to work with and to present. Therefore we try to "capture" the essential information about the distribution by using summary statistics, such as the mean, median, or quartiles, and the standard deviation or interquartile range (see below). While it is often essential to compress a distribution, curve, or more complex picture into a number or two, care must be taken that the necessary simplification does not distort the resulting computation, presentation, and interpretation. Indeed, it may be the persons at one end of the distribution who are most important or informative in respect to health consequences.

If the data are distributed in a familiar fashion, we can adequately characterize the entire distribution by its parameters (e.g., the mean and standard deviation for a "normal" [Gaussian] distribution). But it can be hazardous to assume that the data conform to any particular distribution without verifying that assumption by examining a histogram (e.g., see *Statistics for Clinicians*, Figure 7-7, for several distributions with identical mean and standard deviation but dramatically different appearance).

Common summary statistics for description and comparison

Mean – The "average" value of the variable

Median – The middle of the distribution of the variable – half of the values lie below and half lie above

Quartiles – The values that demarcate the 1st, 2nd, and 3rd quarter of the distribution of the variable [the median is the 2nd quartile]

Percentiles – The values that demarcate a percentage of the distribution, e.g., the 20th percentile (also called the second decile) is the value below which the lowest 20% of the observations fall.

Standard deviation – Roughly speaking, the distance of a typical observation from the mean of the distribution (more precisely, the square root of the average of the squared distances of observations from the mean) [Not to be confused with the **standard error**, which is a measure of the imprecision of an estimate.]

Interquartile range – The distance between the 1st and 3rd quartiles.

Skewedness – The degree of asymmetry about the mean value of a distribution. Positively skewed or right-skewed means that the distribution extends to the right; in a positively-skewed distribution, the mean (overall average) lies to the right of the median, due to the influence of the outlying values.

Kurtosis – The degree of peakedness of the distribution relative to the length and size of its tails. A highly peaked distribution is "leptokurtic"; a flat one is "platykurtic".

When interpreting summary statistics, it is important to consider whether the summary statistics represent the most relevant features of the distributions that underlie them. Several examples:

Community health promotion:

Suppose that surveys before and after a community alcohol control program find a reduction in mean alcohol consumption of 1 drink/day in the target population. That reduction could reflect either:

- a 5 drink/day reduction for each person in the highest consumption 20 percent of the population

or

- a 1.25 drink/day reduction for all people but those in the highest consumption 20%,

with very different implications for health.

Black-white differences in birth weight:

The distribution of birth weight has an approximate Gaussian ("normal") shape, with a range from about 500 grams (the lower limit of viability) to about 5,000 grams and a mean of about 3,000 grams. Statistically the distribution is smooth and reasonably symmetrical. However, the biological implications vary greatly across the distribution, since the majority of infant deaths occur for babies weighing less than 2,500 grams. For babies weighing 1,000-2,000 grams, the mortality rate is 33%; for babies weighing less than 1,000 grams, the mortality rate is 75%.

The birth weight distributions for Black and White Americans are generally similar, with that for Blacks shifted slightly to the left. But that slight shift to the left translates into a substantially greater proportion below 2,500g, where mortality rates are much higher.

Per capita income:

> Should health care resources for poor people be allocated on the basis of per capita income of counties? At least one study has found that the barriers to health care experienced by the poor in the U.S. appear to be similar in wealthy counties and in other counties, so that per capita income (i.e., mean income per person) is not as good a criterion for determining the need for public health care programs as is the number of poor persons in the area (Berk M, Cunningham P, Beauregard K. The health care of poor persons living in wealthy areas. *Social Science in Medicine* 1991;32(10):1097-1103).

The moral: in order to interpret a change or difference in a summary measure it is necessary to know something about the shape of the distribution and the relationship between the variable and the relevant health outcome.

Heterogeneity and distributions of unknown factors – any summary is a weighted average

Since populations differ in characteristics which affect health, an overall number, such as a proportion or mean, often conceals subgroups that differ meaningfully from the overall picture. Even when we cannot identify these subgroups, we should be mindful of their likely existence. Because most diseases vary across subgroups, epidemiologic measures are more interpretable with knowledge of the composition of the group they refer to, at least in terms of basic demographic characteristics (notably age, sex, geographical area, socioeconomic status, employment status, marital status, ethnicity) and important exposures (e.g., smoking).

> E.g., a workforce experiences 90 lung cancer deaths per l00,000 per year: To know what to make of this it is essential to know the age distribution of the workforce and if possible the distribution of smoking rates.

Virtually any measure in epidemiology can be thought of as a weighted average of the measures for component subgroups. We can use "specific" measures (e.g., "age-specific rates," "age-sex-specific rates") where the overall ("crude") measure is not sufficiently informative. Also, we can produce "adjusted" or "standardized" measures in which some standard weighting is used to facilitate comparisons across groups. Adjusted measures are typically weighted averages – the weights are key. The concept of weighted averages is fundamental and will resurface for various topics in epidemiology. (Rusty on weighted averages? See the Appendix on weighted averages.)

Types of epidemiologic measures

Purpose of the measure:

There are three major classes of epidemiologic measures according to the question or purpose. We use **measures of frequency or extent** to address questions such as "How much?", "How many?", "How often?", "How likely?", or "How risky?". We use **measures of association** to address questions about the strength of the relationship among different factors. We use **measures of impact** to address questions of "How important?".

Availability of data:

We can also categorize epidemiologic measures according to the type of data necessary to obtain them:

1. Measures derived from routine data collection systems, e.g., vital events registration, cancer registries, reporting of communicable diseases.

2. Measures derived from data collected in epidemiologic studies or for related purposes (e.g., clinical studies, health insurance records).

3. Measures derived from theoretical work in biometry - no data necessary! e.g., Risk of disease in exposed = Pr[D | E]

$$\text{Incidence density} = -\frac{d(N_t)}{N_t\, dt}$$

The usefulness of the third class of measures is in refining measurement concepts and in advancing understanding. Measures in the first two classes generally involve compromises between the theoretical ideal and practical reality. Epidemiology is fundamentally a practical field. In the rest of the chapter we will touch on the first class and then dwell on the second.

Measures derived from routinely collected data

In this area come the vital statistics data compiled by health authorities and statistical agencies, such as the World Health Organization, the U.S. National Center for Health Statistics, state health departments, and their counterparts in other countries. Examples of measures published from such data are:

- total death rates
- cause-specific death rates
- birth rates (births per 1,000 population)
- infant mortality rates
- abortion/live birth ratio
- maternal mortality rate

[See Mausner and Kramer, ch 5; Remington and Schork, ch 13.]

The denominator for vital statistics and other population-based rates (e.g., death rates, birth rates, marriage rates) is generally taken from population estimates from the national census or from other vital events data, as in the case of the infant mortality rate:

$$\text{Infant mortality rate} = \frac{\text{Deaths of children < 1 year of age in one year}}{\text{Total live births in one year}}$$

Results are usually scaled so that they can be expressed without decimals (e.g., 40 deaths per 1,000 or 4,000 deaths per 100,000).

Optional aside – Assessing precision of an estimated rate, difference in rates, or ratio of vital statistics rates

If r is a rate (e.g., an infant mortality rate) and n is the denominator for that rate (e.g., number of live births), then a 95% confidence interval for r can be constructed using the formula:

$$r \pm 1.96 \times \sqrt{(r/n)}$$

E.g., in an area with 30 infant deaths and 1,000 live births, r = 30/1,000 = 30 per 1,000 or 0.03. The 95% confidence interval for r is:

$$0.03 \pm 1.96 \times \sqrt{(0.03/1{,}000)} = 0.03 \pm 0.0107 = (0.0193, 0.0407),$$

or between 19.3 and 40.7 per thousand

The 95% confidence interval for the difference, D, between two rates, r_1 and r_2, based, respectively, on number of deaths d_1 and d_2, and denominators n_1 and n_2, is:

$$(r_1 - r_2) \pm 1.96 \times \sqrt{(r_1/n_1 + r_2/n_2)}$$

The 95% confidence interval for the ratio, R, of r_1 and r_2 is:

$$R \pm R \times 1.96 \times \sqrt{(1/d_1 + 1/d_2)}$$

where d_2 (the number of deaths for the denominator rate) is at least 100.

Source: Joel C. Kleinman. Infant mortality. Centers for Disease Control. National Center for Health Statistics *Statistical Notes*, Winter 1991;1(2):1-11.

The basis for the above can be stated as follows. The number of rare events in a large population can often be described by the Poisson distribution, which has the notable feature that its mean is the same as its variance. For a Poisson distribution with mean d (and variance d), if the number of events is sufficiently large (e.g., 30), then 95% of the distribution will lie within the interval $d \pm 1.96\sqrt{d}$. If we divide this expression by the population size (n), we obtain the 95% confidence interval for the rate as:

$$d/n \pm (\sqrt{d})/n = r \pm \sqrt{(r/n)}$$

Reporting systems and registries for specific diseases, hospital admissions, and ambulatory care visits provide data on incidence or health care utilization for some conditions. Communicable diseases have long been reportable, though the completeness of reporting is quite variable. Major investments in state cancer registries are creating the basis for a national cancer registry system in the

U.S. Several states have reporting systems for automobile collisions. For the most part, however, data on non-fatal disease events are less available and complete than mortality data.

Remember: All rates, ratios, and other measures can be:

Specific to a group defined by age, sex, and/or other factors.

Adjusted for age, sex, or other relevant variable(s);

Crude (i.e., neither specific nor adjusted).

These terms apply with respect to particular variable(s) and are therefore not mutually exclusive. For example, a rate can be adjusted with respect to age, specific with respect to gender, and crude with respect to ethnicity, geographical region, etc. (e.g., an age-adjusted rate for women of all ethnicities and all geographical regions).

The basic concept underlying adjustment procedures is that of the **weighted average**. The limitations of adjusted measures derive from this aspect – validity of comparison depends upon the similarity of the component weights; validity of interpretation depends upon the numerical and conceptual homogeneity of the component specific measures.

Measures derived from data collected in epidemiologic studies

For most epidemiologic studies, routinely collected data are not adequate, so data must be collected specifically for the study purposes. The reward for the time, effort, and expense is a greater opportunity to estimate measures that are more suited for etiologic and other inferences. Three principal such measures are prevalence, incidence, and case fatality.

Prevalence – the proportion of cases within a population
$$\text{Prevalence} = \frac{\text{Cases}}{\text{Population-at-risk}}$$

Prevalence – a kind of "still life" picture – is the most basic of epidemiologic measures. It is defined as the number of cases divided by the population-at-risk. Note that:

- Prevalence is a proportion, so must lie between 0 and 1, inclusive.
- Population at risk (PAR) means "eligible to have the condition".
- Prevalence can be used to estimate the probability that a person selected at random from the PAR has the disease [Pr(D)]

Example:

$$\text{Prevalence} = \frac{\text{No. of persons with senile dementia at a given time}}{\text{No. in study population at risk for senile dementia}}$$

$$= \frac{175}{1{,}750} = 0.10 = 10\%$$

Optional aside – Assessing precision of an estimated prevalence.

Since prevalence is a proportion, a confidence interval can be obtained using the binomial distribution or, where there are at least 5 cases, the normal approximation to the binomial distribution. The variance of a point binomial random variable is pq (where p is the probability of a "success" and q=1—p), so the standard error for the estimated probability is $\sqrt{(pq/n)}$. Thus, the 95% confidence interval for a prevalence estimate p is: $p \pm 1.96 \sqrt{[p(1—p)/n]}$. For the preceding example, the 95% confidence limits are $0.10 \pm 1.96\sqrt{[(0.10)(0.90)/1750]} = (0.086, 0.114)$. When there are fewer than 5 cases, an exact procedure is required.

Prevalence has three components:

1. Existing cases
2. Population "at risk" to have the condition
3. Point (or sometimes a period) in time to which the prevalence applies

Incidence – the occurrence of new cases
$\text{Incidence} = \frac{\textit{New}\text{ cases}}{\text{Population-at-risk } \textit{over time}}$

Incidence – a "motion picture" – describes what is happening in a population. Incidence is defined as the number of new cases divided by the population at risk over time. Incidence therefore includes three components:

1. New cases
2. Population at risk.
3. Interval of time.

Note that:

- Incidence involves the passage of time.

- Incidence may be expressed as a proportion or as a rate.
- Incidence can be used to estimate the risk of an event during a stated period of time.

Example:

$$\text{e.g., Cumulative incidence} = \frac{\text{New cases of senile dementia in 5 years}}{\text{No. of persons at risk}}$$

In infectious disease epidemiology, this measure is often termed the **attack rate** or **secondary attack rate**, especially when referring to the proportion of new cases among contacts of a primary case.

Case fatality is a measure of the severity of a disease. Though often called the case fatality "rate", the measure is generally computed as a proportion:

Case fatality – proportion of cases who die
$\text{5-year case fatality} = \frac{\text{Deaths from a condition}}{\text{Number of persons with the condition}}$

If the time period under discussion does not encompass the entire period of risk of death from the condition, then the time period must be stated explicitly or the statistic is uninterpretable. The case fatality rate for AIDS increases with every year following diagnosis, but that for an episode of influenza or for a surgical procedure does not change after a month or so.

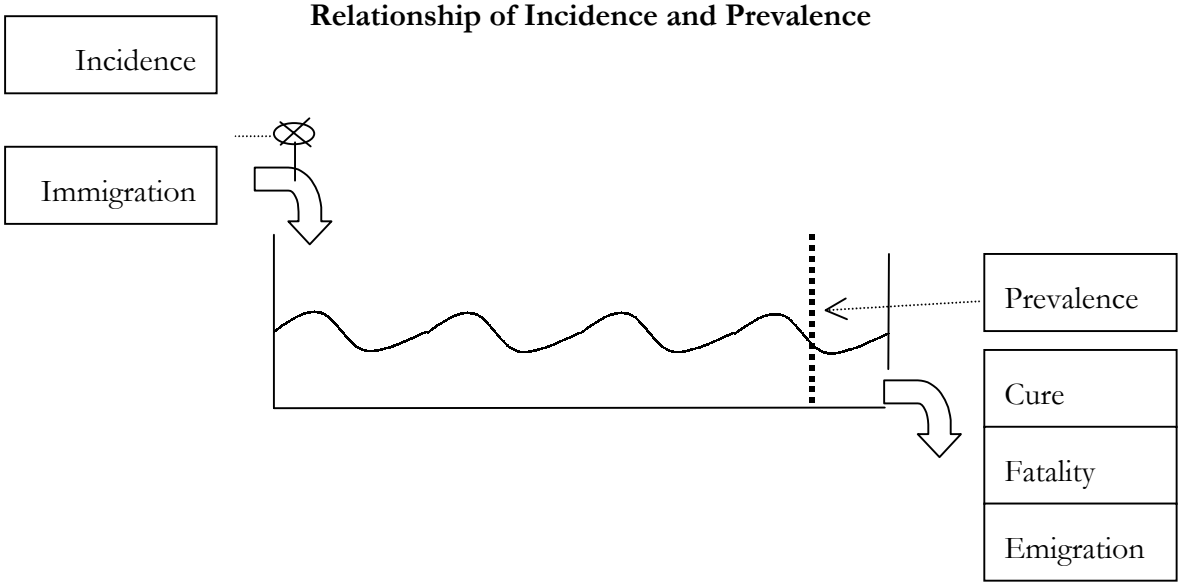
Example:

$$\text{Case fatality rate} = \frac{\text{Deaths from senile dementia in 5 years}}{\text{Number of persons diagnosed with senile dementia}}$$

Relationship of incidence and prevalence

Incidence, mortality, and prevalence are intimately related, of course, just as are births, deaths and population size. Demographers study the latter phenomena, and their techniques are used in epidemiology (under other names, naturally, to "protect the innocent"!).

Relationship of Incidence and Prevalence



In a stationary population, in which there is no migration of cases or noncases, if the incidence, prevalence, and duration of a condition remain constant then the number of new cases that occur must be balanced by the number of existing cases that leave the population through death or cure. In such a situation, the prevalence is a function of incidence and the average duration of being a case. For a rare disease, prevalence ≈ incidence × duration (see "Incidence and prevalence in a population", below).

Influences on the relation of incidence and prevalence

The relationships among incidence, mortality, and prevalence are affected by such factors as:

Virulence of the disease - Is it rapidly fatal?

Health care - When do cases come to medical attention?

Can cases be cured?

Does earlier detection alter prognosis?

Behavior - Do people recognize and act promptly on symptoms?

Do patients comply with treatment?

Competing causes of death - Are people with the disease likely to die of other causes?

Migration - Are people with the disease likely to leave the area?

Are people with the disease like to migrate to the area?

Because prevalence is affected by factors (e.g., duration and migration) that do not affect the development or detection of a disease or condition, measures of incidence are generally preferred over measures of prevalence for studying etiology and/or prevention. Both incidence and prevalence are useful for various other purposes (surveillance and disease control, health care

planning). Also, prevalence may be more readily estimated than incidence and may be looked to for etiologic inferences despite its limitations.

It is important to note, however, that although incidence itself is not affected by factors unrelated to etiology, observed incidence reflects the influence of a variety of nonetiologic factors (how quickly the disease produces symptoms that prompt a health care visit, access to health care, whether the health care provider selects the correct diagnostic maneuver, accuracy of the exam result and its interpretation, and accuracy and promptness of reporting). There are, accordingly, great difficulties in interpreting reported incidence of many diseases and conditions (e.g., Alzheimer's disease, AIDS, HIV, other sexually transmitted infections, Lyme disease, and prostate cancer, to name but a few).

An example of how disease natural history distorted trends in observed incidence comes from the early years of the AIDS epidemic, when AIDS case reporting was the primary means of tracking the HIV epidemic. Due to the considerable variability in the time between HIV infection and development of opportunistic infections signaling the onset of AIDS, the upward trend in AIDS cases exaggerated the upward trend in HIV infections. The mechanism for this effect can be illustrated as follows. Suppose that the numbers of new HIV infections during the first four years of the epidemic were 500, 1,000, 1,500, 2000, respectively, indicating a linear increase of 500/year. Suppose that 5% of HIV infections progress to AIDS during each year following infection, for a median time-to-AIDS of 10 years. During the first year 25 cases of AIDS will occur (5% of 500 infections). During the second year 75 cases of AIDS will occur (5% of 500 plus 5% of 1,000). During the third year 150 cases of AIDS will occur (5% of 500 plus 5% of 1,000 plus 5% of 1,500). During the fourth year 250 cases of AIDS will occur, so the trend in AIDS (25, 75, 150, 250) will initially appear to increase more steeply than the trend in HIV (HIV infections double in year 2, but AIDS cases triple) and then will appear to level off despite no change in the HIV incidence trend. There will also be a change in the ratio of AIDS to HIV, as also occurred during the early years of the epidemic. (The phenomenon was described in an article in the *American Journal of Epidemiology* in about 1987; I am looking for the citation.)

Prevalence versus incidence

	Prevalence	Incidence
Cases	Entities	Events
Source population (PAR)	At risk to be a case	At risk to become a case
Time	Static (point)	Dynamic (interval)
Uses	Planning	Etiologic research

Considerations relevant for both prevalence and incidence

Cases

1. **Case definition** – What is a case?

Examples: arthritis, cholelithiasis, cardiovascular disease, diabetes, psychiatric disorder, epidemiologic treatment of syphilis or gonorrhea, prostate cancer

2. **Case development** – When is a case?

   Issues: induction, latency, progression, reversibility

   Examples: atherosclerosis, cancer, cholelithiasis, diabetes, hypertension, AIDS

3. **Case detection** – When is a case a "case"?

   Issues: Detectability is a function of technology and feasibility. What can be detected is not the same as what is detected.

   Examples: Atherosclerosis, breast cancer, cholelithiasis, osteoporosis, asymptomatic infections, prostate cancer

Source population [Population at risk (PAR)]

1. What is the relevant population — who is really "at risk"?
   E.g., age (most diseases), sex (breast cancer), STD's and sexual activity, uterine cancer and hysterectomy, gallbladder cancer and cholecystectomy, genotypes?
2. What about previous manifestations?

   Of the same disease? (influenza, tumors, injuries)

   Of a related disease (stroke after CHD, cancer at a different site)
3. What about death from other causes? (competing risks)
   E.g., deaths for diabetes reduce the rate of death from coronary artery disease, heart disease deaths reduce the rate of death from lung cancer to the extent that smokers are at excess risk for both

Choosing the right denominator

The choice of the most appropriate denominator can be complex. For example, what is the most appropriate denominator for motor vehicular injuries or deaths?

Total population?

Population age 16 years and above?

Licensed drivers?

Registered vehicles?

Vehicle miles?

Passenger miles?

Which one to choose depends upon whether the question of interest concerns:

Injury risk by age and/or sex (population denominator?)

Effect on risk of seat-belt use (passenger-miles?)

Effect on deaths of 55 mph limit (passenger-miles?)

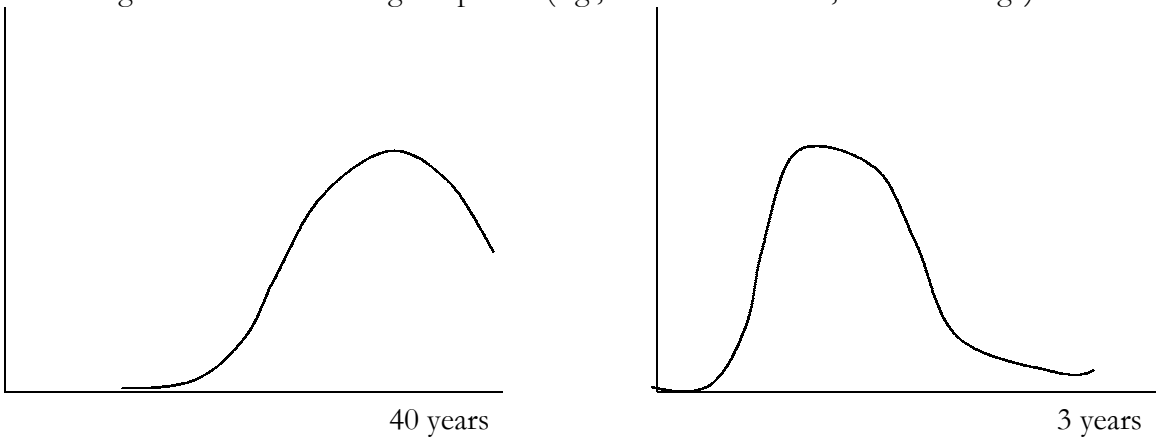
Role of alcohol in motor vehicular fatalities

Evaluation of alternate transportation policies

For example, older drivers have a higher crash rate per 100 million vehicle miles traveled than teen drivers do. But the rate of crashes per licensed driver is no higher for older drivers, because older drivers limit their driving.

Passage of time [incidence only] – what period of observation?

1. Natural history of the disease - period of risk versus period of observation
   E.g., atom bomb survivors and solid tumors, motor vehicle injury, congenital malformations
2. Different periods of observation for different subjects (does 1 person observed for 2 years = 2 people observed 1 year?)
3. Changes in incidence during the period (e.g., seasonal variation, secular change)



Cancer in atomic bomb survivors

Congenital malformations

Types of source populations for incidence

Source populations can be defined in various ways, including residence in a geographical area, employment in a company or industry, attendance in a school or university, membership in an organization, seeking health care from a given set of providers, or explicit recruitment into a study. Incidence involves the passage of time and therefore implies some type of follow-up of population. A key characteristics of a source population is in what ways its membership can change over time. Rothman and Greenland (1998) present a detailed discussion of types of populations and terminology that has been used to describe these. The primary distinction we will make here is that between a **fixed cohort**, whose membership changes only through attrition, and a **dynamic population** (Rothman and Greenland call this an **open cohort**), whose membership can change in various ways. (The fixed cohort versus dynamic population terminology come from Ollie Miettinen by way of Kleinbaum, Kupper, and Morgenstern.)

Cohort – entrance into the population is defined on the basis of some aspect or event in the lives of members of the study population (e.g., living in a geographical area when a major environmental event occurred, start of employment in a worksite or industry, receipt of a medical or surgical treatment, onset of a condition, start of an exposure, or simply enrollment into a study). Exits from the cohort (from death, out-migration, dropout) are problematic; entrances into the cohort are permitted only in relation to the qualifying event that defines the start of follow-up for that person. Note that once recruitment has been completed a cohort will become smaller over time due to attrition, and the entire age distribution will become older.

Variants:

Retrospective or historical cohort - the population is defined at some time in the past (e.g., based on employment records) and then followed forward in time towards the present by the use of available records.

"Dynamic cohort" – follow-up time is counted from the time of entrance into the study or in relation to some event that occurs at different times for different people (e.g., a medical procedure), so that accrual to the cohort continues over a period of time. In a classic cohort study, follow-up time for each subject and calendar time are identical; in a dynamic cohort, each participant's follow-up time may take place over a different interval of calendar time (this does not appear to be a widely-used term).

Dynamic population – a population is defined over a period of time and their experience is monitored during that period. The study population may be defined in the same way (e.g., geographical residence, employment, membership, etc.). In a dynamic population, however, both entrances and exits are expected and accommodated. For example, the population of a geographical area will experience births, deaths, and possibly substantial migration. Over time, a dynamic population can increase or decrease in size, and its age distribution can change or remain the same.

Special case:

A dynamic population is said to be **stable** or **stationary** when its size and age distribution do not change over time. The assumption of stationarity is often made, since it greatly simplifies analysis. (See Rothman and Greenland, 1998 for more on this.)

Types of incidence measures: cumulative incidence (incidence proportion) and incidence density (incidence rate)

There are two major types of incidence measures, differing primarily in the way in which they construct the denominator: **cumulative incidence** and **incidence density** (again, this is Olli Miettinen's terminology, adopted by Kleinbaum, Kupper, and Morgenstern; Rothman and Greenland use **incidence proportion** and **incidence rate**, respectively). Cumulative incidence (CI) is simply the proportion of a population that experience an event or develop a condition during a stated period of time. Incidence density (ID) is the rate at which new cases develop in a population, relative to the size of that population.

Cumulative incidence (incidence proportion)

$$CI = \frac{\text{New cases during stated period}}{\text{Number of persons at risk}}$$

Incidence density (Incidence rate)

$$ID = \frac{\text{New cases during stated period}}{\text{Population-time}}$$

Cumulative incidence (CI), a.k.a. Incidence proportion (IP)

The definition of CI is based on the following "ideal" scenario:

1. A population known to be free of the outcome is identified at a point in time (a cohort);
2. All members of the cohort are at risk of experiencing the event or outcome (at least once) for the entire period of time;
3. All first events or outcomes for each person are detected.

For example, consider a study of the risk that a rookie police officer will suffer a handgun injury during his first six months on patrol duties. Data are collected for a cohort of 1,000 newly-trained police officers entering patrol duties with the San Francisco Police Department (SFPD). During their first six months with the SFPD, 33 of the officers suffer a handgun injury. The other 967 officers have carried out patrol duties during the six-month period with no handgun injuries. The 6-months CI of handgun injury is 33/1,000 = 0.033. We use this observed CI to estimate the six-month risk of handgun injury to new patrol officers in San Francisco.

This example conforms to the ideal scenario for CI: there is a population "at risk" and "in view" for the entire period, and all first events were known. For the moment we assume away all of the reasons that might result in a member of the cohort not remaining "at risk" (e.g., transfer to a desk job, extended sick leave, quitting the force) and "in view" (e.g., hired by another police department).

Some things to note about CI:

1. The period of time must be stated (e.g., "5-year CI") or be clear from the context (e.g., acute illness following exposure to contaminated food source);
2. Since CI is a proportion, logically each person can be counted as a case only once, even if she or he experiences more than one event;
3. As a proportion, CI can range only between 0 and 1 (inclusive), which is one reason it can be used to directly estimate risk (the probability of an event).

Sample calculation:

200 people free of chronic disease X observed over 3 years

10 cases of X develop

3-year CI = 10 cases / 200 people = 10/200 = .05

Thus, the 3-year risk for one of the 200 people to develop disease X, conditional on not dying from another cause, is estimated as 0.05 or 5%.

> **Optional aside** – Assessing precision of an estimated cumulative incidence
>
> Since cumulative incidence is a proportion, a confidence interval can be obtained in the same manner as for prevalence (see above).

Risk and odds

In epidemiology, the term "risk" is generally taken to mean the probability that an event will occur in a given stated or implicit time interval (be alert for other uses, though). In its epidemiologic usage, risk is a conditional probability, because it is the probability of experiencing an event or becoming a case conditional on remaining "at risk" (eligible to become a case) and "in view" (available for the event to be detected).

Any probability can be transformed into a related measure, the "odds". **Odds** are defined as the ratio of the probability of an outcome to the probability of another outcome. When the only outcomes are (case, non-case), then the odds are the ratio of the probability of becoming a case to the probability of not becoming a case. If the risk or probability of becoming a case [Pr(D)] is p, then the odds of becoming a case are p/(1-p). If the risk, or probability, of developing disease X is 0.05 (5%), then the odds of developing disease X are .05/.95 = 0.0526 (the odds always exceed the risk, especially for large risks).

The mathematical properties of odds make them advantageous for various uses. Whereas probabilities are restricted to the 0 – 1 interval, odds can be any nonnegative number. Odds = 1.0 ("fifty-fifty") corresponds to probability = 0.5, the middle of the set of possible values. The logarithm of the odds can therefore be any real number, with log(odds) = 0 corresponding to the middle of the set of possible values. The natural (Naperian) logarithm of the odds (called the "logit", for "logarithmic transformation") is widely used in biostatistics and epidemiology. For the above example, with risk = 5%, odds = 0.0526, the ln(odds), or logit = -2.944; since the ln(odds) is zero when the risk is .5, a risk smaller than 0.5 yields a negative logit. [Rusty on logarithms? See the Appendix on logarithms and exponents.]

Cumulative incidence when there is loss to follow-up

In the example above, all 200 people who were originally free of disease X were observed over all 3 years. What if instead 20 of the people had died of other causes before developing X? Then not all 200 would have been "at risk" for the entire 3 years.

There are four principal alternatives to estimating the 3-year CI:

1. Ignore the deaths:

   3-year CI = 10/200 = .05

2. Ignore the people who died (analyze only the people followed for all 3 years):

   3-year CI = 10/(200-20) = .056

3. Compromise by counting the 20 people who died as if they were 10 people who were at risk for the full 3 years:

   3-year CI = 10/(200-20/2) = .053

4. Use a lifetable, in which (a) CI is computed for each segment of the period (e.g., annually) to estimate the risk during that segment; (b) risks are converted to survival probabilities (1-risk); and (c) risks are multiplied to obtain the 3-year survival probability and therefore the 3-year risk (1 - survival probability).
5. Take the inverse of the Kaplan-Meier estimated survival proportion. This method is the same as the previous one except that the segments are made so short that only a single case occurs in any one segment. Segments with no cases have 100% survival, so the K-M survival estimate is the product of the proportion surviving during each interval when a case occurs.

Each of these methods makes certain assumptions about when the disease occurs during the three-year period, whether it will be detected when it occurs, and whether the people who die of other causes were more or less likely to develop X had they lived.

Incidence density (ID)

$$\text{ID} = \frac{\text{New cases during stated period}}{\text{Number of person-years of observation}} \qquad \text{(person months, etc.)}$$

Note that:

- ID is a relative rate, not a proportion.

- The units of time must be stated, since otherwise the numeric value is ambiguous (e.g., 15 cases/100,000 person-years = 15 cases/1,200,000 person-months).*
- Ideally, incidence density is the instantaneous rate of disease occurrence at each moment in time. In practice, epidemiologists generally compute average ID during one or more periods.

Interpretation:

ID addresses the question "How rapidly is the disease occurring in the population, relative to its size?", or "What is the intensity with which the disease is occurring?". It has been argued that ID has no interpretation at the individual level (see Morgenstern H, Kleinbaum, DG, Kupper LL, 1980). However, it is possible that ID can be thought of as at least indirectly addressing the question, "How soon might this happen to me?").

Sample calculation:

In our original example for CI, we had 10 cases of chronic disease X develop in 200 people initially free of X and observed over 3 years with no loss to follow-up. Here are the values of CI and ID for this example:

3-year CI = 10 cases / 200 people = 10/200 = .05

ID ≈ 10 cases / (200 people × 3 years) = 10 / 600 person-years

≈ 0.167 cases per person-year (py) = 0.167 / py = 167 / 1000py

The reason for the approximation is that, as we shall see, people stop contributing person-time when they develop the disease so the denominator must be reduced accordingly. The more nearly correct calculation is 10 / (200 × 3 years – 10 × 1.5 years) = 10/585 = 0.17/py, assuming that cases occurred uniformly during the 3 years.

Calculating ID

In calculating ID, we use the same cases as for CI except that we may want to allow multiple events per person. If we regard the recurrences as independent of one another, then we can simply count

---

* The importance of stating units can perhaps be appreciated from the following:
"On Sept. 23, 1999, NASA fired rockets intended to nudge its Mars Climate Orbiter into a stable low-altitude orbit. But after the rockets fired, NASA never heard from its expensive spacecraft again, and scientists later concluded that it had either crashed on the Martian surface or had bounded away, escaping the planet completely.
"The reason for the debacle, scientists concluded months later, was that the manufacturer, the Lockheed Martin Corporation, had specified the rocket thrust in pounds, while NASA assumed that the thrust had been specified in metric-system newtons." Browne, Malcom W. Refining the art of measurement. Science Times, New York Times, 3/20/2001, page D6¶.

them as new cases. If not, we can define the disease as the first occurrence. Other considerations can also affect the choice.

There are several methods used to compute population-time.

1) If individuals are being followed over time, so that the period of disease-free observation is known for each person, we simply add up the disease-free time for all persons:

   population-time = Σ (disease-free time for each person)

2) If a fixed cohort is being followed, but not in sufficient detail to know the period of disease-free time for each individual, we can estimate population time as follows:

   population-time = average population size during the period
   × length of the period of observation

   If there are N_0 disease-free people at the beginning of the period, and during the period there are "C" cases, "D" deaths from causes other than the disease of interest, and "W" persons whose disease status is unknown ("withdrawals"), then the number of disease-free persons at the end of the period is $(N_0 - C - D - W)$. The average number of disease-free people, assuming that cases, deaths, and withdrawals occur uniformly during the period, is:

$$\frac{N_0 + (N_0 - C - D - W)}{2} = (N_0 - C/2 - D/2 - W/2)$$

   and the population-time at risk can be estimated as:

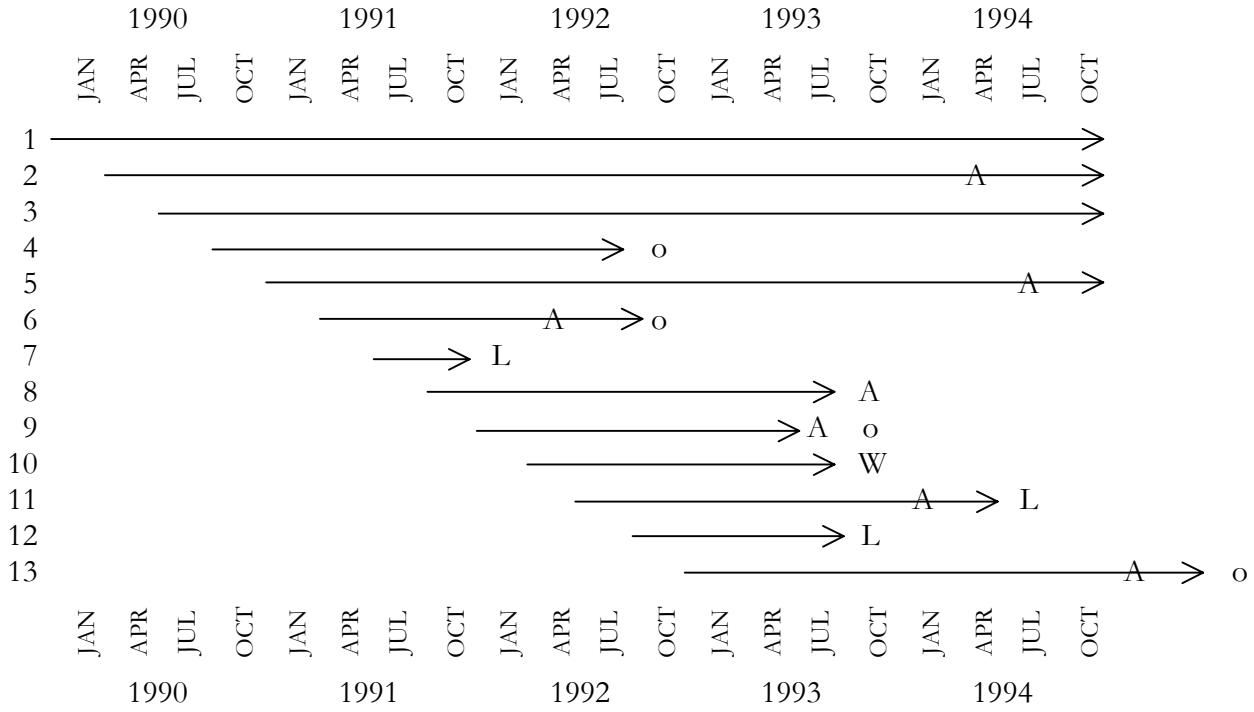
$$(N_0 - C/2 - D/2 - W/2) \times (\text{time interval})$$

3) If we are following a dynamic population (a.k.a. "open cohort") instead of a fixed cohort, we can use the same strategy of multiplying the average size of the disease-free population by the time period. It may be possible to estimate the average number of disease-free people by taking the average of the number of disease-free people at the beginning and end of the period. If we can assume that the population is "stable" (the number of disease-free people who are lost to the population through out-migration, death, and developing the disease of interest is balanced by in-migration), then the number of disease-free people is approximately constant. If we have any usable estimate of the average number of disease-free persons (N_0), then we estimate population time as $N_0 \times$ (time interval)

   If the disease is rare, then the number of disease-free persons (N_0) will be approximately equal to the total number of persons (N), which is more likely to be known. In that case, we can estimate population time as $N \times$ (time interval), where N is the average population size without regard to disease status. Annual death rates and other annual vital statistics rates are typically computed using the estimated mid-year (July 1) population as the denominator, which is approximately the average size of the population on any day in the year if the population is approximately constant or changing in a monotonic fashion.

Calculation of person-time in a cohort
when individual follow-up times are known

Graph of hypothetical follow-up experience for 13 advanced Alzheimer's patients being cared for at home during January 1990 - December 1993 and followed until December 31, 1994 for admittance to a nursing home, in order by study entrance date (after Kleinbaum, Kupper, and Morgenstern, 1982).



Key:

A = admitted to nursing home care

L = lost to follow-up

W = withdrew

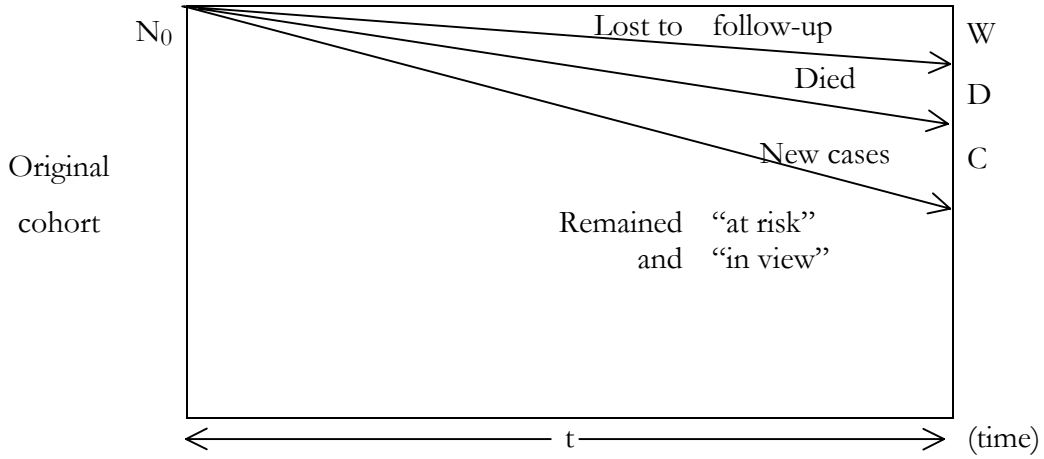
o = died

$$ID = \frac{\text{Cases}}{\text{Sum of disease-free follow-up over all individuals}}$$

Subject	Cases	Follow-up
1		5.0
2	1	4.0
3		4.5
4		2.0
5	1	3.5
6	1	1.0
7		0.5
8	1	2.0
9	1	1.5
10		1.5
11	1	1.5
12		1.0
13		2.0
Total	6	30.0

$$ID = \frac{6}{30 \text{ person-years}} = 0.20 \text{ patients admitted per year}$$

Calculation of person-time in a cohort
when individual follow-up times are not known

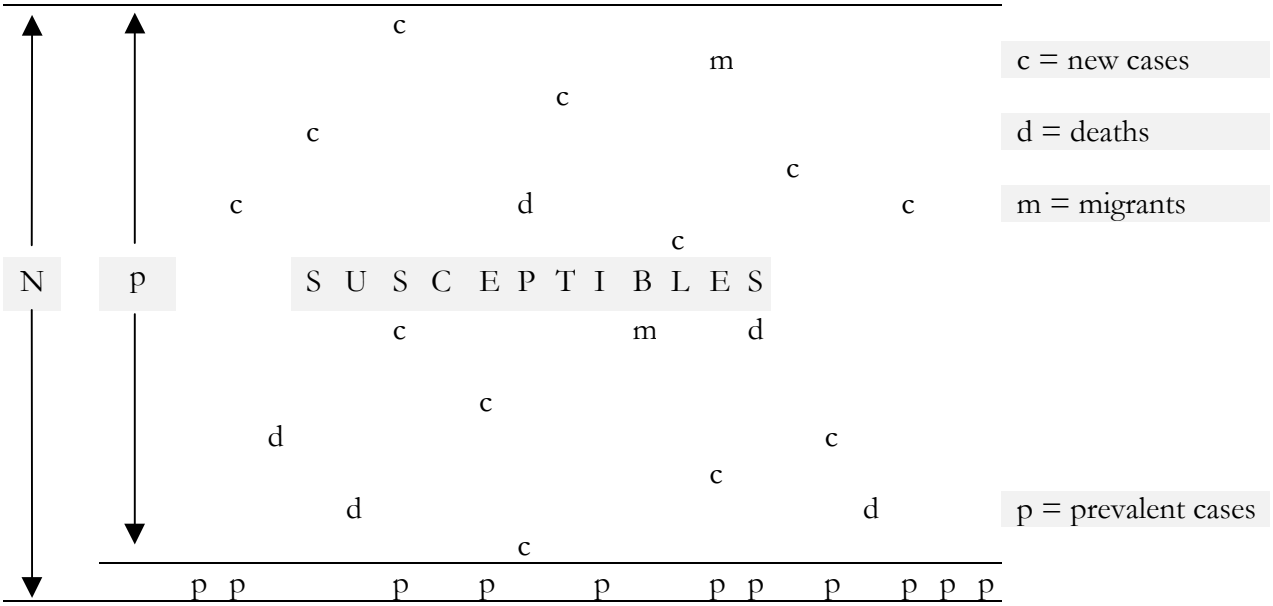


$$ID = \frac{C}{(N_0 - W/2 - D/2 - C/2)\,t}$$

(t = time interval)

(Since the area of a triangle = base × height/2, the person-time lost to follow-up can be estimated by one half times the number of withdrawals [the base of the triangle] times the length of the time interval [the height]. The procedure is the same for follow-up time lost due to deaths and to incident cases. These estimates assume that cases are detected as they occur and that only the first case per subject is counted.)

Calculation of person-time in a stable, dynamic population



Processes at work: Immigration of cases, noncases
Out-migration of cases, noncases
Death of cases, noncases
Development of new cases

$$ID = \frac{\text{cases}}{N_0 t} \quad \text{or} \quad ID = \frac{\text{cases}}{Nt}$$

(t = time interval)

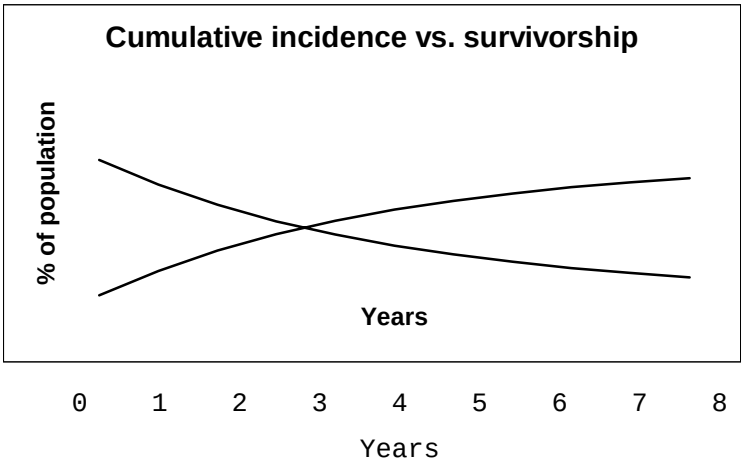
Relationship of CI and ID

Both ID and CI are actually old acquaintances who have changed their outfits. When we calculated life expectancy in the first topic, we used the terms death rate, hazard, cumulative mortality, cumulative survival. ID is essentially the hazard, now applied to events other than death. CI is essentially the cumulative mortality proportion, now applied to events of any variety. Both represent different summary statistics from survivorship analysis (known in engineering as failure-time analysis).

ID is the rate at which the size of the unaffected population is changing, relative to the size of the unaffected population; CI is the proportion of the original population that has been affected by time t. CI is a cumulative measure from a baseline time to a specific later point in time. CI estimates the

average risk for a member of the cohort. In principle, ID can apply to an instant in time, though it can be computed only as an average over some interval. ID is sometimes referred to as the "force of morbidity", in analogy to the hazard function (the "force of mortality").

The following figure shows the relationship between CI and its inverse, the proportion unaffected (survivorship). ID is the relative rate of decline in the survivorship curve.



0 1 2 3 4 5 6 7 8

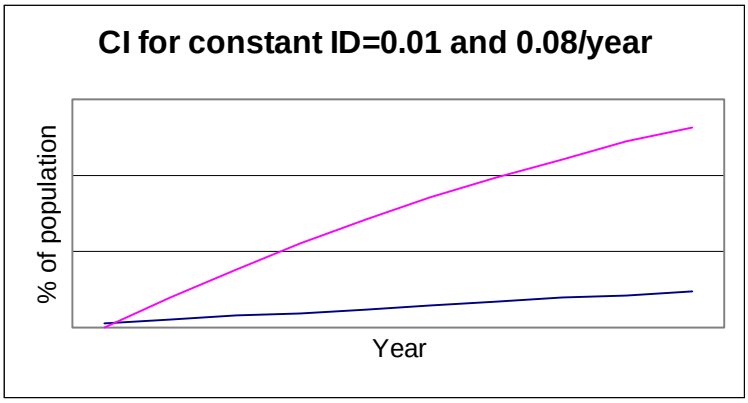
Years

The incidence of AIDS in gay men in San Francisco from 1984 might look something like the left half of this graph.

The mathematical relationship between CI and ID over time can be seen by considering an incurable disease in a hypothetical fixed cohort defined at a point in time and with no entrances or exits other than from the disease in question. Assuming that ID_t (the force of morbidity) is constant over time, cases will develop throughout the follow-up period. However, since the number of unaffected (at risk) cohort members is diminishing, the number of new cases will be smaller in each successive time interval. Because the number of cases is smaller in each interval, the slope of the curve for CI will tend to flatten out as it approaches 1.0 (its maximum value), at which time the entire cohort has developed the disease. The proportion unaffected (the inverse of CI: 1-CI) also becomes less steep. ID_t, of course, we have assumed to be constant. In this situation, the mathematical relationship between CI and ID is:

$$CI = 1 - \exp[-\int(ID_t dt)] = 1 - \exp(-ID\ \Delta t)$$

For a rare disease with a constant ID (or during a sufficiently short time interval): $CI \approx ID \times \Delta t$ (where Δt is the time interval), because since the cohort does not become depleted, the number of new cases in each time interval remains about the same.



Example:

- ID = 0.01/year (1 case per 100 person-years)
- In 5 years, CI will be 0.049, or about the same as ID×5 (=0.05); 95% of the cohort remains disease free and therefore exposed to the 0.01/year ID.
- In 10 years, CI will be .096, only slightly below ID×t (=0.10); 90% of the cohort remains disease free.
- ID = 0.05/year (5 cases per 100 person-years)
- In 5 years, CI will be 0.226, slightly smaller than ID×5 (=0.25); 77% of the cohort remains disease free.
- In 10 years, CI will be 0.40, while ID×t (=0.50); only 60% of the cohort remains disease free.

CI vs. ID - a real-life example

(courtesy of Savitz DA, Greenland S, Stolley PD, Kelsey JL. Scientific standards of criticism: a reaction to "Scientific standards in epidemiologic studies of the menace of daily life", by A.R. Feinstein. *Epidemiology* 1990;1:78-83; it was actually Charles Poole who spotted this *faux pas* [Poole C, Lanes SF, Davis F, *et al*. "Occurrence rates" for disease (letter). *Am J Public Health* 1990; 80:662]; the specific issue being discussed is the effect of alcohol on breast cancer risk)

> ". . . substantially different occurrence rates of breast cancer: about **6.7 per thousand** (601/89,538) in the nurses cohort and about **18.2 per thousand** (131/7,188) in the NHANES cohort." (Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 1988;242:1259 quoted in Savitz DA et al., p.79, emphasis added)
>
> Implication:
>
> (1) Different rates suggest errors in ascertainment of breast cancer
>
> (2) With under/overascertainment, there may be biased ascertainment
>
> (3) The bias may produce more complete or overdiagnosis among drinkers

However:

Nurses: 601 cases/89,538 women over 4 years

CI = 6.7 per thousand (4 years)

ID = 1.68 per 1,000 women-years

NHANES: 121 cases/7,188 women over 10 years (10 cases should have been excluded by Feinstein)

CI = 16.8 per thousand (10 years)

ID = 1.68 per 1,000 women-years

This example illustrates the importance of stating the follow-up period for a CI and the problem that can arise in comparing CI's for different amounts of follow-up.

Two complementary measures of incidence: CI and ID

Cumulative incidence (CI)

1. increases with period of observation (i.e., it is "cumulative")
2. has problems with:
   - multiple events in one subject
   - differing follow-up times for subjects

But

3. it is not necessary to know exact time of onset of the disease
4. directly estimates risk

Incidence density (ID)

1. suggests ability to extrapolate over time - "duration free";
2. accommodates:
   - multiple events in one subject
   - different follow-up times for subjects
3. does not require a cohort to estimate or interpret
4. may be more appropriate for etiologic inference

Choosing between CI and ID

A. Objective

Estimate rate or risk

B. Natural history

Does the period of interest fit within the period of observation? (restricted versus extended risk period)?

E.g., If one wanted to analyze the relative longevity of men and women, the lifetime risk (CI) of death would be useless.

C. Availability of data, e.g.

Fixed cohort, dynamic cohort, dynamic population

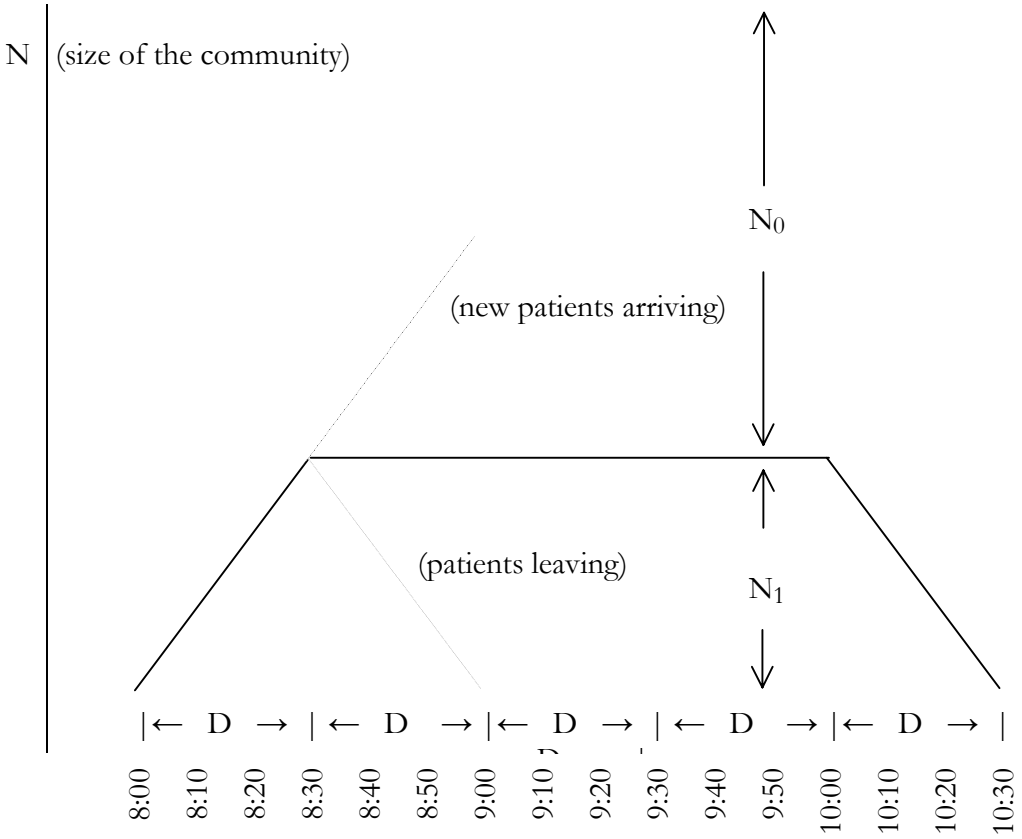
Different follow-up times

Knowing when events occur may favor one method or the other.

Incidence and prevalence in a population

The relationship between incidence and prevalence is the population-level analog for many familiar situations, such as the number of people on line at the grocery store check-out, the number of patients in a waiting room or a hospital, or the number of simultaneous log-ins for an internet service provider.

Incidence, prevalence, and duration: patient flow in a community-based clinic



If a clinic opens at 8:00am, a patient arrives every 10 minutes (6/hour), and it takes 30 minutes for a patient to be seen and treated, then the number of patients in the clinic will rise for the first 30 minutes and then remain constant at 3 patients until the clinic closes and the last 3 patients are

treated. If the rate at which patients arrive were to increase to 10/hour, then in the half-hour it takes to treat the first patient 5 more will arrive, so the number of patients in the clinic will stabilize at 5, instead of 3. Similarly, lengthening the treatment time from 30 to 60 minutes would cause the number in the clinic to increase for the first hour, for a total of 6 patients in the clinic at any time until closing.

With the original assumptions, 6 patients arrive at the clinic every hour during 8:00am-10:00am, and 6 patients leave the clinic each hour during 8:30am-10:30am. During 8:30am-10:00am the clinic is in equilibrium, with 3 patients there at any given time. This equilibrium number, N_1, equals the arrival rate (6/hour) times the average time a patient remains (0.5 hours):

$$N_1 = \text{arrival rate} \times D$$

where D is average duration of a clinic visit.

If the clinic is the only one in a community of size N (or is the approved source of care for N people), then we can express the arrival rate as a function of the size of the community:

$$\text{Arrival rate (patients/hour)} = I \times N_0$$

where I is the incidence of visiting the clinic and N_0 is the number of people available to go to the clinic (N minus the N_1 people already in the clinic, which assumes that people can return to the clinic as soon as they leave or that they immediately leave the community and are replaced by other people eligible to go to the clinic). We can also express the number of patients in the clinic, N_1, as a function of the size of the community, using P as the population "prevalence" of clinic attendance.

$$N_1 = P \times N$$

Making use of these three equations, we can write:

$$\begin{aligned} N_1 &= \text{arrival rate} \times D \\ &= (I \times N_0) \times D \\ P \times N &= (I \times N_0) \times D \end{aligned}$$

$$P = \frac{N_0}{N} \; I \times D$$

Prevalence odds = incidence × average duration

If the number of visitors to the clinic is small in relation to the size of the community, then $N_0/N \approx 1$, and we have the approximation **prevalence = incidence × average duration**.

Otherwise the relationship can be written as **prevalence odds = incidence × average duration**, since:

$$P = \frac{N_0}{N} \; I \times D = \frac{N - N_1}{N} \; I \times D$$

$$P = (1 - P) \times I \times D \quad \text{and} \quad \frac{P}{(1 - P)} = I \times D$$

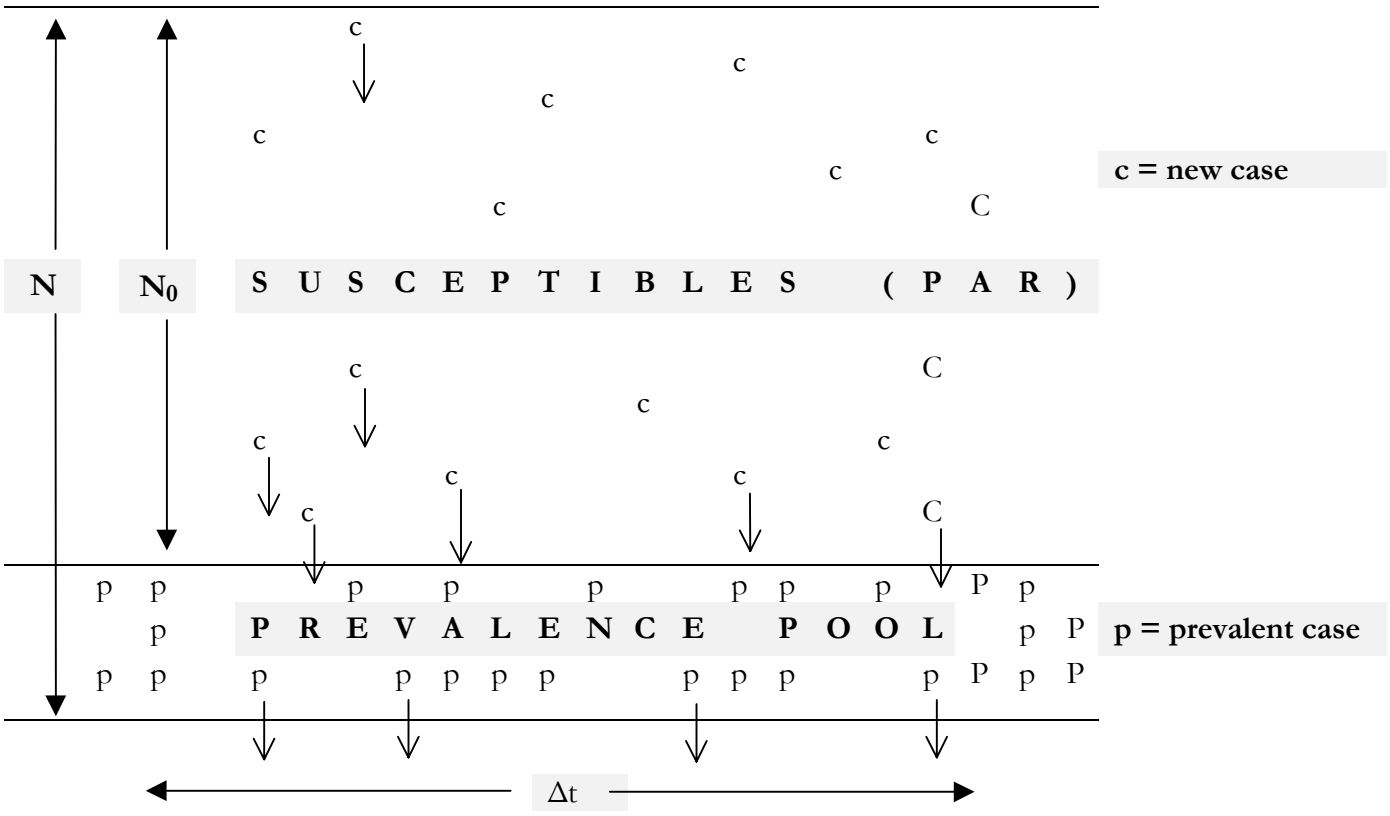
Odds are defined as the ratio of two probabilities, most often the ratio of a probability divided by its inverse (probability for/probability against). The prevalence of a condition is an estimate of the probability that a randomly selected member of the population is a case [Pr(case)]. If the prevalence is p, then the **prevalence odds** are **p/(1-p)**. So the prevalence odds, i.e., the odds that a randomly selected person in the population has the disease (i.e., is a prevalent case) are:

$$\begin{aligned} \text{prevalence odds} &= \text{prevalence} \; / \; (1 - \text{prevalence}) \\ &= (N \times \text{prevalence}) \; / \; (N - N \times \text{prevalence}) \\ &= (N \times \text{prevalence}) \; / \; N_0 \;\; = \;\; (N/N_0) \times \text{prevalence} \end{aligned}$$

Incidence, prevalence, and duration in a stationary population

The following diagram displays the above process as it might appear for cases of a disease occurring in a population followed during an interval of time, in equilibrium with respect to disease incidence, duration, and entrances and exits from the population. An alternate derivation of the relation prevalence odds = incidence × duration follows. (See Rothman and Greenland, 1998 for more on this topic.)

Incidence and prevalence in a population of size N observed for a time interval Δt



c's are incident (new) cases

p's are prevalent (existing) cases

Δt indicates the time interval

↓ indicates exits from unaffected population or from prevalence pool

Size of the population = N = disease-free persons + existing cases = N_0 + prevalence pool

The assumption that incidence and prevalence are constant means that:

$$\text{New cases} = \text{Terminations}$$

$$(\text{Incidence} \times N_0) \times \Delta t = (\text{Prevalence} \times N \times \text{Termination rate}) \times \Delta t$$

$$\text{Prevalence} \times \frac{N}{N_0} = \frac{\text{Incidence}}{\text{Termination rate}}$$

Since the termination rate is the rate at which existing cases leave the prevalence pool, this rate is the reciprocal of the average duration of a case. To see this, consider the termination rate for a single case:

$$\text{Termination rate} = \frac{\text{Terminations}}{\text{No. of cases} \times \Delta t}$$

For a single case,

$$\text{Termination rate} = \frac{1}{1 \times \Delta t} = \frac{1}{\Delta t}$$

$$\text{Average duration (i.e., } \Delta t) = 1 \text{ / Termination rate}$$

Thus, in the above relationship between incidence and prevalence, we can substitute Duration (D) for 1 / Termination rate:

$$\text{Prevalence} \times \frac{N}{N_0} = \text{Incidence} \times \text{Duration}$$

So in a population that is in a steady state with respect to a given condition, the prevalence odds of that condition equals the incidence times the average duration (the prevalence does too, if it is sufficiently small). Conversely, if we observe that the prevalence odds of a condition remains constant (and can assume a stable population with no net migration of cases), then the incidence must balance the loss of cases due to death or cure. Since prevalence is often easier to ascertain than is incidence, we can make use of this relationship to draw inferences about incidence.

Estimating incidence from prevalence data

This relation has been used as the basis for estimating HIV seroincidence from seroprevalence data, using a seroassay procedure designed to identify recently-infected persons (Janssen et al., 1998). This technique makes use of the fact that ELISA tests for HIV antibody have become considerably more sensitive since they were first developed. People who test HIV-positive with a current (highly sensitive) HIV antibody test are then re-tested with a "detuned" version of an older, less-sensitive test. Since it takes time for the anti-HIV antibody titer to increase to the level that it can be detected with the less sensitive test, there is a period of time (about four months) during which the less sensitive test will be negative. The discordant results of the two HIV antibody tests defines a short-lived (average duration 129 days) "condition" whose prevalence can be used to estimate occurrence of new HIV infections. Solving the relation Prev odds = I × D yields I = Prev odds / D ≈ P/D for

small prevalence. So if the seroprevalence of recent infection in a stable population is 2%, the incidence of new HIV infections is approximately 0.02/129 days = 0.057/year. = 5.7/100py.

However, the stable population assumption is often not met in practice, and the above model is also grossly simplified in that it treats the entire population as a homogenous entity, ignoring the influence of age (see Rothman and Greenland, 1998). When we examine the relationship between incidence and prevalence within a specific age-band, we need to consider the effect of entrances and exits due to aging into or from the age band of interest. For example, the U.S. armed forces have conducted serologic testing for HIV antibody of all recruits, active duty military, and reserves since the antibody test became available. Within each age group, the seroprevalence of HIV antibody has been approximately constant over a period of years. If we could ignore the effect of age, then using the relationship prevalence odds = incidence × average duration, we could conclude that HIV incidence should equal the (small) proportion of the population who leave the prevalence pool each year due to discharge or death. However, another manner of exiting from the prevalence pool of a given age group is to age out of it into the next one. Since HIV seroprevalence increases with age (up to about age 35 years), it can be inferred that infections (incident cases) are occurring more rapidly than necessary to balance deaths and discharges among cases. The reason is that each year, some of the persons in each age group are replaced by persons from the next younger age group, a group with lower seroprevalence. If infections were not occurring at a rate sufficient to balance this outflow of prevalent cases, then the prevalence in each age group would decrease over time, as the lower prevalence groups move up in age (see David Sokol and John Brundage, Surveillance methods for the AIDS epidemic, *NYS J Medicine* May 1988).

Furthermore a meaningful incidence measure still requires identification of a cohort or source population. Although the detuned serologic assay for recent HIV infection has been used to estimate HIV "incidence" among clinic patients, the interpretation of those estimates is highly problematic (Schoenbach, Poole, and Miller, 2001).

Bibliography

Bailar, John C., III; Elaine M. Smith. Progress against cancer? *N Engl J Med* 1986; 314:1226-32. (included in an assignment)

Elandt-Johnson, Regina C. Definition of rates: some remarks on their use and misuse. *Am J Epidemiol* 1975;102:267-271.

Gable, Carol Brignoli. A compendium of public health data sources. *Am J Epidemiol* 1990; 131Z:381-394.

Gaffey, WR. A critique of the standardized mortality ratio. *J Occupational Medicine* 1976;18:157-160.

Glantz, Stanton H. *Primer of biostatistics*. NY, McGraw-Hill, 1981.

Hook, Ernest B. Incidence and prevalence as measures of the frequency of birth defects. *Am J Epidemiol* 1983;116:743-7

Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. JAMA 1998; 280:42-48.

Morgenstern H, Kleinbaum, DG, Kupper LL. Measures of disease incidence used in epidemiologic research. *Int J Epidemiology* 9:97-104, 1980.

Remington, Richard D. and M. Anthony Schork. *Statistics with applications to the biological and health sciences*. Englewood Cliffs, NJ, Prentice-Hall, 1970.

Rothman, Kenneth J. Clustering of disease. Editorial. *Am J Public Health* 1987; 77:13-15.

Victor J. Schoenbach, Charles Poole, William C. Miller. Should we estimate incidence in undefined populations? *Am J Epidemiol* 2001;153(10).935:937.

Smouse, Evan Paul; Martin Alva Hamilton. Estimating proportionate changes in rates. *Am J Epidemiol* 1983; 117:235-43.

Zeighami EA, Morris MD. The measurement and interpretation of proportionate mortality. *Am J Epidemiol* 1983; 117:90-7.

Appendix on weighted averages

Because epidemiology studies populations, and populations contain various subgroups, weighted averages figure prominently in epidemiology. Nearly any population-based measure can be regarded as a weighted average of the value of that measure across the subgroups that comprise the population. Weighted averages are used to standardize or adjust crude measures to make them more comparable across populations with different subgroup proportions. Both the concept and the mathematics are fundamental.

A weighted average is like an ordinary mean except that the components being averaged can have more or less influence (weight) on the resulting average. For example, suppose we measure systolic blood pressure on 10 occasions and obtain the following values (mmHg): 95, 100, 100, 105, 105, 105, 110, 110, 115, 120. If we want the mean (average) systolic blood pressure, we simply sum the individual measurements and divide by the number of readings: 1,065/10 = 106.5 mmHg. Since some of the readings occur more than once, we could achieve the same result by using a weighted average:

Number of readings	Value	Weighted sum
1	95	95
2	100	200
3	105	315
2	110	220
1	115	115
1	120	120
10		1,065

Average = 1,065 / 10 = 106.5 mmHg.

A small business might use a layout like this to compute the average price paid for some commodity over some time period. In that situation, the first column might show the number of sacks purchased, the second column the price per sack, and the third column the total dollar amount.

With a little generalization (to permit the "number of readings" to be a fractional number), we have the procedure for creating a weighted average. Familiar examples are grade-point averages (course grades weighted by credit hours), average cost per share of a stock purchased in multiple buys, and average price per gallon for gasoline purchased on vacation.

Mathematically, a weighted average is a linear combination where the coefficients (p_i) are proportions whose sum is 1.0. Several equivalent formulations are:

$$\frac{w_1a_1 + w_2a_2 + \ldots + w_na_n}{w_1 + w_2 + \ldots + w_n} = \frac{w_1a_1 + w_2a_2 + \ldots + w_na_n}{W}$$

$$= \frac{w_1a_1}{W} + \frac{w_2a_2}{W} + \ldots + \frac{w_na_n}{W} = \sum\left(\frac{w_ia_i}{w_i}\right)$$

$$= p_1a_1 + p_2a_2 + \ldots + p_na_n = \sum(p_ia_i)$$

where $W = w_1 + w_2 + \ldots + w_n$ and $p_1 + p_2 + \ldots + p_n = 1$

For the gasoline price example, the w_i represent the amount purchased at each stop and the a_i represent the price of each purchase.

Appendix on exponents and logarithms

(Adapted from Defares JG and Sneddon IN. *An introduction to the mathematics of medicine and biology*. The Netherlands, North-Holland, 1960)

Some simple facts:

$2^2 = 2 \times 2 = 4$

$2^3 = 2 \times 2 \times 2 = 8$

Square root of 4 = 2

Cube root of 8 = 2

Exponents:

b^x means b raised to the x^{th} power; x is referred to as an exponent.

If x is 2, then $b^x = b^2 = b \times b$. If x is 3, then $b^x = b^3 = b \times b \times b$. From this we can reason that:

1) $b^m \times b^n$ must be equal to $b^{(m+n)}$

(The product of a number raised to the m-th power multiplied by the same number raised to the n-th power equals that number raised to the sum of the powers.)

2) b^m / b^n must be equal to $b^{(m-n)}$

(The quotient of a number raised to the m-th power divided by the same number raised to the n-th power equals that number raised to the difference of the powers (numerator power minus denominator power.)

3) $(b^m)^n$ must be equal to $b^{(m \times n)}$

(The m-th power of a number raised to the n-th power equals that number raised to the (m × n)-th power.)

For exponents that are not positive integers, we define b^x in order to preserve the above three rules. So $b^0=1$ and $b^{-x} = 1 / b^x$.

When the base number (b in the above examples) is e, a transcendental number that is approximately 2.7183, then we write e^x or (for typographical convenience) exp(x). e and Naperian logarithms have special properties that recommend them for use in mathematics and statistics.

Logarithms:

If for a number b (greater than 1.0), it is possible to find a number x such that:

$$y = b^x$$

then we say that x is the logarithm of y to the base b:

$$x = \log_b y$$

Taking the logarithm is the inverse of exponentiation, so that if $y=b^x$:

$$\log_b(y) = \log_b(b^x) = x \qquad \text{and}$$

$$b^x = b^{(\log(y))} = y$$

To preserve consistency with the rules for exponents, above, we see that:

1) $\log_b(xy) = \log_b x + \log_b y$

(the logarithm of a product is the sum of the logs)

2) $\log_b(x/y) = \log_b x - \log_b y$

(the logarithm of a quotient equals the logarithm of the numerator minus the logarithm of the denominator), and

3) $\log_b(x^n) = n \log_b x$

Logarithms are defined so that these rules generalize to the case of fractional and negative exponents. The log of a negative number, however, is undefined.

The base b must be a positive number greater than 1.0, and is usually 10 (for "common logarithms") or e (for "natural" or Naperian logarithms). The latter are most often seen in mathematics, statistics,

and epidemiology. The notation ln(x) or simply log(x) is often used when Naperian logarithms are understood.

Note that for base e (= 2.7183), exp(x) (a) must be greater than zero, (b) will equal 1 when x=0, and (c) will increase very rapidly for large x. In contrast, ln(x) (a) will be negative for x<1, (b) will equal 0 when x=1, and (c) will be positive for x>1. So if x is a positive ratio whose null value is 1.0, ln(x) will represent a transformation of x with the null value at 0 and other values distributed symmetrically around it. These properties of logarithms are useful for transforming variable distributions and for the analysis of ratios.

Measuring disease and exposure - Assignment

1. The graph below shows the trends in incidence and prevalence for chronic disease Q over a 50-year period. Which of the following interpretations is consistent with the graph below? Circle as many as could logically be correct.

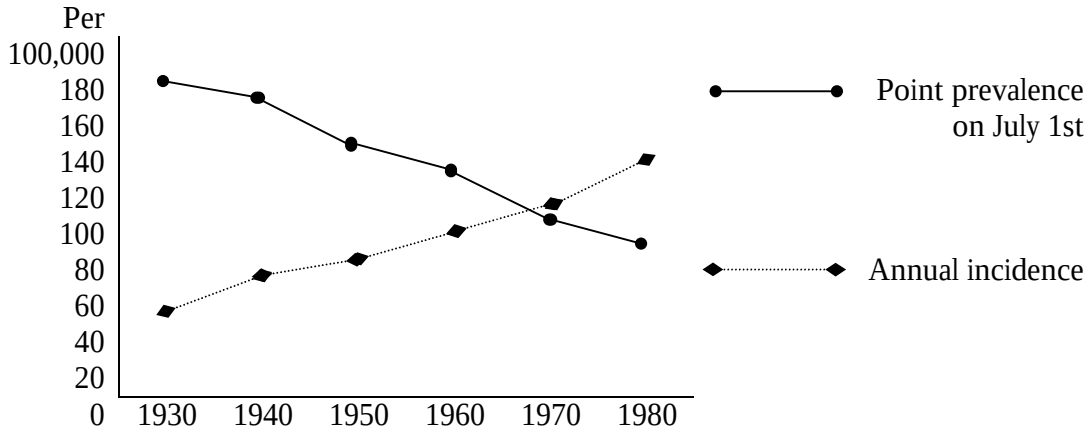
   A. The disease may be becoming more chronic with lower case-fatality rate;

   B. The disease may be becoming more rapidly fatal (i.e., it kills patients sooner than before);

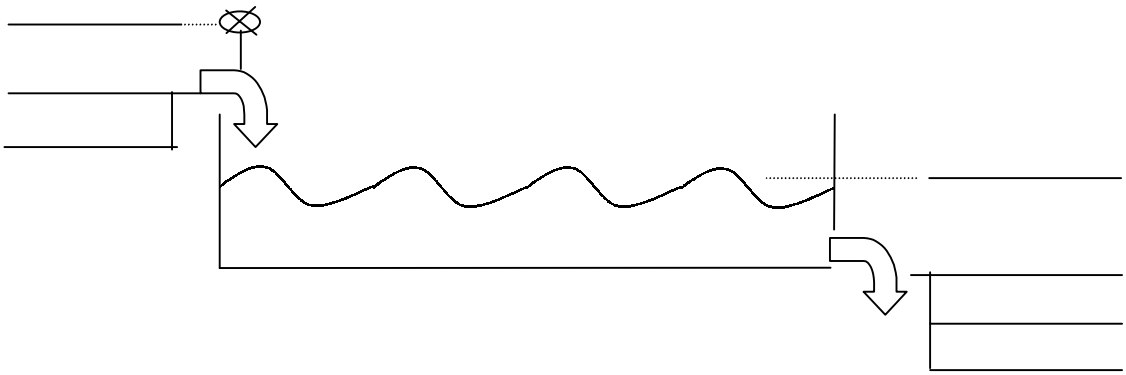
   C. The disease may be becoming shorter in duration due to better medical treatment;

   D. The disease may be becoming more rare due to better preventive public health programs.

Incidence and prevalence of disease Q



2. Fill in the blanks in the following diagram, using the terms "incidence", "prevalence", "case fatality", "recovery", "prevention", "inmigration", "outmigration".



3. For the following hypothetical data on viral upper respiratory infections(URI), calculate the epidemiologic measures listed below. Assume that:

   - each infection lasts 10 days and confers no immunity
   - infections begin at 12:01 A.M. of the date shown
   - there are no deaths from URI or other causes and no loss to follow-up
   - "thirty days has September, April, June, and November. All the rest have thirty-one."
   - a person is not at risk of a new URI until he/she has recovered from an existing episode

Person	Dates of onset of URI episodes
A	(none)
B	August 24, October 5
C	September 12
D	(none)
E	(none)
F	November 26
G	September 2, November 29
H	(none)

   First draw a time-line chart of the illness episodes for all subjects. Then calculate:

   a. Point prevalence of URI on September 1: _____

   b. Point prevalence of URI on November 30: _____

   c. Person-days at risk (total) for the period September 1 through November 30, inclusive:

      _____

   d. Average ID of URI for the period of September 1 through November 30, inclusive.

      _____

   Be sure to show units where applicable.

4. Regina Elandt-Johnson gives the following definitions of epidemiologic "rates":

   Ratio: the result of dividing one quantity by another. More specifically, the numerator and denominator are two separate and distinct quantities, and may be measured in the same or different units. Examples:

   Sex ratio = (No. of males) / (No. of females)

Fetal death ratio = (no. of fetal deaths) / (No. of live births)

Proportion: a ratio in which the numerator is included in the denominator, i.e., [p = a/(a + b)]. Example:

Proportion of males = (# males)/[(# males) + (# females)]

Proportions must have values between 0 and 1 (inclusive) and can be used to estimate probabilities, or risks.

Rate: a measure of change in one quantity per unit of another quantity on which the first depends. Three kinds are discussed:

$$\text{absolute instantaneous rate of change in y per unit time} = \frac{dy}{dx}$$

$\frac{dy}{dx}$ represents the derivative of y with respect to x

In calculus, the derivative is shown to the slope of the function relating Δy to Δx ["Δ" means "change"]. The derivative is defined as the limit of the change in y divided by the change in x as the change in x becomes infinitessimally small. Calculus is not required for this course.)

$$\textit{Absolute}\text{ average rate of change in y per unit time} = \frac{\Delta y}{\Delta t}$$

$$\textit{Relative}\text{ average rate of change in y per unit time} = \frac{\Delta y}{y(\Delta t)}$$

[Regina Elandt-Johnson. Definition of rates: some remarks on their use and misuse. *Am J Epidemiol* 1975; 102(4):267-271.]

For each of the following ratios, indicate whether it is a rate (R) or a proportion (P) or neither (N). If a rate, indicate whether it is absolute or relative.

a. _____ 3 cases / 25 person-years

b. _____ 3 cases / 25 persons

c. _____ 6 fatalities / 24 acute MI admissions

d. _____ 200 abortions / 1000 live births

e. _____1,000 new cases of diarrhea / day in 500,000 people

5. In 1960, investigator A took a simple random sample of 1,050 adults from an urban community of 100,000 (i.e., each adult had an equal, 1,050/100,000 chance of being chosen for the sample). After examining the entire study population of 1,050, she had detected 50 cases of disease Q, a chronic disease for which there is no recovery or cure.

   In 1965 (5 years later), investigator A re-examined all of the survivors from her original study population and determined the cause of death in those who had died since the first examination. Of the 50 subjects in whom disease Q was detected in 1960, 40 had died prior to being re-examined in 1965. Of those who did not have disease Q in 1960, 100 subjects developed it between 1960 and 1965 including 50 subjects who died prior to reexamination (presumably due to disease Q). Among the subjects who did not contract disease Q, 15% had died between the 1960 and 1965 examinations.

   a. Draw a flow diagram for the study.

   b. Calculate estimates of the following measures:

      i. Point prevalence of disease Q among adults in the community at the initial examination

      ii. 5-year cumulative incidence of disease Q (make no adjustment for deaths from causes other than disease Q). What is the impact on this measure of the deaths among individuals who did not develop disease Q?

      iii. Average incidence density for disease Q in the cohort followed. (Be sure to state the units.)

      iv. The 5-year case fatality rate for disease Q (as a proportion of those diagnosed as having disease at the initial examination--see (i) above).

      v. The prevalence of disease Q among subjects alive at the time of the re-examination (i.e., 1965).

   c. Most of the measures computed above are proportions. What are the theoretical lower and upper limits of this class of measures?

   d. Which of the above measures is (are) not a proportion?

e. The case fatality rate was originally devised to assess the virulence (severity) of an infectious disease. If another investigator reported a value for the case fatality rate for disease Q, what assumption about the duration of the disease among cases at the beginning of the study is involved in comparing the two case fatality rates?

f. Which of the above measures would you use to estimate the average risk of developing disease Q? State that risk estimate in terms of the language of probability.

6. Give a one-sentence definition, in terms that you might employ in an article for the educated but non-professional public, of:

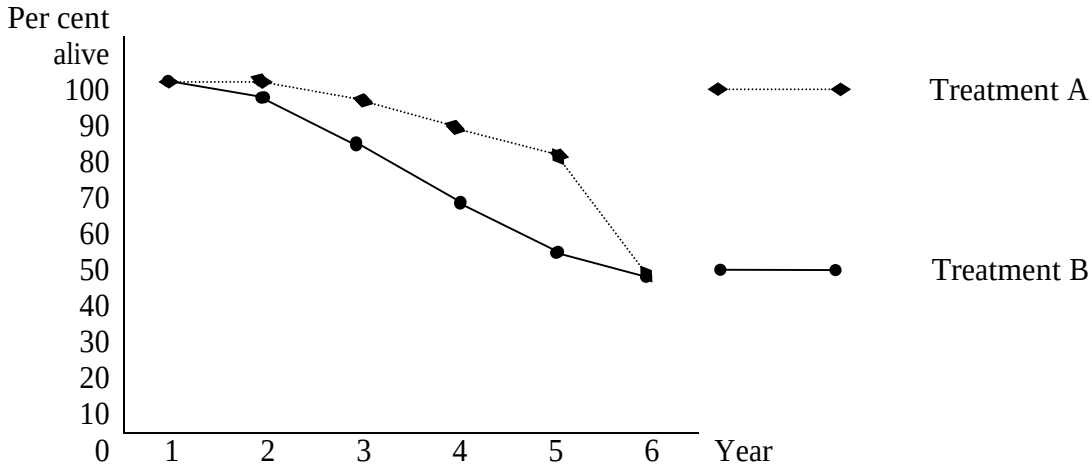
a. Cumulative incidence

b. Incidence density

c. Prevalence

7. What are the three basic constituents or components of the concept of incidence?

8. The following graph shows the results of a controlled clinical trial of two treatments of a highly fatal cancer:



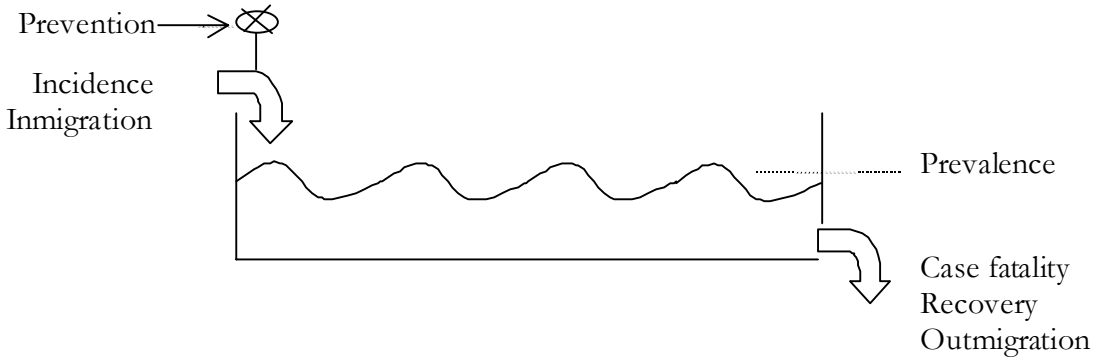
a. Assuming that the apparent differences at years 4 and 5 are statistically significant, which treatment was superior in prolonging life?

b. Why would survivorship analysis methods be preferable to the use of 5-year survival ratios or similar measures for the analysis and interpretation of the results of this trial?

Measuring disease and exposure - Assignment solutions

1. "b" & "c" are correct; shorter duration can lower prevalence despite rising incidence. "a" is incorrect, as the prevalence would increase, not decrease, with increasing chronicity. "d" is incorrect, as prevention should reduce the incidence.

2.



3.

a. 0.125 (1 case with 8 persons at risk)

$$\text{Prevalence} = \frac{\text{Cases present in a population at a specified time}}{\text{Number of persons in that population at that time}}$$

b. 0.250 (2 cases with 8 persons at risk)

c. person days at risk = 689:

Total person days = 91 days (3 mos.) x 8 persons = 728.

There are 39 days within this 3-month period when individuals are not at risk because they are already ill (B loses 12 days *within the period of observation* 9/1 - 11/30 inclusive, C loses 10 days, F loses 5 days, and G loses 12 days): 728 - 39 = 689 person-days

d. Incidence density:

$$\text{Average incidence density} = \frac{\text{Number of new cases}}{\text{Population time at risk}} = \frac{5}{689}$$

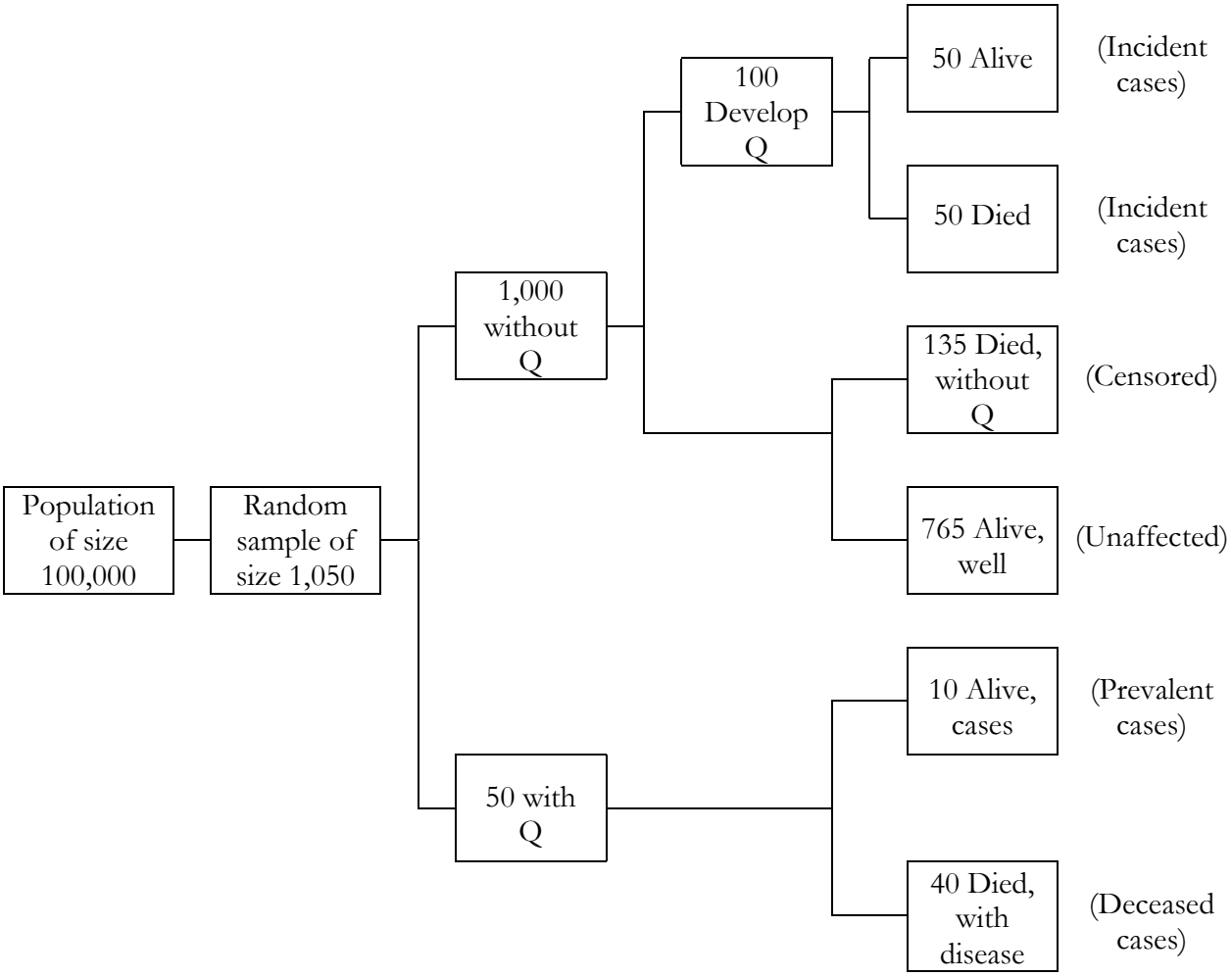
$$= 0.0073 \text{ cases per person-day}$$

Specification of units for incidence density is essential, since the number has no meaning in itself (for example, the incidence density could be expressed per person-week, per person-month, etc., with a different numerical value for the incidence density in each case). In contrast, proportions have no units, though a scaling factor is often used in order to write the number in a more readable fashion, e.g., 153 per 100,000 is a more easily read number than 0.00053, but either form is correct and complete for prevalence or incidence proportion.

4.

a. Rate (relative)
b. Proportion
c. Proportion
d. Neither - this is (only) a ratio
e. Rate (relative) - change in cases / change in time relative to population

5. a. Flow Diagram



f. (i) point prevalence at the initial examination:

50/1050 = .048, or 48 cases per thousand

(ii) 5-year cumulative incidence:

$$\text{Cumulative incidence} = \frac{\text{Number of new cases}}{\text{Population at risk}}$$

There were 100 new cases and 1000 disease-free persons at the start of the period. Therefore:

$$CI = \frac{100}{1.000} = 0.10, \text{ or } 100 \text{ per } 1{,}000$$

However, 135 persons died of other causes than X and therefore were not actually "at risk" of developing disease Q, at least not throughout the 5 years. Omitting them gives:

$$CI = \frac{100}{865} = 0.116, \text{ or } 116 \text{ per } 1{,}000$$

The former CI (0.10) probably underestimates the "true" CI, since it implicitly assumes that none of the 135 persons who died of other causes would have developed disease Q had he lived. The latter CI may overestimate the "true" CI since, after all, the 135 who died were available to get disease Q and be detected during part of the follow-up period.

A compromise solution is to estimate the CI by taking into account the follow-up time on those subjects who died of other causes (or who withdrew from the study for other reasons). One method is:

$$CI = \frac{Q}{(N - W/2)} = \frac{100}{(1{,}000 - 135/2)} = 0.107$$

Where: Q = new cases of disease Q

N = initial cohort (disease free)

W = withdrawals

This method assumes that:

- subjects withdrew (died) evenly throughout the period (i.e., that they withdrew, on the average, at the midpoint).
- subjects were in fact at risk of disease (and detection of disease) prior to withdrawal - e.g., if they had developed disease Q, it would have been noted at the time of their death.

If the loss to follow-up is small, the results of each method will be about the same. An intensive search for a random sample of those originally lost to follow-up can be invaluable in assessing bias.

(iii) Average incidence density

$$ID = \frac{\text{New cases}}{\text{Population time at risk}} = \frac{Q}{\frac{1}{2}(N_1 + N_0)(\Delta t)}$$

Where: Q = new cases

N_1 = size of initial cohort

N_0 = number alive and well at follow-up

Δt = length of follow-up

So that:

$$ID = \frac{100}{\frac{1}{2}(1{,}000 + 765)(5)} = 0.023/\text{year} = 23 \text{ cases per } 1{,}000 \text{ py}$$

The same result can be obtained from:

$$ID = \frac{Q}{\frac{1}{2}(N_1 + N_0)(\Delta t)} = \frac{100}{(1{,}000 - \frac{1}{2}[100] - \frac{1}{2}[135])(5)}$$

(iv) 5 yr case fatality rate:

$$\text{5-year CFR} = \frac{\text{Deaths from Q}}{\text{Cases of Q at initial exam}} = \frac{40}{50} = 0.80, \text{ or } 80\%$$

(v) Prevalence of disease at the reexamination (1965):

$$\text{Prevalence} = \frac{60}{825} = 0.073 = 73 \text{ cases per } 1{,}000$$

a. The lower and upper limits of proportions are 0 and 1, respectively.

b. Incidence density is an average rate, not a proportion.

c. The assumption is that the distribution of duration of the disease is similar between the two case groups. Information on age, sex, and other potentially relevant characteristics would also be desirable.

d. Cumulative incidence would be used to estimate risk. In probability terms, Pr(D|at risk for 5 years)=0.107, or an individual in the study population had a 10.7% chance of developing disease Q in the next 5 years if he does not first die of another cause during that period.

6. Definitions:

   a. Cumulative Incidence - the proportion of new cases that develop in a population at risk of getting the disease, over a stated period of time.
   b. Incidence Density - the rate at which new cases develop per unit time, relative to the size of a population at risk of getting the disease.
   c. Prevalence - the number of existing cases of a disease as a proportion of a defined population at a specified point in time (or short period of time).

7. The three basic components of incidence are:

   a. the number of new cases
   b. the population at risk
   c. the period of observation or follow-up.

8.

   a. Treatment A was superior in prolonging life. Even though the proportion of patients dying by year 6 was the same for each treatment, patients receiving treatment A tended to survive longer (die later during the follow-up period).
   b. The value of a survival ratio would depend upon the (arbitrary) choice of time period. For example, in the graph shown, the 3-year survival advantage for treatment A is very small, the 5-year advantage is quite large. Survivorship analysis considers the time-to-death for patients in the two groups, providing a fuller basis for comparison. After all, by the end of a long enough follow-up period, all subjects will be dead! The aim of medical treatment (and health promotion) is, among other things, that we should die later.

6. Standardization of rates and ratios*

Concepts and basic methods for deriving measures that are comparable across populations that differ in age and other demographic variables.

Overview

Epidemiologists are always mindful of population diversity. Virtually every large population is heterogeneous in regard to sociodemographic (e.g., age, gender, education, religion), geographic, genetic, occupational, dietary, medical history, and innumerable other personal attributes and environmental factors related to health. A population can be viewed as a composite of diverse subgroups (ultimately, subgroups of size one, i.e., individuals, but epidemiologic measures break down at that point). Any overall measure or statistic reflects the value of that measure for each of the subgroups comprising the population.

An overall measure that does not take explicit account of the composition of the population is called **crude**. Its value will be an average of the values for the individual subgroups, weighted by their relative sizes. The larger the subgroup, the more influence it will have on the crude measure (i.e., "democracy"). Thus, the death rate for a population is a weighted average of the death rates for its component subgroups. Suppose we consider a population of size N as consisting of five age groups, or **strata**. Each age **stratum** will have a specific number of people, say n_i (i=1 to 5). During the following year, each stratum will experience some number of deaths, say d_i. The total population size, N, is therefore Σn_i, the total number of deaths, D, is Σd_i, and the crude mortality rate is D/N, which can also be written as a weighted average of the **stratum-specific mortality rates**, d_i/n_i, as follows:

$$\frac{D}{N} = \frac{\Sigma d_i}{N} = \frac{\Sigma\, n_i\,(d_i/n_i)}{N} = \Sigma(n_i/N)(d_i/n_i) = \Sigma w_i(d_i/n_i)$$

where w_i are the weights (note that $\Sigma w_i = \Sigma(n_i/N) = (\Sigma n_i)/N = \Sigma n_i/\Sigma n_i = 1$).

The crude rate is the simplest and most straightforward summary of the population experience. But mortality is strongly related to age, so the stratum-specific mortality rates will differ greatly from one another. The summary provided by the crude rate glosses over this heterogeneity of stratum-specific mortality rates.

---

* (An earlier version of the chapter was prepared by Timothy Wilcosky, Ph.D.)

This issue is particularly relevant when we compare rates across populations or time periods, because if the populations differ in composition, then at least some of what we observe may be attributable to these differences. For example, suppose you and a friend each agree to bring 10 pieces of fruit to a picnic. You stop at a fruit stand and buy 8 mangoes ($1.00 apiece) and 2 apples ($0.50 apiece). Meanwhile your friend goes to the supermarket and buys 2 mangoes ($1.75 apiece) and 8 apples ($0.45 apiece). Which is the more expensive purchase? From one perspective, the first purchase is the more expensive, since $9.00 is certainly greater than $7.10. But from another perspective, the second purchase is more expensive, since the supermarket charged a much higher price for the mangoes and only slightly less for the apples.

Which of these perspectives you choose depends on the purpose of your question. More often than not, the epidemiologist (and the serious shopper) would ask whether the prices were higher in the fruit stand or the store and by how much. We can answer that question by simply comparing the price lists. But what if you also bought oranges, melons, grapes, and bananas? What if you bought two dozen varieties of fruit? It would certainly be more convenient to have a summary measure that permitted an overall comparison. The trouble with total cost ($9.00 versus $7.10) or average price ($0.90/piece of fruit versus $0.71 per piece) is that the fruit stand average price gives more weight to the price of mangoes, because you bought more mangoes, whereas the supermarket average price gives more weight to the price of apples because your friend bought more apples. We're comparing apples to mangoes, instead of fruit stand to supermarket.

Clearly what we need is a procedure that averages the prices in the same way for each vendor, so that both averages give the same proportionate weighting to mangoes. The average prices will depend upon the weighting we use, but at least we will be comparing (proportionally speaking) apples with apples and mangoes with mangoes. However, it's also clear that at least in this example, the weights will determine which seller is favored by the comparison. The fruit stand owner will prefer a higher weight for the price of mangoes, so that her prices will seem the better bargain. But the supermarket owner will prefer a very low weight on the mangoes. He might argue, in fact, that mangoes are a specialty item and not really worth considering in the comparison. He might argue for assigning zero weight to the mangoes, so that his average price will be 0.45/piece (the summary is simply the price of the apples), which is less than the fruit stand charges for apples.

Which set of weights is the right one to use? People who don't like mangoes might agree with the supermarket owner. People who like mangoes – or fruit stands – would not. For the most part, the choice of weights (a.k.a. the **standard population**) is based on convention, the intended and potential comparisons, and various other considerations. There is often no absolute correct choice, and there can easily be different opinions about the best one. But it helps to have a rationale for the choice other than that it happens to give you a result you like. Finally, nothing you do about weights is going to change the fact that your purchase **did** cost more than your friend's, so the crude summaries are not irrelevant.

Adjustment and standardization

The terms "adjustment" and "standardization" both refer to procedures for facilitating the comparison of summary measures across groups. Such comparisons are often complicated by

differences between the groups in factors that influence the measures of interest but which are not the focus of attention. Adjustment attempts to remove the effects of such "extraneous" factors that might prevent a "fair" comparison.

"Adjustment", the more general term, encompasses both standardization and other procedures for removing the effects of factors that distort or *confound* a comparison. Standardization refers to methods of adjustment based on weighted averages in which the weights are chosen to provide an "appropriate" basis for the comparison (i.e., a "standard"), generally the number of persons in various strata of one of the populations in the comparison, an aggregate of these populations, or some external relevant population. Other kinds of adjustment, some of which also employ weighted averages, will be discussed in the chapter on **Confounding**.

Most textbooks of epidemiology present the topic of rate standardization in relation to adjusting for age. This tendency is not coincidental, since virtually all mortal or morbid events occur with different frequencies among groups of different ages. But the same principles and procedures apply to subgroups defined by other variables. The following example illustrates how these varying frequencies can affect a summary measure. Table 1 indicates that in 1970, 5,022 out of the 562,887 white women in Miami died, and that 285 of the 106,917 white Alaskan women died. The respective overall (crude) death rates are 8.92 per 1,000 and 2.67 per 1,000. Is life in Alaska more conducive to longevity than life in Florida?

Although the crude rates suggest that the force of mortality is stronger in Miami than in Alaska, Table 1 reveals that for any given age the two populations have very similar mortality rates. What then accounts for the difference in the crude death rates? A look at the age distributions in Miami and Alaska provides the answer. Compared to Alaska, Miami has a much greater proportion of women in older age groups, where mortality is high. Since the data from larger strata dominate the crude death rate, the Miami death rate is heavily influenced by the high mortality in older ages. In contrast, in Alaska the crude death rate reflects the low mortality rates among young women, who account for a much larger proportion of the Alaska population than they do of the Florida population.

Two populations may have the same overall size and identical age-specific death rates, but different total numbers of deaths and different overall death rates, due to differences in their age distributions. Standardization (and other adjustment procedures) seeks to provide numbers and comparisons that minimize the influence of age and/or other extraneous factors.

Table 1
Population and Deaths by Age in 1970 for White Females in Miami, Alaska, and the U.S.

	Miami			Alaska			U.S.		
Age	Pop.	Deaths	Rate*	Pop.	Deaths	Rate*	Pop.+	Deaths	Rate
< 15	114,350	136	1.19	37,164	59	1.59	23,961	32	1.34
15-24	80,259	57	0.71	20,036	18	0.90	15,420	9	0.58
25-44	133,440	208	1.56	32,693	37	1.13	21,353	30	1.40
45-64	142,670	1,016	7.12	14,947	90	6.02	19,609	140	7.14
65+	92,168	3,605	39.11	2,077	81	39.00	10,685	529	49.51
	562,887	5,022		106,917	285		91,028	740	
Crude death rate*			8.92			2.67			8.13

* Deaths per 1,000 population + in thousands

Standardization of rates by the direct method

In the above example, the difference in crude death rates between Alaska and Miami results from differences in their respective age distributions rather than differential age-specific death rates. It follows intuitively that if Miami had the same age distribution as Alaska, or vice-versa, their crude death rates would be similar to each other. As a matter of fact, if Miami and Alaska had the same age distribution, regardless of what that distribution might be, their crude death rates would be similar, since their age-specific rates are similar.

In direct standardization the stratum-specific rates of study populations are applied to the age distribution of a standard population. (In the above example, each age group is a stratum.) Consequently, if Alaska happened to have the same age distribution of white females as the 1970 U.S. white female population, and Miami also had this same age distribution, then the crude death rates for Alaska and Miami would be similar. In other words, direct standardization applies the same set of weights to the age-specific rates of Alaska and Miami, and the summary (age-adjusted) death rate is therefore independent of differences in the age distribution of the two populations. The directly age-standardized death rates are equivalent to the crude death rates which Miami and Alaska "would have experienced" if they had had the same age distribution as the 1970 U.S. white female population.

Computationally, direct standardization of rates is straightforward:

$$\text{Directly standardized rate} = \frac{\sum (\text{stratum-specific rates} \times \text{standard weights})}{\sum (\text{standard weights})}$$

$$\text{Directly standardized rate} = \frac{(r_1 N_1 + r_2 N_2 + r_3 N_3 + \ldots + r_n N_n)}{(N_1 + N_2 + N_3 + \ldots + N_n)}$$

$$R_s = \frac{\sum(r_k \times N_k)}{\Sigma(N_k)} = \sum (r_k \times \frac{N_k}{\Sigma(N_k)}) = \sum (r_k \times \frac{N_k}{N})$$

$$R_s = \sum (r_k W_k)$$

where:

r_k = rate in k-th stratum of the ***study*** population

N_k = number of persons in k-th stratum of the ***standard*** population

N = <u>total</u> number of persons in the ***standard*** population (ΣN_k)

W_k = weight for each stratum (equal to N_k/N)

$\sum$ means summation over the k strata.

This formula shows that, when the same standard is used, if two study populations have the same age-specific rates (i.e., for each k their R_k's are equal) then their directly standardized rates will be identical, independent of the age distributions in the study populations. The standardized death rate for white Miami women using the 1970 U.S. population of white women as the standard is:

$$\text{Directly standardized rate} = \frac{(1.19 \times 23{,}961) + (0.71 \times 15{,}420) + \ldots + (39.11 \times 10{,}685)}{91{,}208}$$

$$= 6.92 \text{ deaths/thousand}$$

The corresponding standardized rate for Alaska is:

$$\text{Directly standardized rate} = \frac{(1.59 \times 23{,}961) + (0.90 \times 15{,}420) + \ldots + (39.00 \times 10{,}685)}{91{,}208}$$

$$= 6.71 \text{ deaths/thousand}$$

(Results can be expressed as decimal fractions or scaled to aid in their intuitive meaningfulness, e.g., 0.00134 = 1.34 per thousand = 134 per hundred thousand.)

After adjusting for age, the difference in death rates between Alaska and Miami is nearly eliminated.

Some points to consider

There are several things to consider about the above formula and computation. First, the directly standardized rate is a weighted average. Since each W_k is the proportion that the k-th stratum is of the total standard population, the weights are simply the proportional age distribution in the standard population. The crude death rate in a population, which represents the total number of deaths divided by the total number of persons, can be regarded as an average of the population's stratum-specific death rates (R_k) weighted by its own age distribution.

Similarly, a directly standardized rate corresponds to the crude rate that would be observed in the standard population if the standard population had the same stratum-specific rates as does the study population. (To put the foregoing in terms of the above data for Alaska, Miami, and the U.S. population, the crude death rate for Miami (8.92/1,000) can be expressed as a weighted average of the age-specific death rates (1.19, 0.71, etc. per 1,000) for Miami, where the weights are the population proportion in each age stratum (114,350/562,887, 80,259/562,887, etc.). Similarly, the crude U.S. death rate (8.13/1,000) can be expressed as a weighted average of the U.S. age-specific death rates (1.34, 0.58, etc. per 1,000) with weights consisting of the age distribution in the U.S. population (23,961/91,028, 15,420/91,028, etc.). Therefore, if the U.S. as a whole had experienced the death rates shown above for Alaska, then the crude 1970 U.S. death rate would be 6.71 deaths/thousand, i.e., the directly standardized death rate for Alaska.

[Aside: A technical issue that Rothman and Greenland point out but which we will not worry about is that when the above rates are computed using person-years, rather than people, changes in the death rates can lead to changes in person-years. Unless the death rates are the same across all age strata or the changes in person-years do not change the proportional age distribution, then hypothetical statements such as "if the U.S. as a whole had experienced the death rates shown above for Alaska" require the assumption that replacing the death rates would not alter the proportional age distribution.]

Reasons for standardizing rates

Two main motivations encourage the use of standardized rates. First, summary indices from two or more populations are more easily compared than multiple strata of specific rates. This becomes

especially important when comparing rates from several populations or when each population has a large number of strata. Second, small numbers in some strata may lead to unstable specific rates. When sample populations are so small that their strata contain mostly unstable rates and zeroes, the direct standardization procedure may not be appropriate and an alternate procedure (see below) becomes desirable.

Although standardized rates can summarize trends across strata, a considerable amount of information is lost. For example, mortality differences between two populations may be much greater in older ages, or rates for one population compared to another may be lower in young ages and higher in older ages. In the latter case, a single summary measure obscures valuable information and is probably unwise. Furthermore, different standards could reverse the relative magnitude of the standardized rates depending on which age groups were weighted most heavily. The trade-off between detailed information and useful summarization runs through epidemiologic data analysis methods.

Simultaneous adjustment

Rates can be standardized for two or more variables simultaneously. Table 2 compares age and baseline diastolic blood pressure (DBP)-specific incidences of elevated blood pressure (DBP > 90 mm Hg) in light and heavy subjects (relative weight greater and less than 1.25, respectively) among individuals with DBP previously below 90 mm Hg. The combined population is used as the standard to adjust for age and baseline blood pressure differences in the two weight categories. Computations for simultaneous adjustments are essentially identical to those for the single case:

Standardized rate for low weight subjects

$$= [(0.14 \times 80)+(0.31 \times 59)+...+(0.11 \times 36)] / 349 = 0.14$$

Standardized rate for heavier subjects

$$= [(0.30 \times 80)+(0.30 \times 59)+...+(0.59 \times 36)] / 349 = 0.36$$

In this example, the directly standardized rates differ little from the crude rates.

Table 2
Incidence of High Blood Pressure by Baseline Relative Weight, Blood Pressure, and Age in Evans Co., Ga.

Age	Baseline Diastolic Blood Pressure	Relative weight								
		Light			Heavy			Total		
		No.	Cases	Rate	No.	Cases	Rate	No.	Cases	Rate
25-34	Low	70	10	0.14	10	3	0.30	80	13	0.16
	Normal	49	15	0.31	10	3	0.30	59	18	0.31
	Moderate	13	5	0.38	5	4	0.80	18	9	0.50
35-44	Low	67	3	0.04	5	2	0.40	72	5	0.07
	Normal	66	4	0.06	18	4	0.22	84	8	0.10
	Moderate	19	2	0.11	17	10	0.59	36	12	0.33
Total		284	39	0.14	65	26	0.40	349	65	0.19

Spreadsheets

Those of you who are familiar with spreadsheet program (e.g., Lotus 123®, Quattro Pro®, Microsoft Excel®) will readily see the resemblance of the above layout to a spreadsheet. Indeed, spreadsheets are a very convenient method for carrying out a modest number of standardizations. Spreadsheet neophytes will certainly want to learn this method, and even experienced spreadsheet users (who will no doubt want to try this on their own before reading further) may find that creating an age standardization worksheet helps them to learn and understand standardization methods better.

To create the above table in a spreadsheet program, copy the layout, the columns and rows that contain the labels ("35-44", "Moderate", "Light", etc.) and fill in the cells in the first two columns labeled "No." and the two columns labeled "Cases" — but for simplicity of exposition below, do not set aside rows for blank space or horizontal rules or blank columns as separators. If the age categories are placed in column A and the DBP categories in column B, then columns C, D, F, and G (leaving E for the first "Rate" column and "H" for the second) will contain the data for number of participants and number of cases. I will assume that the first row of data (for ages 25-34 years, low diastolic blood pressure) is row **14** (allowing some blank rows for labels and documentation).

To compute the total columns, insert the formula "=C14+F14" into cell **I14** (this corresponds to the number 80 in the table). Upon completing this operation you should see that number appear. Then copy this formula to the rest of the cells in this column (I15-I19) and in the next one (J14-J19). Now, have the spreadsheet compute the row containing the totals for these columns, by using your spreadsheet's summation function to sum the cells in each column. If no rows have been skipped, the summation function will go into cell **C20** and will look something like "@SUM(C14..C19)" [Lotus 123®] or "=SUM(C14:C19)" [Excel®]. Once again you should see the correct total. Then

copy this cell to the other columns to be totaled (D20, F20, G20, I20, J20). (Note to spreadsheet neophytes: spreadsheet programs generally use "relative addressing" by default, so when you copy the formula the program generally adjusts the row and/or column numbers accordingly. Sometimes that's *not* what you want to happen, but in this case it is.)

Then fill in the columns labeled "Rate" by inserting a formula for the ratio of the appropriate "Cases" cell and "No." cell. A simple method for doing this, though not the most elegant, is the following. If the top row of data (25-34 years old, low) is row **14**, the first "No." column is **C**, and the first "Cases" column is **D**, then insert the formula "=D14/C14" into cell **E14** (no rocket science that!). Copy this formula to the remaining cells in the column (E15-E19), and then copy this column to the other two columns labeled "Rate". Your worksheet should now look the same as the table, and you are ready to compute the directly standardized rates.

There are several equivalent ways to proceed. Try this one and then see if you can figure out some others. In the first "Rate" column (i.e., E), a few lines below the "Total" row (e.g., row 26), type in the formula "=E14*I14" (this goes in cell **E26**). This formula multiples the rate for participants who are younger, have low DBP, and are in the lighter relative weight category (**E14**) by the total number of participants who are age25-34 years and have low DBP (**I14**). Then copy E26 to cells E27-E31 and H16-H31.

Each of the latter cells now shows what we might term the "expected number of cases that would occur in each age-DBP stratum of the total participant group if the total group experienced the incidence rates for the lighter-weight participants [for the values in column E] or for the heavier-weight participants [for the values in column H]". Thus, we have only to sum these expected numbers and divide by the total population size. Copy one of the cells that contains a summation function (e.g., C20) to the cell (**E32**) just under the first new column and then copy it (from either C20 or E32) to H32. If the relative addressing works properly, the summation functions should become "=SUM(E26:E31)" and "=SUM(H26:H31)" (or their equivalent in your spreadsheet program). Finally, perhaps on the following row, insert the formulas "=E32/I19" in column E (i.e., in cell **E33**)and "=H32/I19" in column H. You should see the directly standardized rates 0.14 and 0.36, respectively.

If you have faithfully followed the above instructions, you will probably think this is a lot of work to go through for a several-minute task with paper, pencil, and a calculator — even if you have not encountered any difficulties (of your own making or mine). However, this spreadsheet can easily be modified to compute standardized rates for other data, so if you can find it when the need arises it may come in very handy. For now, though, it's probably worthwhile using both calculator and spreadsheet in order to master the computations and concepts.

Standardized ratios and differences

Rates that have been standardized by the direct method, using the same standard population, may be compared in relative or absolute terms (i.e., as a ratio or as a difference). For example, we can obtain a "Standardized Rate Ratio" ("SRR") by dividing the (directly) standardized rate for Miami by that of Alaska. Using the values computed above:

$$SRR = \frac{\text{directly standardized rate for Miami}}{\text{directly standardized rate for Alaska}} = \frac{6.92}{6.71} = 1.03$$

Similarly, the difference of the two rates would be a "standardized rate difference" (SRD = 6.92–6.71=0.21 [per 1,000 – the ratio has no need for the scaling factor, but the difference does). Since the rates are virtually identical, the SRR is close to 1.0, and the SRD is close to zero, all give the same message: the mortality experience in Alaska, Miami, and the total U.S. are all about the same when the differences due to age structure are eliminated.

In addition, a directly standardized rate can be compared to the crude rate in the population from which the weights were taken (the "standard population"). The reason that this works is that, as noted above, the crude rate for a population can be expressed as a weighted average of the population's stratum-specific death rates (R_k) weighted by its own age distribution. Therefore the crude rate and the directly standardized rates are all weighted averages based on the same set of weights (the proportional age distribution in the standard population). So the following SRR is legitimate:

$$SRR = \frac{\text{directly standardized rate for Alaska}}{\text{directly standardized rate for total U.S.}} = \frac{6.92}{8.13} = 0.852$$

Standardized ratios and differences are also weighted averages [optional]

It may or may not be of interest to know that the standardized ratios and differences obtained by taking the ratios and differences of directly-standardized rates are also weighted averages. For example, the SRR can be written as:

$$SRR = \frac{\sum (r_k W_k)}{\sum_j (r'_j W_j)} = \frac{\sum (r_k / r'_k)(r'_k W_k)}{\sum_j (r'_j W_j)} = \frac{\sum [(RR_k)(r'_k W_k)]}{\sum_j (r'_j W_j)}$$

$$= \sum_k \left[(RR_k) \left(\frac{(r'_k W_k)}{\sum_j (r'_j W_j)} \right) \right] = \sum_K (RR_k W'_k)$$

where the RR_k are the stratum-specific rate ratios and the expression in parenthesis is the stratum-specific weight, W'_k for the SRR.

Nonuniformity of stratum-specific rates

Before computing and reporting standardized measures, we should ask the question that applies to any summary measure: does the summary conceal important heterogeneity. If one population has higher rates in some strata but lower rates in others, and stratum sizes are large enough for these differences to be worth paying attention to, then a comparison of standardized rates for the two populations could conceal an important feature of the data. In such a situation, it is important to report the nonuniformity of the stratum-specific rate comparisons and to consider whether computing standardized rates and ratios serves any purpose.

Sparse data

Even though standardized rates can be computed, they are not always meaningful. Use of the same set of weights to average the stratum-specific rates guarantees comparability, but for the comparisons to be meaningful there must also be large enough numbers in all important strata ("important" means those constituting substantial weight in the standardization procedure). Otherwise the stratum-specific rate estimates will be too unstable (i.e., imprecise), and weighting them may only amplify that instability. For example, a rate of 0.10 based on two cases becomes only half as large, 0.05, if two more cases are found. Although the difference between these two rates is small, if they happened to fall in a stratum for which the standard population had a particularly large proportion, then this small difference would be magnified (relative to the other rates) in the standardized rate. There are various rules of thumb for what constitutes "large enough", such as at least 10 or 20 events (e.g., deaths, cases) and a denominator of at least 100, though a specific situation might call for substantially larger numbers.

Indirect standardization

When stratum-specific numbers are small, as is often the case in such populations as a single industrial plant or a small city, stratum-specific rate estimates are too susceptible to being heavily influenced by random variability for the direct standardization method to be satisfactory. Instead, an "indirect" standardization procedure is often used and a "standardized mortality ratio" ("SMR") computed. (The standard mortality difference, computed as the indirectly standardized rate minus the crude rate from the standard population, is also theoretically of interest).

Indirect standardization avoids the problem of imprecise estimates of stratum-specific rates in a study population by taking stratum-specific rates from a standard population of sufficient size and relevance. These rates are then averaged using as weights the stratum sizes of the study population. Thus, the procedure is the mirror-image of direct standardization. In *direct standardization*, the study population provides the rates and the standard population provides the weights. In *indirect standardization*, the standard population provides the rates and the study population provides the weights. (For this reason Ollie Miettinen employs the terms "externally standardized" and "internally standardized", respectively, for what we are calling direct standardization and indirect standardization.)

	Study population	Standard population
Directly-standardized rate	Rates	**Weights**
Indirectly-standardized rate	**Weights**	Rates

We have seen that directly-standardized rates (computed using the same standard population) can be readily compared to each other and to the standard population, because all are based on the set of same weights (those from the *standard population*). **However**, comparison of indirectly-standardized rates can be problematic, because each study population's standardized rate is based on its own set of weights. In fact, the only comparison that is always permissible is the comparison between the study population and the standard population, since these indirect rates are both based on weights from the study population.

Directly-standardized rates are based on one set of weights;
indirectly-standardized rates are based on multiple sets of weights

	Study pop. A	Study pop. B	Standard population
Directly-standardized rate	Rates-A	Rates-B	**Weights**
Indirectly-standardized rate	**Weights-A**	**Weights-B**	Rates

As the above table illustrates, the directly-standardized rates for the three populations are based on the same set of weights, but the indirectly-standardized rates for each study population are based on its own age distribution. This lack of comparability of indirectly standardized rates (or of SMR's), and as long as the study populations have similar age distributions then there is not necessarily a practical problem. However, if the age (or other) distributions *do* differ across the study populations, then comparison of the indirectly-standardized rates could be no better than comparison of the crude rates themselves.

Carrying out indirect standardization

Indirect standardization can be thought of as taking the observed number of deaths or events in the study population and comparing that number to an "expected" number of deaths, i.e., the number of deaths that would be expected in the study population if its mortality experience (its stratum-specific rates) were the same as for the standard population. The ratio of observed to expected deaths is termed the Standardized Mortality Ratio (or Standardized Morbidity Ratio if disease, rather than death, is the outcome), abbreviated SMR, and it, rather than standardized rates, is the usual product of the indirect standardization procedure.

The **expected number of deaths** is obtained as follows:

$$\text{Expected number of deaths} = \sum \left(\begin{array}{c}\text{[Stratum-specific rates from the standard population]}\end{array} \times \begin{array}{c}\text{[stratum sizes from the study population]}\end{array} \right)$$

$$= \sum (R_k n_k)$$

and the observed number of deaths is $\sum d_k$

$$\text{so that} \qquad \text{SMR} = \frac{\text{Observed deaths}}{\text{Expected deaths}} = \frac{\sum d_k}{\sum (R_k n_k)}$$

where d_k = number of deaths in the k-th stratum of the ***study*** population ("observed deaths")

n_k = size of the k-th stratum of the ***study*** population

R_k = death rate in the k-th stratum of the ***standard*** population

The number of observed deaths can also be expressed as the sum of stratum-specific death rates multiplied by stratum sizes:

$$\textbf{Observed number of deaths} = \sum \left(\begin{array}{c}\text{[Stratum-specific rates from}\\ \text{the } \textit{study} \text{ population]}\end{array} \times \begin{array}{c}\text{[stratum sizes from}\\ \text{the } \textit{study} \text{ population]}\end{array} \right)$$

$$= \sum (r_k n_k)$$

where: r_k = death rate in the k-th stratum,

Thus, the SMR can be readily expressed as a ratio of two weighted averages of stratum-specific death rates, where the weights are the proportionate stratum sizes of the ***study*** population:

$$\text{SMR} = \frac{\text{Observed deaths}}{\text{Expected deaths}} = \frac{\sum (r_k n_k)}{\sum (R_k n_k)} = \frac{\sum (r_k w_k)}{\sum (R_k w_k)}$$

where n_t is the total size of the study population and w_k gives the proportionate stratum sizes, computed as n_k/n.

The SMR indicates the relative excess or decrement in the actual mortality experience in the study population with respect to what might have been expect had it experienced the force of mortality in the standard (or reference) population. [The denominator of the SMR is not precisely the "expected mortality" when the stratum sizes are in person-years (see Rothman and Greenland, 1998:234, but for our purposes it is close enough.]

Comparison of SMR's

As noted above, the comparison of SMR's (or, equivalently, indirectly-standardized rates) from different study populations is complicated by the fact that the weights used in obtaining the indirectly standardized rates are the stratum sizes of the individual study populations rather than of a (common) standard population. Technically, therefore, one cannot compare SMR's unless the distribution of the standardization variable (e.g., age) is identical across the study populations, in which case standardization is unnecessary since the crude death rates could have been compared directly. Even if two populations have identical stratum-specific rates and therefore their *directly* standardized rates are identical, their *indirectly* standardized rates can be quite different (see example below). Remember, however, that the usual reason for using indirect standardization is that the stratum-specific rate estimates are very imprecise, making directly standardized rates problematic.

Strictly speaking, SMR's can be validly compared across populations with different age distributions in only *one special case*—the situation where the stratum-specific rates in each population are uniform, i.e., they do not vary by age. In this case the weights or age distribution is irrelevant: the average of a set of identical rates will always be the same regardless of the set of weights that are used. If the stratum-specific rates or ratios are reasonably uniform—and if they are widely disparate the usefulness of a single average is somewhat questionable—then a comparison of indirectly standardized rates may be reasonable though admittedly technically improper. If the rates are uniform, however, then the weighting will make little difference so there may be no need to standardize at all.

The following example provides a numerical illustration of the problem of comparing SMR's:

Table 3
Death rates by age in two occupations and a standard population

	Occupation A			**Occupation B**			**Standard population**		
Age	Persons	Deaths	Rate	Persons	Deaths	Rate	Persons	Deaths	Rate
40-49	1,000	2	0.002	5,000	10	0.002	30,000	30	0.001
50-59	5,000	20	0.004	1,000	4	0.004	40,000	120	0.003
Total	6,000	22		6,000	14		70,000	150	

SMR	$\frac{22}{(0.001)(1{,}000)+((0.003)(5{,}000)}$	$\frac{14}{(0.001)(5{,}000)+(0.003)(1{,}000)}$
	1.38	1.75

Though both occupations have exactly the same stratum-specific rates, their SMR's differ, due to the substantially different age distributions for the two occupations. However, the directly standardized rates for both occupations are, reassuringly, the same:

Directly standardized rate for A = (0.002 x 30,000 + 0.004 x 40,000) / 70,000 = 0.0031

Directly standardized rate for B = (0.002 x 30,000 + 0.004 x 40,000) / 70,000 = 0.0031

Similarly, the SRR for each occupation relative to the standard population is 0.0031/0.0021 = 1.48, indicating a 48% higher age-standardized rate of death in each occupational population compared to the standard population. However, the apparent equivalence of the directly standardized rates is misleading. With so few deaths in the younger age stratum in Occupation A and in the older age stratum in Occupation B, the rate estimates are very unstable. In other words, we cannot really estimate some of the rates, so direct standardization is a dubious procedure. Given the substantial uncertainty about what the stratum-specific rates really are, the only conclusion we can be confident of is that both occupations have elevated mortality rates compared to the standard, or reference population. Without assumptions or additional information, we have no evidence from standardization to conclude that one of the occupations is more hazardous (or is not more hazardous) than the other.

Indirectly Standardized Rates (optional topic)

Though not commonly seen, an indirectly standardized rate can be obtained from an SMR as follows:

$$\text{Indirectly-standardized rate} = \text{SMR} \times \left(\begin{array}{c} \text{Crude death rate in} \\ \text{the } \textit{standard} \\ \text{Population} \end{array} \right)$$

The logic for this relationship is that the SMR gives a standardized comparison of the mortality experience in a study population compared to that in the standard population0. So, for example, if the study population has twice the mortality rate of the standard population, the standardized rate for the study population should be twice the observed (crude) death rate in the standard population.

An alternate (and algebraically equivalent) strategy is to multiply the crude death rate from the study population by a "standardizing factor" consisting of the ratio of the crude rate in the standard population to an "index death rate". This "index death rate" is the death rate that would be expected in the study (index) population, due to its age distribution, if in each stratum the corresponding death rate from the standard population applied, i.e., the expected number of deaths divided by the study population size.

$$\text{Indirectly-standardized rate} = \begin{array}{c} \text{Crude death rate} \\ \text{in the } \textit{study} \\ \text{Population} \end{array} \times \text{Standardizing factor}$$

$$= \begin{array}{c}\text{Crude death rate}\\ \text{in the } \textit{study}\\ \text{population}\end{array} \times \frac{\text{Crude death rate in the standard population}}{\text{Index death rate}}$$

where the index death rate is:

$$= \sum \left(\begin{array}{c}\text{Stratum-specific}\\ \text{rates in the}\\ \textit{standard} \text{ population}\end{array} \times \frac{\text{Stratum sizes from study population}}{\text{Total size of study population}} \right)$$

Algebraically, this may be written:

$$\text{Indirectly-standardized rate} = r \times \frac{R}{\sum (R_k n_k)/n} \left[\frac{\text{Crude death rate in the standard population}}{\text{Index death rate}} \right]$$

and may be reformulated:

$$\text{Indirectly-standardized rate} = \frac{d}{\sum (R_k n_k)} \times R$$

$$\text{Indirectly-standardized rate} = SMR \times R$$

where:

R = crude death rate in the standard population

R_{k_i} = death rate in the k-th stratum of the standard population

r = crude death rate in study population

n_k = size of the k-th stratum of the study population

n = size of the study population

d = total deaths in the study population

Example:

If we use the U.S. rates (from table 1) as a standard, the indirectly standardized death rate for Miami is:

$$\text{Indirectly standardized rate} = \frac{5{,}022}{(1.34^* \times 114{,}350) + (0.58^* \times 80{,}259) + \ldots + (49.51^* \times 92{,}168)} \times 8.13$$

$$= 6.84 \text{ deaths/thousand}$$

*(Per 1,000 population)

For Alaska, the indirect standardized rate is:

$$\text{Indirectly standardized rate} = \frac{285}{(1.34^* \times 37{,}164) + (0.58^* \times 20{,}036) + \ldots + (49.51^* \times 2{,}077)} \times 8.13$$

$$= 7.32 \text{ deaths/thousand}$$

*(Per 1,000 population)

The indirectly standardized rate can be viewed as the study population's crude death rate standardized for the relative "*a priori* mortality proneness" of the study population versus the standard population.

(Returning to basics here)

Table 4
Crude and Age-Standardized* 1970 Death Rates Per 1000 for White Females in Alaska, Miami, and the U.S.

	Alaska	Miami	U.S.
Crude	2.67	8.92	8.13
Direct	6.71	6.92	-
Indirect	7.23	6.84	-

*Standard population is 1970 U.S. white females

Table 4 summarizes the results and indicates that the type of standardization makes a modest difference in this example; the directly standardized rates for Miami and Alaska are closer than their indirect counterparts.

Notice that the age-specific rates from Alaska and Miami do not enter the indirect standardization computations at all. The information which they contain enters indirectly (hence the procedure name), since the observed number of deaths is partly determined by the age-specific rates. But the observed number of deaths is also determined by the stratum sizes.

Choice of Standard Population

Standardized measures describe a hypothetical state of affairs, which is a function of the standard population chosen. For direct age-standardization, the total U.S. population from the previous census is especially common. Since rates standardized to the same external standard are comparable, the selection of a commonly used standard has advantages when comparing rates across different studies. Sometimes investigators compute directly standardized rates based upon one of their own study populations as the standard or by combining two or more study populations to create a standard. But rates standardized to a specific study population are not as readily compared to rates from other studies.

When a study involves a comparison with a "control" population, the choice of a standard should reflect the study goals. For example, an examination of county mortality variation within a state might compare county mortality to the state as a whole. A clean industry may be a good standard for an industrial population exposed to suspected occupational health hazards. Since indirectly standardized measures require knowledge of stratum-specific rates in the standard, data availability constrains the choice.

The choice of a standard population is not always obvious, and there may not be a "best" choice. For example, in comparing syphilis rates across counties in North Carolina, Thomas et al. (1995) decided to standardize the rates by age and sex to reduce the influence of different age-sex distributions in different counties. One obvious choice for a set of weights was the age-sex distribution of North Carolina as a whole. However, another possible choice was to use the age-sex distribution for the U.S. as a whole, so that other investigators could more readily compare syphilis rates in their states to the rates presented in the article. Was there a "right" answer? In this case the choice between the two standards could be regarded as a choice between greater "relevance" and broader comparability. The net result makes little difference, however, since the age-sex distribution of North Carolina and the entire U.S. are very similar. In other situations, however, the choice of standards can indeed change the message conveyed by the results.

Just as the growth of knowledge leads to revisions to disease classification systems, thereby complicating comparisons across revisions, changes in the age distribution over decades creates the dilemma of switch to a new standard population to reflect the present reality versus retaining the existing standard to preserve comparability across time. For this reason mortality rates in the United States have been standardized to the 1940 population distribution almost to the end of the 20th century. Other standards (1970, 1980) were also in use, however, complicating comparisons of mortality statistics. During the 1990's, the U.S. National Center for Health Statistics (NCHS/CDC) coordinated an effort among federal and state agencies to adopt the year 2000 projected U.S. population for standardization of mortality statistics. In August 1998 all U.S. Department of Health and Human Services (DHHS) agencies were directed to use the 2000 Standard Population for age adjusting mortality rates beginning no later than data year 1999 (Schoenborn et al., 2000).

Since the age distribution in 2000 is shifted to the right (older ages) compared to the 1940 population, mortality rates standardized to the 2000 population will be higher than if they were standardized to the 1940 census because they will assign more weight to older age strata, where mortality rates are high. In the same way, comparisons (e.g., ratios) of standardized rates will reflect

the situation among older age groups more than in the past. To be sure, the switch will make comparisons to past data problematic, though NCHS will recompute age-standardized mortality rates for past years based on the 2000 population standard.

The opposite result will occur when at some point it is decided that in a global society all countries should standardized their rates to the World population, to facilitate comparison across countries. Since the large majority of the world's population live in developing countries and is much younger than the population of the U.S. and other developed countries, standardization using a world standard will yield lower standardized rates for most causes of death. As illustrated by the fruit stand example in the beginning of this chapter, different standards can give different, but correct, results. Comparisons, the usual goal of examining rates, may be less affected then the rates themselves, as long as the patterns (e.g., rise in mortality rate with age) are the same in the populations being compared. When that is not the case, then the question of whether it is meaningful to compare summary measures at all becomes more important than the question of which weights to use.

Key concepts

- Populations are heterogeneous – they contain disparate subgroups. So any overall measure is a summary of values for constituent subgroups. The underlying reality is the set of rates for (ideally homogenous) subgroups.
- The observed ("crude") rate is in fact a weighted average of subgroup-"specific" rates, weighted by the size of the subgroups.
- Comparability of weighted averages depends on similarity of weights.
- "Standardized" (and other kinds of adjusted) measures are also weighted averages, with weights chosen to improve comparability.
- Crude rates are "real", standardized rates are hypothetical.
- The "direct" method (weights taken from an external standard population) gives greater comparability but requires more data.
- The "indirect" method (weights taken from the internal study population) requires fewer data but provides less comparability.
- Choice of weights can affect both rates, comparisons of rates, and comparability to other populations, so the implications of using different possible standard populations should be considered.
- Any summary conceals information; if there is substantial heterogeneity, the usefulness of a summary is open to question.

Bibliography

Rothman, Kenneth - *Modern Epidemiology*, Chapter 5 and pages 227-2290. Lilienfeld and Lilienfeld - *Foundations of epidemiology*, Measures of mortality, pp. 71-80; Mausner & Kramer - *Epidemiology: an introductory text*, pp. 338-344.

Inskip, Hazel; Valerie Beral, and Patricia Fraser. Methods for age adjustment of rates. *Statistics in Medicine* 1983; 2:455-4660.

Gaffey WR: A critique of the standardized mortality ratio. *J Occup Med* 18:157-160, 1976.

Schoenborn, Charlotte; Richard Klein, Virginia Fried. Age adjustment of National Center for Health Statistics data using the 2000 projected U.S. population with emphasis on survey data systems. Division of Health Interview Statistics, NCHS/CDC. To appear in the *Healthy People Statistical Note* series (http://www.cdc.gov/nchs/products/pubs/workpap/ageadjust.htm, 9/10/00).

Thomas, James C.; Alice L. Kulik, Victor J. Schoenbach. Syphilis in the South: rural rates surpass urban rates in North Carolina. *Am J Public Health* 1995; 85:1119-1122

Tsai SP, Hardy RJ, Wen CP. The standardized mortality ratio and life expectancy. *Am J Epidemiol* 1992; 135:824-831.

Appendix on Standardized Mortality Ratios

(courtesy of Raymond Greenberg, M.D.,Ph.D.

I. DEFINITION. The Standardized Mortality Ratio (SMR) is a measure of mortality in a study population, relative to mortality in a reference population. The SMR answers the following question: "How does the number of observed deaths compare with the expected number of deaths, if our study group had the age-specific mortality rates of the reference population during these study years?"

II. CALCULATION. In principle, any reference population yielding sufficiently precise rates can be used to obtain the expected death number, but it is customary to use the general population. The SMR is given by the following expression:

$$\text{SMR} = \frac{\text{Observed deaths in study population}}{\text{Expected deaths in study population}}$$

The SMR is usually scaled up by multiplying it by 100. An SMR over 100 indicates that more deaths were observed than expected (i.e., the study population had a relatively poor outcome). An SMR less than 100 means that fewer deaths were observed than expected (i.e., the study population had a relatively favorable outcome). Obviously, the value of the SMR will depend on the choice of the reference population used for the comparison mortality rates. If the reference population is healthy, they will have low mortality rates and thereby increase the SMR. Conversely, if the reference population is unhealthy, they will have high mortality rates and thereby decrease the SMR. It is therefore crucial to choose an appropriate reference population or at least to know in which direction the reference population differs from an appropriate one.

III. HEALTHY WORKER EFFECT. The SMR is frequently used to examine mortality in an industrial plant or industry. However, when workers are compared to the general population, it is common to find lower mortality rates in the workers (SMR less than 100). The reason is thought to be that the general population includes people who are too sick to work. The elevated mortality in such people raises the mortality rate of the general population, so that mortality in the general worker population is lower. This phenomenon is called the healthy worker effect. The healthy worker effect is an important consideration primarily for mortality from diseases, such as cardiovascular disease, where an extended period of physical limitation or disability frequently precedes death and thus affects entrance into and remaining in the workforce.

IV. SAMPLE CALCULATION: Suppose you are studying male textile workers between the ages of 20 and 39 years between the years 1960 and 1979.

a. Observed deaths

	Period		
Age	1960-1969	1970-1979	Total
20-29	1	2	3
30-39	2	3	5
Total			8

b. Person-years of exposure

	Period		
Age	1960-1969	1970-1979	Total
20-29	1,000	500	1,500
30-39	500	1,000	1,500

c. Mortality rates from reference population

	Period	
Age	1960-1969	1970-1979
20-29	1/1,000py	2/1,000py
30-39	2/1,000py	4/1,000py

d. Expected deaths (b x c)

	Period		
Age	1960-1969	1970-1979	Total
20-29	1	1	2
30-39	1	4	5
Total			7

$$\text{SMR} = \frac{\text{Observed Deaths in Study Population}}{\text{Expected Deaths in Study Population}} \times 100 = \frac{8}{7} \times 100 = 114$$

or a 14% elevation in mortality.

V. CAUTIONS IN USE OF SMR:

a. An SMR is an indirect standardization procedure (standard rates applied to study population) and therefore two SMR's cannot be compared, unless their respective populations have the same age distribution (in which case, why standardize). [If the age distributions are not markedly different or the relationships in mortality rates between the populations are similar

across age strata, then the damage is not great. The latter possibility can only rarely be checked, of course, since SMR's are typically computed in situations where there are too few deaths in each stratum to calculate meaningful stratum-specific rates.]

b. SMR's do not readily translate into life-expectancy (though recent work provides an approximation).

c. As length of follow-up increases, an SMR based on cumulative mortality tends toward 100.

(See Gaffey WR: A critique of the standardized mortality ratio. *J Occup Med* 18:157-160, 1976

Standardization of rates and ratios - Assignment

1. From the data in the table below, compute for each sex separately (for Rateboro) and for the United States (both sexes) the following measures. Write your answers (rounded to 4 decimal places) in the table; show all work for (c) and (d).

   a. crude death rates

   b. age-specific death rates

   c. directly-standardized death rates for Rateboro males and females (separately) using the U.S. population as a standard.

   d. indirectly standardized death rates as in (c).

Population and Deaths in 1980 in Rateboro
Adults by Age and Sex and U.S. Total
(hypothetical data)

	Rateboro						United States		
	Males			Females			Both Sexes		
Age	Pop.	Deaths	Rate	Pop.	Deaths	Rate	Pop*	Deaths*	Rate
18-34	900	6		800	1		60,000	90	
35-59	800	3		800	5		45,000	270	
60-74	300	15		500	10		20,000	600	
75 +	200	22		500	38		15,000	1500	
Total	2200	46		2600	54		140,000	2460	

(*In thousands. Population and deaths for Rateboro are actual figures.)

Direct standardized rate:

Indirect standardized rate:

2. Based on the results for question 1.:

   a. Do males or females have a more favorable mortality experience in Rateboro? Cite the rates or other figures on which you have based your decision.

   b. How do you account for the similarity in the crude death rates for Rateboro males and females?

   c. Briefly discuss the reasons for and against (i) rate adjustment, and (ii) direct versus indirect methods--in these data.

---

* Thanks to Barbara Richardson, Ph.D. for the first version of this question.

d. How would you feel about the conclusion, by an experienced epidemiologist, that "the Rateboro data are generally consistent with the typical finding of a more favorable mortality experience of U.S. females; the anomolous result for the 35-59 year-old group, with the high death rate among females (more than 50% greater than the rate for males) is evidence that the Rateboro environment is more suitable for males in the age range 35-59 than for females."

3. The following extract from "Breast cancer in women after repeated fluoroscopic examinations of the chest" (John D. Boice, Jr., and Richard R. Monson, *J Natl Cancer Inst* 59:823-832, 1977) describes their adjustment procedure:

> "...Expected breast cancer cases were determined with the use of age-calendar year specific incidence rates of Connecticut (refs), a neighboring State whose cancer registry has been in existence since 1935. The years at which a woman was at risk for breast cancer development (i.e., the years after sanitarium admission or fluoroscopy exposure) were computed separately for each 5-year age group, each 5-year period since start of observation, and each quinquennium from 1930 to 1970 through 1974 and for the six month period from January 1975 through June 1975. Multiplication of the age-calendar year specific WY [women-years] at risk by the corresponding Connecticut incidence rates determined the number of expected breast cancers."

a. What method of adjustment is being used, direct or indirect?

b. The following tables show hypothetical data from a follow-up study like that done by Boice and Monson. Why is it not possible to calculate from the information below the number of breast cancer cases expected for the period 1950-1969 with the method used by Boice and Monson (as described above)? (Note: this is a "sticky" question. Do not try to calculate or derive numbers.)

Distribution of Women-Years (WY) among exposed subjects

	Period			
Age	1950-54	1955-59	1960-64	1965-69
30-34	1900	--	--	--
35-39	1800	1700	--	--
40-44	1700	1600	1500	--
45-49	1600	1500	1400	1300

Average breast cancer incidence rates from the Connecticut Cancer Registry (1950-1969), by age (rate per 1000 WY)

Age (years)	Rate
30-34	.2
35-39	.4
40-44	.8
45-49	1.2

c. What advantage does this adjustment procedure have over simple age adjustment?

4. Tuberculosis (TB) has been called the "captain of all men of death" because of its ability to decimate populations. Improvements in the physical conditions of life in the present century, especially nutrition, housing, and the work environment, greatly reduced this scourge even before the advent of effective chemotherapy for the mycobacterium. The discovery of isoniazid and its effectiveness in reducing infectiousness led to the application of public health measures for tracing and treating active cases, thereby effectively controlling TB in the United States and other developed countries. Indeed, U.S. public health policy has set the elimination of TB by the year 2010 as a goal.

   However, TB incidence in U.S. minority populations has never been reduced to the same extent as the overall U.S. incidence, and the ratio of TB risk in nonwhites to whites has grown steadily from about 3 in the mid-1950s to over 5 in the mid-1980s. In 1986, however, the long-term decline in TB was reversed, with an estimated 9,226 cases in 1985-87 beyond those projected from the 1981-84 trend. The 25-44 year age group had the largest 1985-87 increase, made up of a 17% increase among non-Hispanic blacks and 27% among Hispanics. The HIV epidemic has been implicated in the upswing in tuberculosis; poverty, homeless, and immigration of persons from higher TB areas may also have a role. [Source: Reider HL, Cauthen GM, et al. Tuberculosis in the United States. *JAMA* 1989 (July 21); 262(3):385-389.]

   In this question, you are asked to interpret data from three North Carolina counties. The following tables show the number of TB cases during the period January 1, 1986 to December 31, 1990, the mean population during that time period, and the corresponding U.S. TB rates.

Cases of tuberculosis in three N.C. counties during January 1, 1986 - December 31, 1990

County	White males	White females	Nonwhite males	Nonwhite females
Johnston	11	8	43	13
Orange	5	3	3	4
Wilson	6	10	51	27

Source: NC TB Control Branch

Mean population sizes of three N.C. counties during January 1, 1986 - December 31, 1990

County	White males	White females	Nonwhite males	Nonwhite females
Johnston	31,721	33,955	6,910	8,078
Orange	34,542	37,649	7,510	8,753
Wilson	19,844	22,259	10,692	12,788

Source: (Log Into North Carolina [LINC] database)

Mean annual incidence of tuberculosis,
United States, January 1, 1986 to December 31, 1990

	White males	White females	Nonwhite males	Nonwhite females
Cases per 100,000	7.4	3.6	39.2	19.8

Source: Centers of Disease Control, *Tuberculosis in the United States*

Your interpretation should compare the counties to each other and to the U.S. Is there a greater-than-expected TB incidence in any of the counties? Is an increase confined to particular race-sex-groups?

Suggestions:

a. Compute the race-sex-specific TB rates for each county and overall.

b. Compute an SMR comparing each county to the national TB rates.

5. .This question is optional. If you like it, do it; if you don't like it, forget it! Show that:

   a. if age-specific rates for group A are all equal and age-specific rates for group B are all equal (but not equal to those in group A, i.e., $r_{ai} = r_a$ and $r_{bi} = r_b$ for all i), then:

$$\frac{\text{Directly standardized rate for A}}{\text{Directly standardized rate for B}} = \frac{\text{Crude rate for A}}{\text{Crude rate for B}}$$

   Under what conditions will this ratio equal the ratio of indirect standardized rates?

   c. if age-specific rates in groups A and B are not all equal, but for each stratum

$$\frac{r_{ai}}{r_{bi}} = K \qquad \text{[Where K is the same for all strata]}$$

   then SMR (for A using B as the standard) = K

   d. If the proportional age distributions in two populations are identical, then direct adjustment, indirect adjustment, and crude rates are all comparable between the two populations.

6. (Optional) Solve problem #1 using a computer spreadsheet.

Standardization of Rates and Ratios - Assignment solutions

1. a & b

Population and Deaths in 1980 in Rateboro Adults by Age and Sex and U.S. Total (hypothetical data)

	Rateboro						United States		
	Males			Females			Both Sexes		
Age	Pop.	Deaths	Rate	Pop.	Deaths	Rate	Pop*	Deaths*	Rate
18-34	900	6	.0067	800	1	.0013	60,000	90	.0015
35-59	800	3	.0038	800	5	.0063	45,000	270	.0060
60-74	300	15	.0500	500	10	.0200	20,000	600	.0300
75 +	200	22	.1100	500	38	.0760	15,000	1500	.1000
Total	2200	46	.0209	2600	54	.0208	140,000	2460	.0176

(*In thousands. Population and deaths for Rateboro are actual figures.)

Calculations:

c. Directly standardized death rates for Rateboro males and females (separately) using the U.S. population (both sexes) as a standard.

$$\text{Directly standardized rate} = \frac{\Sigma(r_t N_t)}{N_t}$$

$$\text{Male rate} = \frac{[(.0067 \times 60{,}000) + (.0038 \times 45{,}000) + (.05 \times 20{,}000) + (.11 \times 15{,}000)]}{140{,}000}$$

$$= 0.0230, \text{ or } 23 \text{ deaths per thousand}$$

$$\text{Female rate} = \frac{[(.0013 \times 60{,}000) + (.0063 \times 45{,}000) + (.02 \times 20{,}000) + (.076 \times 15{,}000)]}{140{,}000}$$

$$= 0.0136, \text{ or } 13.6 \text{ deaths per thousand}$$

d.

$$\text{Indirectly standardized rates:} = \frac{d_t}{\Sigma(R_i n_i)} R_t$$

$$\text{Male rate} = \frac{46}{[(0.0015 \times 900) + (0.006 \times 800) + (0.03 \times 300) + (0.1 \times 200)]} (.0176)$$

$$= 0.0230, \text{ or } 23 \text{ deaths per thousand}$$

[the similarity to the directly-standardized rate is coincidental.]

$$\text{Female rate} = \frac{54}{[(0.0015 \times 800) + (0.006 \times 800) + (0.03 \times 500) + (0.1 \times 500)]} (.0176)$$

$$= 0.0134, \text{ or } 13.4 \text{ deaths per thousand}$$

[the similarity to the directly-standardized rate is coincidental.]

2.

a. Females have a more favorable mortality experience. Although the crude death rates for males and females are very close (20.9/1000 vs. 20.8/1000), when age-standardized (direct or indirect) rates are compared, the lower death rates for women are clear.

   i. direct: 23 deaths/1000 (men) vs. 13.6 deaths/1000 (women)

   ii. indirect: 23 death/1000 (men) vs. 13.4 deaths/1000 (women)

b. The similarity in the crude death rates is a function of the respective age distributions of males and females in Rateboro. A greater proportion of women are found in the older age groups, where the morality rates are higher. The crude death rate gives more weight to these larger strata.

c. i. Reasons for rate adjustment are:

- adjustment procedures attempt to permit valid comparisons by minimizing the effect of extraneous variables (e.g., age) that are differientially distributed across the populations of interest;
- summary indices from two or more populations are more easily compared than multiple strata with specific rates; and
- small numbers in some strata may lead to unstable rates.

ii. Disadvantages of adjustment are:

- information is lost when summary measures are used (opposing trends in subgroups may be masked);
- the absolute and relative magnitudes of the standardized rates will depend on the standard used (i.e., the age groups weighted most heavily); and
- standardized rates are fictitious – they do not estimate any "true" parameter.

ii. Direct vs. indirect methods: indirect methods of adjustment are used when the numbers of deaths in the individual strata are too small to yield meaningful rates. The major disadvantage of indirectly standardized rates is that they can properly be compared only to the crude rate in the standard population (that is, it is technically incorrect to compare the indirectly standardized rates for males to the indirectly standardized rates for females as was shown in 2.a.2 above). Conversely, the major advantage of using direct adjustment is that the standardized rates are comparable to one another if they were based on the same standard weights. However, in several of the strata the numbers of observed deaths are small (e.g., 1,3 , 5 and 6), so the estimates of the rates for those strata are imprecise (likely to be heavily influenced by random error) and therefore weighting them in a direct adjustment procedure is hazardous.

d. Agree with the first part (consistency of Rateboro experience and U.S.) but question the second part (Rateboro environment more suitable for males age 35-59) since the rates cited are based on only 3 male and 5 female deaths and are therefore too imprecise to warrant such a conclusion.

3.

a. Indirect adjustment was used, as age-calendar-year-specific rates from a standard population (Connecticut) were applied to the age-calendar-year distribution (of women-years) in the study population. Here is a detailed explanation:

For the indirect standardization or adjustment procedure, "standard rates" were obtained from the Connecticut population. These rates were both age-specific and calendar-year specific, to control for changes in incidence over time. Thus, a table of standard rates like the following would have been used:

Breast cancer incidence (per 100,000 Connecticut women per year) (hypothetical data)

	Period					
Age	1935-39	1940-44	1945-49	1950-54	1955-59	etc.
30-34	20	22	26	28	30	
35-39	30	33	35	38	40	
40-44	50	54	57	59	62	
45-49	70	72	75	78	81	

Source: Connecticut Cancer Registry (1950-1969)

The second ingredient for an standardized rate is the weight. The weight could be population or population-time (person-years, or in this case, women-years). Boice and Monson tell us that they computed women-years within 5-year age groups and 5-year calendar time intervals (quinquennia) (which is why the above table is constructed as it is). Boice and Monson also divided the follow-up period for each woman into 5- (their lucky number!?) year intervals since the start of observation (sanitarium admission or fluoroscopy exposure) for the women. Dividing up the follow-up period is not part of the adjustment procedure, but enables the investigators to analyze the results for different lengths of follow-up after exposure. Thus the investigators can allow for latency periods in cancer development.

Suppose that the distribution of women-years for all women followed between 11 and 15 years after admission or exposure was:

Distribution of Women-Years (WY) among exposed subjects
between 11 and 15 years (inclusive) following admission or exposure
(hypothetical data)

	Period					
Age	1935-39	1940-44	1945-49	1950-54	1955-59	etc.
30-34	1900	1800	--	--	--	
35-39	1800	1700	1600	--	--	
40-44	1700	1600	1500	1400	--	
45-49	1600	1500	1400	1300	1200	

Source: U.S. Census

With the rates and the weights, the next step is: "Multiplication of the age-calendar year specific WY [women-years] at risk by the corresponding Connecticut incidence rates determined the number of expected breast cancers."

So the expected number of breast cancer cases would be:

0.00020	×	1900	+
0.00022	×	1800	+
0.00030	×	1800	+
0.00033	×	1700	+
0.00035	×	1600	+
0.00050	×	1700	+
0.00054	×	1600	+
0.00057	×	1500	+
0.00059	×	1400	+
etc.			

This expected number of breast cancer cases (expected if the women in the study had the same age- and calendar-year-specific breast cancer incidence as women in Connecticut) would be compared to the number of breast cancer cases actually observed.

b. It is not possible to calculate by the method used by Boice and Monson, since their method requires age-calendar-year specific incidence rates whereas the rates given in the question are not specific for calendar year.

c. The advantage of this more complex adjustment procedure is that it controls for secular changes in breast cancer incidence.

4. a. Race-sex-specific and overall TB rates for the three counties:

Incidence of tuberculosis, per 100,000, in three N.C. counties during January 1, 1986 - December 31, 1990

County	White males	White females	Nonwhite males	Nonwhite females	Overall
Johnston	7.0	4.7	124.5	32.2	18.6
Orange	2.9	1.6	8.0	9.1	3.4
Wilson	6.0	9.0	95.4	42.2	28.7

E.g., mean annual TB incidence for nonwhite females in Johnston county = 13 / (8,078 × 5) = 32.2 per 100,000. The 5 in the denominator is needed to obtain the annual incidence, since the numerator contains cases accumulated during 5 years.

Overall annual TB incidence in Johnston county =

75 / (80,664 × 5) = 93 / 5 = 18.6 per 100,000

b. SMR's:

SMRs for tuberculosis in three N.C. counties during January 1, 1986 - December 31, 1990

County	Expected	Observed / Expected	SMR
Johnston	11.7 + 6.1 + 13.5 + 8 = 39.3	75 / 39.3	1.8
Orange	12.7 + 6.8 + 14.8 + 8.7 = 43	15 / 43	0.35
Wilson	7.34 + 4 + 21 + 12.7 = 45	94 / 45	2.1

E.g., overall SMR for Johnston County:

Expected (over 5 years) based on national rates =

Group	US rate /100,000		County pop.		5 years		Expected cases in 5 yrs	
WM	0.000074	×	31,721	×	5	=	11.74	+
WF	0.000036	×	33,955	×	5	=	6.11	+
NM	0.000392	×	6,910	×	5	=	13.54	+
NF	0.000198	×	8,078	×	5	=	8.00	+
							39.39	

SMR = Observed/Expected = 75 / 39.39 ≈ 1.9

Interpretation: Both Johnston and Wilson Counties have higher TB incidence than the U.S. average. The greater TB incidence in these counties is apparently due to the higher rates in nonwhites of both sexes than in the U.S. as a whole. In Johnston County there are 56 cases in nonwhites vs. 21.5 expected; in Wilson County there are 78 cases in nonwhites vs. 33.7 expected. There is also a slight increased incidence in whites in Wilson County: 16 white cases observed vs. 11 expected. Note that the incidence of TB in Johnston County is nearly 18 times as great in nonwhite males compared to white males.

In this case comparison of the SMR's between Johnston and Orange counties is not problematic, since the race-sex population distributions (i.e., the "weights") are similar for the two counties. The population distribution in Wilson County is different, however, so comparing its SMR to the others is indeed problematic.

5. a. Intuitively, we know this assertion to be true, since:

   i. a directly standardized rate is a weighted average of stratum-specific rates;

   ii. the crude rate is a weighted average of stratum-specific rates, weighted in this case by the stratum sizes of the study population;

   iii. a weighted average of identical rates will be equal to the value of those rates, no matter what weights are used.

Using the notation from the Standardization chapter of the *Evolving Text*, with subscript "a" or "b" referring to group A or B, respectively, we have the directly-standardized rate for group A (from the formula under "Standardization of rates by the direct method" and using the information in the problem):

$$\text{Directly standardized rate for A} = \frac{\sum(r_{ai}\, N_i)}{N_t} = \frac{\sum(r_a\, N_i)}{N_t}$$

$$= \frac{r_a \sum N_I}{N_t} = \frac{r_a\, N_t}{N_t} = r_a$$

$$\text{Crude rate for A} = \frac{\sum r_{ai}\, n_{ai}}{n_t} = \frac{r_a \sum n_{ai}}{n_t} = r_a$$

So the directly standardized rate equals the crude rate (equals the stratum-specific rates). The same can be shown, in a identical manner, for B. Therefore the ratio of directly-standardized rates equals the ratio of crude rates.

Moral: if there is no variation in your stratum-specific rates, you don't need to adjust--the crude is fine.

c. This question asks about the situation in which there is a constant rate ratio between groups A and B within each age stratum. Since the SMR is calculated using the rates in the standard population (in this case, r_{bi}) for the denominator (the "expected" deaths), that denominator will be 1/K times the observed deaths, since the rates from the standard population are 1/K times the rates observed in the study population.

Using the formulas on pages 4 and 8:

$$\text{SMR} = \frac{\text{Observed deaths}}{\text{Expected deaths}} = \frac{\sum (r_{ai} n_{ai})}{\sum (r_{bi} n_{ai})} = \frac{\sum (r_{ai} n_{ai})}{\sum \left(\frac{r_{ai}}{K} \right) n_{ai}}$$

$$= \frac{\sum (r_{ai} n_{ai})}{\frac{1}{K} \sum (r_{ai} n_{ai})} = K$$

This exercise illustrates the underlying rationale for the SMR, i.e., in a situation in which there are too few data to make meaningful judgments about specific rates, we assume that each is a constant multiple of the specific rates in a standard population and then estimate that constant multiple with the SMR. The assumption of a constant multiple may not hold in reality, but it may be reasonably correct with study group we are examining. In any case it is the best we can do given the limited amount of data.

d. Intuitively, if two populations are alike in terms of a particular variable, then that variable cannot be responsible for observed differences between them.

Directly standardized rates are comparable, regardless of age distributions, because the specific rates in each population are weighted by the same external standard.

Crude rates are comparable because the crude rate for each group may be thought of as a weighted average of the group's specific rates, with weighting by the proportional size of the strata:

$$r_a = \frac{\text{deaths}}{n_{at}} = \frac{\sum (r_{ai} n_{ai})}{n_{at}} = \sum \left(r_{ai} \frac{n_{ai}}{n_{at}} \right)$$

$$r_b = \frac{\text{deaths}}{n_{bt}} = \frac{\sum (r_b n_b)}{n_{bt}} = \sum_{r_{bi}} \left(\frac{n_{bi}}{n_{bt}} \right)$$

To say that both groups have the same proportional age distribution is to say that for any age stratum (i.e., stratum "i"),

$$r_b = \frac{n_{ai}}{n_a} = \frac{n_i}{n_{bt}} = p_i$$

So $r_a=\Sigma[r_{ai}p_i]$, $r_b=\Sigma[r_{bi}p_i]$, and the two sets of specific rates are averaged using the same weights, p_i.

Indirectly standardized rates:

From the formula at the top of page 4,

$$\text{Indirectly standardized rate} = r_t \times \frac{R_t}{\sum (R_i n_i)/n_t} = \frac{r_t R_t}{\sum \left(R_i \frac{n_I}{n_t} \right)}$$

$$= r_t \frac{R_t}{\sum (R_i p_i)}$$

Since R_t and R_i come from the standard population and p_i is the same for groups A and B (though it may vary from stratum to stratum) by the conditions of the problem, the indirectly standardized rates for A and B are each equal to their crude rates times a the same constant. So a comparison of indirectly standardized rates in this case is the same as a comparison of their crude rates, which was shown above to be valid.

6. An Excel® spreadsheet for this problem can be found on the EPID 168 web site at www.sph.unc.edu\courses\epid168\public\Standardization.xls.

7. Relating risk factors to health outcomes

Quantifying relationships between two factors or one factor and the occurrence, presence, severity, or course of disease

The "Big Picture"

At this point in the course, it will be good to take stock of where we are and where we are going. After a brief overview of population and health, we have thoughtfully considered the phenomenon of disease in relation to how epidemiologists study disease. Under that topic we examined issues of definition, classification, and natural history. We then turned to the question of how to measure disease frequency and extent in populations. We examined some general issues in numeracy and descriptive statistics, and then took up the fundamental epidemiologic measures of prevalence and incidence, with the latter approached as a proportion or as a rate. From there we took up the topic of standardization, which facilitates comparisons between prevalence and incidence across populations with different demographic composition, and we saw how these various measures and concepts are used in descriptive epidemiology and surveillance.

For the next section of the course we will be concerned with how to investigate associations between health outcomes and potential risk factors. That task involves questions of study design, measures of association, validity, inference and interpretation. The topics of study design and measures of association are so intertwined that whichever one we begin with, it always seems that we should have begun with the other! Analytic studies provide the data for estimating measures of association and impact, but measures of association and impact motivate the design of the studies.

However, the basic epidemiologic approach to relating risk factors to health outcomes is more general than the specifics of either topic. Consider a population in which a disease or some other condition occurs throughout the population but more often in persons with characteristic A. We are likely to be interested in how the existence (prevalence) or occurrence (incidence) of the disease among people with characteristic A compares with that for the population as a whole and for people with some other characteristic B (which could simply be the absence of A). To make this comparison we:

a. Measure the frequency - prevalence, CI, ID - of the disease or condition in each group (and perhaps in the entire population);
b. Compare the frequencies (fairly! - e.g., after standardization if necessary))
c. Quantify the comparison with a measure of association
d. Quantify the potential impact of the characteristic on the condition, if we are willing to posit a causal relationship.

We have already discussed measures of frequency and extent. Now we turn to measures of association and impact.

Measuring the strength of a relationship

The question that summarized the preceding topic could be stated as "How much of a factor is there?" or "How often does a disease (or other phenomenon) occur?". However, much of epidemiology is concerned with relationships among factors, particularly with the effect of an "exposure" on "a disease". Therefore the present topic addresses the question "How strong is the relationship between two factors?" or "How strong is the relationship between a study factor and an outcome?" A relationship may be "strong" without being "causal", and vice versa. Nevertheless, two factors that are strongly associated are more likely to be causally related.

There are a number of ways in which the strength of the relationship between two variables can be assessed. We can, for example, assess the extent to which a change in one variable is accompanied by a change in the other variable or, equivalently, the extent to which the distribution of one variable differs according to the value of the other variable. For this assessment, epidemiologists use a measure of association*.

A second perspective is the extent to which the level of one of the factors might account for the value of the second factor, as in the question of how much of a disease is attributable to a factor that influences its occurrence. Epidemiologists use measures of impact to address this question.

Most of the measures we will cover in this topic apply to relationships between a factor that is dichotomous (binary, having two possible values) and a measure of frequency or extent, in particular, a rate, risk, or odds. Such measures are the most commonly used in epidemiology. We will also touch on measures that are used in other situations.

Measures of association

A measure of association provides an index of how strongly two factors under study vary in concert. The more tightly they are so linked, the more evidence that they are causally related to each other (though not necessarily that one causes the other, since they might both be caused by a third factor).

Association - two factors are associated when the distribution of one is different for some value of the other. To say that two factors are associated means, essentially, that knowing the value of one variable implies a different distribution of the other. Consider the following two (hypothetical) tables:

* Although this term and "measure of effect" have frequently been used interchangeably (e.g., in this text), Rothman and Greenland (2000:58-59) draw the following distinction: associations involve comparisons between groups or populations; effects involve comparisons of the same population [hypothetically] observed in two different conditions; measures of association are typically used to estimate measures of effect.

CHD and oral contraceptives in women age 35 years or more

	OC	No OC	Total
CHD	30	20	50
Non-case	30	70	100
Total	60	90	150

Breast cancer (BC) and oral contraceptives

	OC	No OC	Total
Cancer	15	35	50
Non-case	30	70	100
Total	45	105	150

Consider first the table on the left (CHD and OC). The overall proportion of OC users is 60/150 = 0.40, but that among CHD cases is 30/50 = 0.60 while that among noncases is 30/100 = 0.30. The distribution of values of OC use ("users", "nonusers") is therefore different for different CHD values ("case", "noncase"). Similarly, the distribution of values of CHD is different for different values of OC (30/60 of OC users have CHD; 20/90 of non-users of OC have CHD).

If asked to estimate the proportion of OC users in a sample of 40 women selected at random from the table on the left, would we want to know how many in the sample had CHD and how many did not? Indeed we would.

We know that the proportion of OC users must be no lower than 0.30 (if the sample consists entirely of noncases) and no greater than 0.60 (if the sample consists entirely of cases). In the absence of knowing the proportion of cases, our best estimate would be the overall proportion in the population, 0.40. But is we knew the proportion of cases in the sample, we could move our estimate up (if more than one-third were cases) or down (if fewer than one-third were cases). [Now, verify that to estimate the proportion of CHD cases in a random sample, we would want to know the proportion of OC users. What is the best estimate of the proportion with CHD if the sample consists of 22 OC users and 18 nonusers? – The answer is at the end of this chapter.]

Thus, in the data in the left-hand table, there is an association between OC use and CHD. In contrast, in the table on the right (BC and OC), the distribution of OC use is the same for the cases, the noncases, and the entire group. Therefore, the data in the right-hand table show no association between breast cancer and use of OC's.

Correlation and Agreement

Association is a general term that encompasses many types of relationships. Other terms are used to indicate specific types of association. Two important ones are:

Correlation is a type of association in which the relationship is monotonic, i.e., it goes in one direction - the more of one factor, the more of the other (positive or direct correlation), OR the more of one factor, the less of the other (negative or inverse correlation). Linear correlation (measured by the Pearson product-moment correlation coefficient) assesses the extent to which the relationship can be summarized by a straight line. Nonparametric correlation coefficients, such as the Spearman rank correlation coefficient, assess the extent to which the two factors are correlated

but without regard to the size of the change in one that accompanies a change in the other, simply the direction.

Agreement is a type of correlation in which the two factors (generally two measures of the same phenomenon) are not only directly correlated with each other but have the same actual values. For example, two sphygmomanometers should give the same readings when used on the same person on the same occasion, not merely readings that are correlated. Two measurements of a stable phenomenon should agree with each other, not merely correlate. If one of the measures is known to be highly accurate and the other is being assessed, then we can assess validity of the latter, rather than merely agreement between the two.

ASIDE

Some sociological commentary

Since the factors studied by epidemiologists are often the occurrence of disease and the presence of exposure, the primary epidemiologic measures are proportions and rates of disease across different exposure groups. Indeed, because these measures are so familiar to epidemiologists and clinicians, even when the disease (e.g., blood pressure) and/or exposure are not represented by dichotomous (two-category) variables, it is common convert them into proportions or rates for at least some analyses. We will therefore spend most of our time on measures of association and impact involving rates and proportions. Bear in mind, though, that phenomena (e.g., physiologic measurements, nutrient intake, environmental exposures) that are capable of being measured as quantities are often more properly analyzed without dichotomizing.

The preference for rates and proportions is one reason for the different approaches to statistical analysis used by epidemiologists and social scientists who also study data on populations. But there are other differences in approach that presumably have a different basis, perhaps epidemiologists' focus on biological relationships.

One potential source of confusion – even conflict! – is the difference in the way that epidemiologists on the one hand and social scientists and biostatisticians look at associations. Epidemiologists tend to regard the strength of an association as a separate matter from the quantity of numerical evidence that the association would not easily arise by chance (i.e., its "statistical significance"). Other professions, however, often look first to the statistical significance of an association before considering any other characteristic. Thus, a biostatistician or psychologist might complete dismiss an association that an epidemiologist might characterize as "strong though potentially due to chance". Conversely, a psychologist or biostatistician may characterize as "highly significant" an association that an epidemiologist might dismiss as too weak to be biologically meaningful. As we will see later, various measures of association used in statistics (e.g., chi-squared statistics, correlation coefficients) are in a different category than the measures of association we will discuss now.

END OF ASIDE

Some basic measures

Before diving in to our discussion of how to measure associations, we may wish to begin with some basics. Suppose that an epidemiologist is asked to investigate the possible hazard from an inadequate air filtration system in a large school building in a poor urban neighborhood. The particular concern involves children with asthma, 400 of whom attend the school (school A). The epidemiologist is informed that on a particular day, 12 children suffered an asthmatic attack, whereas at a very similar nearby school (school B) with 500 asthmatic children, only 5 suffered an asthmatic attack on the same day.

The epidemiologist first arranges the data in a 2 x 2 table:

Cumulative incidence of asthmatic attack during one school day

Had an asthma attack	School A	School B	Total
Yes	12	5	17
No	388	495	883
Total	400	500	900

The first step is to compute the incidence in each school:

1-day cumulative incidence in school A: 12 cases / 400 children at risk = 0.03 or 3%

1-day cumulative incidence in school B: 5 cases / 500 children at risk = 0.01 or 1%

School A does in fact have a higher incidence of asthma attacks on the study day.

In order to assess the strength of the association between school and asthma incidence, the next step is to compute a measure of strength of association. The most common measure computed in this situation is the ratio of the two cumulative incidences (the "cumulative incidence ratio", CIR, also called the "risk ratio"). The CIR is simply 0.03/0.01 = 3.0, which is often interpreted as indicating a "moderately strong" association. The epidemiologist cumulative incidence difference (CID) might also compute the difference between the CI's (a "cumulative incidence difference", CID), and report that having inadequate air filtration was associated with a two percentage point greater asthma incidence during the 7-hour school day. Armed with this basic example, let us examine the concepts that underlie these measures.

Absolute versus relative effects

When we have incidence rates or proportions from two different populations (e.g., PC-users and Mac-users), it is easy to tell which rate is larger. But quantifying how much larger raises the question of how to compare the two rates. A basic question is whether or not the amount by which the larger rate exceeds the smaller one should be relative to the size of one of the rates.

If you ask a 10-year old how much older she is than her 5-year old brother, she will probably answer "5 years". But if she is mathematically-inclined, she may say that she is "twice his age" or "100% older". Both statements accurately quantify the amount by which she is older, yet they have different "flavors". Do we have a reason to prefer one or the other?

We might be inclined to prefer the answer "5 years". "Might", because the choice of a measure depends on our purpose, and we have not specified an objective. But two reasons come to mind why we might prefer the absolute difference (5 years) to the relative difference (100% older) or ratio (twice his age).

For one, "5 years" will remain accurate indefinitely, whereas "twice" (or "100% more") are accurate only this year. In that sense "5 years" provides a better summary of the relation between the children's respective ages. For another, human growth and aging, at least from a societal point of view and perhaps from a biological point of view as well, are processes which are marked by absolute increases, not relative ones. For example, we generally think of school entrance and graduation, puberty, eligibility for a drivers' license, presbyopia, and retirement in terms of specific age ranges, not proportional increases. We say "in 15 years you will probably need bifocals", rather than "when your age is 50% greater". In contrast, when adjusting a recipe for a larger or smaller number of guests, we multiply or divide the amounts of each ingredient by a common factor, rather than subtract a common amount from each one. For scaling a recipe, we are interested in proportionate (relative) increases.

Similarly, when we quantify the comparison of two incidences (or two prevalences), we can take the absolute difference (incidence difference) or the relative difference (excess risk). Which one, absolute or relative, is of greater interest to us in quantifying the comparison of two measures of occurrence or extent? This question has inspired no small amount of debate in the early days of modern epidemiology (ca. 1955) and, as so often happens, a case can be made for both approaches. The choice depends on our objective, our concept of the phenomena, and the availability of data.

One problem with using the absolute difference (variously called "risk difference", "rate difference", "cumulative incidence difference", "incidence density difference", "attributable risk", according to fashion, the group of epidemiologists with which the epidemiologist wishes to identify him/herself, the decade in which she/he learned epidemiology, or whether the comparison involves incidence rates, incidence proportions, prevalences, or mortality rates) as a measure of strength of association is that if the incidences themselves are small, as will always be the case for a rare disease, then the difference must also be small. For example, if the annual mortality rate for a rare disease such as esophageal cancer is 60/100,000 in persons with low vitamin C intake and 20/100,000 in persons with high vitamin C intake, the difference is only 40/100,000. In contrast, the difference for an association involving a more common disease, such as vitamin E and CHD, might be 1,200/100,000 for low vitamin E intake and 800/100,000 for high vitamin E intake = 400/100,000, an order of magnitude greater.

The much greater size of the second difference indicates that if these two vitamins are causal factors many more lives could be saved from increasing vitamin E intake than from increasing vitamin C intake. Vitamin E appears to have a greater public health impact. But is it logical to conclude from the greater difference for vitamin E that its association with CHD is stronger than vitamin C's with

esophageal cancer? First, if we did draw that conclusion it would imply that nearly any association involving a common disease must be stronger than all associations involving very rare diseases. Second, since the actual incidence of most conditions varies by all sorts of factors (age, gender, economic resources, smoking, alcohol intake, physical activity, diet, genetics, cofactors), the absolute difference is very likely to vary, possibly greatly, across populations (however, the relative difference may also vary).

In contrast, expressing the incidence differences relative to the size of the actual incidences produces measures of association that appear to be comparable. Thus we can compute a relative difference in incidence of esophageal cancer mortality in relation to vitamin C as $(I_1 - I_0)/I_0$ = (0.00060 – 0.00020)/0.00020 = 2.0 and a relative difference for CHD mortality in relation to vitamin E as $(I_1 - I_0)/I_0$ = (0.01200 – 0.00800 / 0.00800) = 0.50. On this basis, the association involving vitamin C is substantially greater than that involving vitamin E. This relative difference measure is often called the excess risk (or "excess rate", since the data are rates, not proportions). If we add 1.0 to the excess risk or rate, we obtain an even simpler relative measure, I_1/I_0, which is variously termed relative risk, risk ratio, rate ratio, cumulative incidence ratio, incidence density ratio, or, for prevalences, prevalence ratio.

Here are two real-life examples that contrast relative and absolute measures of association. The first is based on data from a follow-up study by Mann *et al.* (presented in a seminar at UNC-CH by Bruce Stadel):

Relative versus Absolute Measures of Association Incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years

Age (years)	Cigarettes/day	Oral contraceptive users	Non-users	RR**	AR***
30-39	0-14	6	2	3	4
	15 +	30	11	3	19
40-44	0-14	47	12	4	35
	15 +	246	61	4	185

Notes:

* RR=relative risk (rate ratio)

** AR=attributable risk (rate difference, absolute difference)

In this table, the incidence of MI is clearly greater for OC users, since in each age-smoking stratum the OC users have a higher incidence (ID) than do the nonusers. Moreover, the ratio of the two incidences (the RR) is nearly constant across strata, a desirable property for a summary measure, whereas the rate difference (AR) varies widely. According to Breslow and Day, the rate ratio tends to be more stable across strata, supporting its desirability as a measure of association. Not all quantitative epidemiologists agree with this assertion.

The second example comes from a follow-up study of lung cancer and coronary artery disease in relation to cigarette smoking:

Mortality rates per 100,000 person-years from lung cancer and coronary artery disease for smokers and nonsmokers of cigarettes

	Smokers	Nonsmokers	Ratio	Difference
Cancer of the lung	48.3	4.5	10.8	44
Coronary artery disease	294.7	169.5	1.7	125

Source: 1964 *Surgeon General's Report on Smoking and Health*, page 110, quoted in Joseph Fleiss, *Statistical methods for rates and proportions*, 2nd edition, page 91

The rate ratio for the relation between smoking and lung cancer mortality is much larger than that between smoking and coronary artery disease mortality, but the rate difference is much larger for coronary artery disease mortality. These figures are usually interpreted to mean that lung cancer mortality is more closely associated with cigarette smoking than is coronary artery disease mortality; elimination of cigarette smoking would lead to a proportionate reduction in lung cancer mortality greater than the proportionate reduction in coronary artery disease mortality. However, the reduction in the number of deaths from lung cancer would be smaller in magnitude than the reduction in deaths from coronary artery disease. These issues will be explored in detail in the section Measures of Impact, later in this chapter.

Concept of relative risk

Nevertheless, for the most part we use relative risk as the basic measure of strength of association between a characteristic and the development of a condition.

General approach - compare observed to "expected".

Concept of relative risk is operationalized by :

a. Cumulative incidence ratio (CIR), also called risk ratio
b. Incidence density ratio (IDR), also called rate ratio
c. Odds ratio (OR), which estimates CIR and IDR under certain circumstances.

General formula:

$$\text{Incidence ratio} = \frac{\text{Incidence in "exposed"}}{\text{Incidence in "unexposed"}} = \frac{I_1}{I_0}$$

You may recall from the chapter on standardization that the SMR can be thought of as a ratio of "observed" to "expected" mortality rates. In fact, the concept of observed and expected can be

brought in here as well. When we contrast the incidence rates in exposed and unexposed groups, we are typically using the unexposed incidence as a barometer of what incidence we might find in the exposed group if exposure had no effect. In that sense, the incidence in the unexposed constitutes an "expected", while the incidence in the exposed group constitutes an "observed".

The concept of relative risk can also be applied in situations where incidence estimates are unavailable or not even of greatest interest. For example, as explained below an estimate of the ratio of incidence in exposed persons to incidence in unexposed persons can be obtained in a case-control study with incident (newly-occurring) cases, as indicated below (see the section on the Odds Ratio) and explained in the chapter on Analytic Study Designs).

In situations where we want to estimate incidences or incidence ratios but only prevalence data are available, the prevalence ratio or prevalence odds ratio may estimate the incidence ratio. The basis for doing so is the relation among prevalence, incidence, and duration presented in the chapter on Measuring Disease and Exposure (in a stationary population, prevalence odds = incidence × average duration, or for a rare outcome, prevalence ≈ incidence × average duration). A key question is whether duration is the same in all groups being compared, since if it is not then the comparison of prevalences will provide a distorted picture of a comparison of incidences.

If the research objective is the comparison of incidence rates or proportions and the available data are cross-sectional, then which prevalence measure of association to use – the prevalence ratio (PR) or the prevalence odds ratio (POR) – has been a matter of some controversy. One recommendation has been to use the PR when the objective of interest is the CIR and to use the POR when the objective of interest is the IDR (always assuming that duration is unrelated to exposure and that the incidences and durations in exposed and unexposed groups have been constant long enough to maintain a steady state condition). In practice the POR has become much more popular, partly because of the greater availability of statistical software packages for fitting logistic regression models (whose coefficients can be converted to estimates of odds ratios) than for fitting regression models whose coefficients can be converted to estimates of risk or prevalence ratios. However, as software tools for estimating PR have become available, the methodological literature increasingly favors it PR as ensuring "truth in advertising" (p273), being more interpretable, and providing a more nearly consistent approximation to the incidence ratio (Thompson, Myers, and Kriebel, 1997).

The PR is also a logical choice for quantifying associations between exposures and conditions whose duration is as or more important than their incidence. For example, a large proportion of a population experience emotions or conditions such as anxiety, fatigue, or unhappiness from time to time. Since point prevalence will count mostly people in whom the condition persists, prevalence may be as or more useful than incidence as a measure of frequency in such cases. (The PR is also the straightforward choice for simple descriptive statements, such as "smoking was twice as common among persons with less than a high school education".)

Interpretation of relative risk

Example: Incidence ratio of 2.0 means that:

- "The incidence in the exposed population is twice that in the unexposed population"

- "The exposure is associated with a 100% increase in incidence."
- "The exposure is associated with a two-fold greater incidence."

Descriptive adjectives for magnitude of association (as commonly used)

1.0	No association (null value)
1.1-1.3	Weak
1.4-1.7	Modest
1.8-3.0	Moderate
3-8	Strong

For inverse associations (incidence ratio is less than 1.0), take the reciprocal and look in above table, e.g., reciprocal of 0.5 is 2.0, which corresponds to a "moderate" association.

Two-by-two tables

The most basic data layout in epidemiology is the two-by-two table:

	Exposure			
Disease	Yes	No	Total	
Yes	a	b	m_1	(a + b)
No	c	d	m_2	(c + d)
Total	n_1 (a + c)	n_0 (b + d)	n	

One major epidemiologic controversy is whether the disease should be shown in the rows, as above, or in the columns. Kleinbaum, Kupper, and Morgenstern use the above format. Hennekens and Buring place the disease categories in the columns and the exposure in the rows. Some authors use one presentation for cohort studies and the other for case-control studies. As you can see, epidemiology is not yet really a discipline (or not yet disciplined).

The above form of the 2 x 2 table is used to present data from a study (e.g., cohort, cross-sectional, case-control) with count data. When the study uses person-years data (e.g., to estimate incidence density), then the "no disease" column is removed and person-time totals (PY_1, PY_0) occupy the right-hand marginal:

	Exposure		
Disease	Yes	No	Total
Yes	a	b	m_1
Person-time	PY_1	PY_0	PY

Armed with our tables (whatever their orientation), we will now define the three major relative risk measures, about which there is much less controversy:

1. Cumulative incidence ratio (CIR)
2. Incidence density ratio (IDR)
3. Odds ratio (OR)

Cumulative incidence ratio (also called "risk ratio" or "relative risk")

The cumulative incidence ratio (CIR) addresses the question "by how many times does the risk in exposed persons exceed that for unexposed persons?" If the CIR is 3, we can say that exposed persons have 3 times the risk of unexposed persons. We can also say that the average exposed individual has three times the risk of disease as the average unexposed individual. This is often just what we want to know. The mathematical definition is:

$$\text{Cumulative incidence ratio} = \frac{\text{Cumulative incidence in "exposed"}}{\text{Cumulative incidence in "unexposed"}} = \frac{CI_1}{CI_0}$$

Since the CIR is based on estimates of CI or risk, the CIR can be estimated directly only from a cohort study. It is, however, possible to estimate it indirectly in other situations.

Incidence density ratio (also called "rate ratio")

The incidence density ratio (IDR) addresses the question "how many times does the rate of disease in exposed persons exceed that in unexposed persons?". If the IDR is 3 we can say the the rate in the exposed is 3 times that in the unexposed. There is not an obvious interpretation at the individual level, but the IDR is of prime importance for studies of dynamic populations and lengthy cohorts. The mathematical definition is:

$$\text{Incidence density ratio} = \frac{\text{Incidence density in "exposed"}}{\text{Incidence density in "unexposed"}} = \frac{ID_1}{ID_0}$$

The IDR is used in situations where the outcome is the length of time until an event (e.g., death) occurs and is mathematically equivalent to the hazard ratio of survivorship analysis. The IDR can be estimated directly in a follow-up study (of a fixed cohort or a dynamic population).

(Risk) odds ratio

The odds ratio (OR) is a ratio of "odds", which are transformations of risks or probabilities.

odds = p/(1-p), where p = probability

The OR addresses the question "how many times greater is the odds of disease for exposed persons than for unexposed persons?" Since odds have a different scale of measurement than risk, the answer to this question can sometimes differ from the answer to the corresponding question about risk. Often, however, we are concerned with rare diseases, for which risk and odds are very close and CIR's and OR's (and IDR's) are very close. Since the OR can be defined in terms of odds of disease among exposed or odds of exposure among cases, there are two mathematical formulations:

$$\text{Odds ratio} = \frac{\text{Odds in "exposed"}}{\text{Odds in "unexposed"}}$$

The odds is simply an algebraic transformation of probability, so any probability (which must, of course, be less than 1.0) can be expressed as "odds". The probability that something may happen, especially something bad, is often referred to as a "risk". Odds derived from a risk are termed, appropriately, risk odds, so that a ratio of two risk odds is a **risk odds ratio**, or **ROR**.

(Exposure) odds ratio

A prevalence is commonly referred to as an estimate of probability (e.g., of exposure). A justification for this usage is that if we were to select an individual at random from the group, the probability that that individual would have a certain characteristic is estimated by the prevalence in the group. Odds that correspond to the probability of exposure are called "exposure odds", so their ratio is an **exposure odds ratio**, or **EOR**. Although conceptually distinct, for a two-by-two table these two odds ratios are algebraically identical, as we shall see. Thus, our ability to estimate an (exposure) odds ratio in a situation where we do not know disease incidence is a powerful tool for examining *associations* involving disease incidence even where we do not have incidence data, as was first presented in a classic paper by Jerome Cornfield (see the chapter on Analytic Study Designs for elaboration).

$$OR_r = \text{Risk odds ratio} = \frac{\text{Risk odds in "exposed"}}{\text{Risk odds in "unexposed"}} = \frac{odds_1}{odds_0} = \frac{CI_1 / (1\text{-}CI_1)}{CI_0 / (1\text{-}CI_0)}$$

$$OR_e = \text{Exposure odds ratio} = \frac{\text{Exposure odds in "cases"}}{\text{Exposure in "noncases"}} = \frac{odds_1}{odds_0}$$

Relation of the odds ratio to the risk ratio

When incidences are small (i.e., the outcome under study is rare in the population), the odds ratio closely approximates both the risk ratio and the incidence density ratio. The conventional guideline for classifying a disease as "rare" is an incidence below 10%. A good way to assess the extent of divergence of the odds ratio and risk ratio is to examine a spreadsheet with sample incidences and computed relative risks and odds ratios (e.g., the guideline suggested by Zhang and Yu [1998] of incidence below 10% and risk ratio below 2.5 allows the odds ratio to be only 20% greater than the risk ratio).

If one feels that the OR exaggerates the strength of association objectionably, it is a simple matter to derive a corresponding risk ratio estimate if one has additional information – overall exposure prevalence, overall disease incidence, disease incidence in the exposed, or disease incidence in the unexposed (Hogue, Gaylor, and Schulz, 1983). The simplest conversion is available if one knows the incidence in the unexposed group, e.g.:

$$RR = \frac{OR}{(1 - CI_0) + (CI_0 \times OR)}$$

where CI_0 is the incidence in the unexposed group [Zhang and Yu (1998), adapted to the notation used here]. A prevalence odds ratio can be converted into a prevalence ratio using the same formula and substituting the prevalence in the exposed in place of CI_0. The divergence between the OR and the IDR will generally be less than that between the OR and the CIR. The reason is that all three measures of incidence (ID, CI, odds) have the identical numerator (new cases), but as incidence increases the denominators of ID and odds decrease, whereas the denominator for CI does not change.

The odds ratio has been somewhat of an unloved stepchild in the epidemiology and medical professions. Authors have termed it "incomprehensible" (Lee, 1994:201), stating that it "lacks intelligibility" (Lee and Chia, 1994). Many beginning students will no doubt find themselves in sympathy with this view! Since the divergence between the OR and risk ratio (and between POR and prevalence ratio) is a popular topic with examiners, students of epidemiology should remain alert to the distinction. If you are caught interpreting an OR as a relative risk for a non-rare outcome, however, you can limit the damage with the argument that (since the divergence between the OR and risk ratio is great only for strong associations) "qualitative judgments based on interpretating odds ratios as though they were relative risks are unlikely to be seriously in error" (Davies, Crombie, and Tavakoli, 1998:991).

Two typically unstated assumptions

Stable exposure status

The above discussion assumes that the population being studied is reasonably stable in respect to exposure status. When this is not the case it may be necessary to change individuals' exposure status during the observation period, assigning their follow-up to one or another exposure group, if the

exposure effect is believed not to persist. For example, a subject may exercise, stop, and begin again. If the effect of exercise is believed to terminate shortly after exercise is stopped and to begin again shortly after resumption of exercise, then follow-up time (person-time) can be accumulated in the appropriate exercise category for each part of the follow-up period of an incidence density measure. (An alternative approach is to place such "switchers" in a category of their own.)

Absence of "contagion"

The above discussion also assumes that exposure and outcome are independent, i.e., one person's disease does not affect another person's risk. This assumption is violated, of course, for contagious diseases, such as sexually transmitted infections, and for anthropod-borne pathogens, e.g. malaria, where humans serve as a reservoir. Here, the spread of disease increases the exposure of unaffected individuals so that their risk increases. These so-called "dependent happenings" can result in distortion, or at least marked variability over time, in the above measures of association (see, for example, Koopman JS *et al.*, 1991). Dependent happenings are by no means confined to communicable diseases, inasmuch as personal and community behaviors are frequently affected by what other people and communities are doing. Some examples are smoking cessation, dietary change, suicide attempts, driving behavior, road safety regulations, and intensity of disease detection and reporting.

Some theoretical background (optional)

Although the risk odds ratio is a measure of association, Greenland (1987) asserts that its relevance for epidemiology derives solely from its ability to estimate of the rate ratio (IDR) or cumulative incidence ratio (CIR). His objection to the odds ratio as a measure of effect lies in the lack of a simple correspondence between the odds for a population and the odds for an individual. Whereas "incidence proportion" (i.e., CI) is equivalent to a simple average of the risk for each individual in the population and incidence density (ID) is equivalent to a simple average of the "hazard" for each individual in the population, incidence odds is not equivalent to a simple average of the disease odds for each individual in the population (Greenland, 1987). Thus, the OR is not a ratio of averages interpretable at the individual level.

However, the OR also has some desirable properties compared to the risk ratio (Walter, 2000). For example, in a smoking cessation trial, the OR for success will be the reciprocal of the odds ratio for failure; the "risk" ratio (CIR) for success, however, will be very different from the CIR for failure.

The popularity of the OR derives from the wide use of multiple logistic regression, which provides estimates for the OR controlling for multiple variables (see chapter on Data Analysis and Interpretation) and, of course, from our ability to estimate the (exposure) OR (OR_e) from a case-control data. Whether the OR in a case-control study estimates the IDR or the CIR depends upon the study design, particularly in relation to the sampling of controls (see Greenland 1987). Even if the disease is not rare, if it has an extended risk period (so that incidence density would generally be preferred) and the controls are obtained through "density sampling" (selected at the same time as matched cases), then the OR from

the case-control study estimates the IDR (see chapter on Analytic Study Designs; ambitious students are referred to Sander Greenland and Duncan C. Thomas, "On the need for the rare disease assumption in case-control studies", *Am J Epidemiol* 1982; 116:547-53 and references therein). The great simplification is that when the CI for the time period of observation is less than 10% in all categories, all three measures – the OR, CIR, and IDR – will be essentially equivalent.

More on risk and relative risk

The **excess risk** gives the proportionate increase in incidence (an analogous measure can be constructed using incidence density or odds). It is a slight modification of the CIR and useful in a variety of circumstances including measures of relative impact, to be discussed shortly. The algebraic definition is:

$$\text{Excess risk} = \text{CIR} - 1 = \frac{CI_1}{CI_0} - 1 = \frac{CI_1 - CI_0}{CI_0}$$

For diseases with an extended risk period, as duration of follow-up increases risk and CI become larger. Being cumulative and predicated on the population remaining at risk, CI is an increasing function whose limit is 1.0 – if we remain at risk forever, then eventually we will all become cases. As CI_1 and CI_0 both increase towards their limit of 1.0, then the CIR also approaches 1.0. Therefore the value of the CIR can change as the duration of follow-up lengthens. It is also possible for the IDR to change with duration of follow-up, but that is a function of the natural history of the disease rather than the the IDR's mathematical properties.

When the CI is low, due to a rare disease and/or short follow-up period:

$$CI \approx ID \times T \quad \text{(where T = follow-up time)}$$

$$OR \approx IDR \approx CIR$$

because if CI is ≈ ID x T, then $CI_1 = ID_1 \times T$ and $ID_0 = ID_0 \times T$, so:

$$CIR \approx \frac{ID_1 \times T}{ID_0 \times T} = \frac{ID_1}{ID_0} = IDR$$

As follow-up time becomes shorter, then CI becomes smaller, eventually reaching 0. But as the CI becomes smaller its value becomes increasingly the same as ID x T. For this reason, the limit of the CIR as the follow-time becomes vanishingly short (T -> 0) is the IDR. For this reason the IDR is sometimes referred to as the "instantaneous CIR".

In a steady-state (constant size and age distribution, constant incidence density, prevalence, and duration of disease) dynamic population:

Prevalence odds = Incidence x Duration (see previous chapter)

From this we can see that the prevalence odds ratio (POR) estimates the IDR if duration is unrelated to exposure - because:

$$POR = \frac{odds_1}{odds_0} = \frac{ID_1 \times T}{ID_0 \times T} = \frac{ID_1}{ID_0} = IDR$$

where T here is duration in exposed and unexposed cases.

Estimating relative risk (via the odds ratio) from data from a case-control study

1. Construct (2x2, four-fold) table

	Exposure			
Disease	Yes	No	Total	
Yes	a	b	m_1	(a + b)
No	c	d	m_2	(c + d)
Total	n_1 (a + c)	n_0 (b + d)	n	

2. Odds of Exposure in cases

$$Odds = \frac{\text{Proportion of cases who are exposed}}{\text{Proportion of cases who are unexposed}} = \frac{a / (a + b)}{b / (a + b)} = \frac{a}{b}$$

3. Odds of exposure in controls

$$Odds = \frac{\text{Proportion of controls who are exposed}}{\text{Proportion of controls who are unexposed}} = \frac{c / (c + d)}{d / (c + d)} = \frac{c}{d}$$

4. Exposure odds ratio (OR_e)

$$OR_e = \frac{\text{Odds of exposure in cases}}{\text{Odds of exposure in controls}} = \frac{a / b}{c / d} = \frac{ad}{bc}$$

If the data had come from a cross-sectional or cohort study, we could instead have estimated the risk odds ratio (OR_r), as the odds of disease in exposed persons divided by odds of disease in unexposed persons. Algebraically, the exposure and disease odds ratios are identical.

Note that the odds ratio can be computed from proportions or percentages as readily as from the actual numbers, since in computing the odds ratio the first step (see above) is to convert the numbers into proportions and then to convert the proportions into odds.

Difference measures

Measures based on the difference between two proportions or rates are the other principal form of comparison for rates and proportions. They are often used as measures of impact, as we will discuss in the next section. The formulas and terms for differences of cumulative incidences (or risks) and incidence rates are:

$CID = CI_1 - CI_0$ ("**Cumulative incidence difference**",

also known as the "Risk difference" or "Attributable risk")

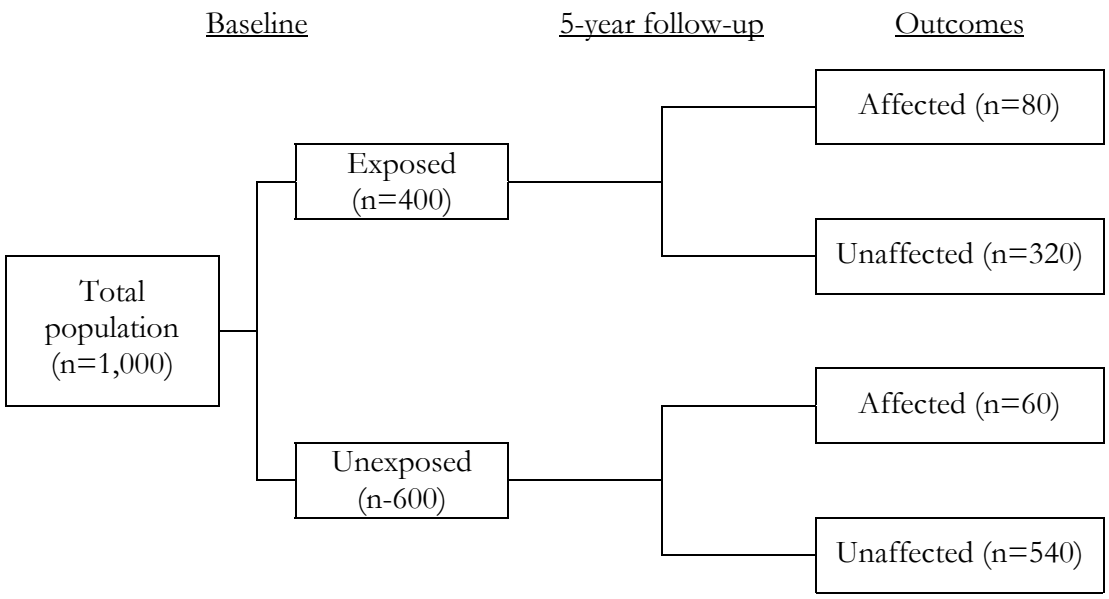
$IDD = ID_1 - ID_0$ ("**Incidence density difference**",

also known as the "Rate difference")

These difference measures, of course, can be derived directly only from a cohort or follow-up study. If we lack information on the size of the population at risk, as in a case-control study with no additional information, we have no way to estimate either CI or ID, so we cannot estimate risk or rate differences. In a cross-sectional study, we cannot estimate incidence at all, though by analogy with CID and IDD we can estimate the **prevalence difference**, $P_1 - P_0$.

Examples of computations

Follow-up of a fixed cohort



(Assume no losses to follow-up, including deaths from other causes.)

The above data are often summarized into a 2 x 2 table:

5 - year incidence of disease

	Exposure		
Disease	Yes	No	Total
Yes	80	60	140
No	320	540	860
Total	400	600	1000

Note on terminology: The four numbers (80, 60, 320, 540) in the interior of the table are referred to as the "cells"; the row and column totals (140, 860, 400, 600) are referred to as the "marginals". The cells are often referred to as "a", "b", "c", and "d" in zig-zag fashion beginning with the upper left cell.

CI (crude) = 140 / 1000 = .14 (i.e., the overall 5-year cumulative incidence was 14/100)

CI_1 = 80 / 400 = .20, CI_0 = 60 / 600 = .10

CIR = CI_1 / CI_0 = .20 / .10 = 2.0 (the exposure was associated with a doubling of risk)

CID = CI_1 - CI_0 = .20 - .10 = .10 (see below for interpretation)

Excess risk = CIR - 1 = 1.0 (i.e., the exposure was associated with a 100% increase in risk)

$$OR_r = \frac{CI_1 / (1 - CI_1)}{CI_0 / (1 - CI_0)} = \frac{0.20 / 0.80}{0.10 / 0.90} = \frac{0.25}{0.11} = 2.25$$

Note that the OR is more extreme than the CIR.

Average incidence density measures could be computed from the above table by making the assumption that cases occurred evenly throughout the period, or equivalently, that all cases occurred at the midpoint of the follow-up period, 2.5 years:

$$ID = \frac{\text{cases}}{\text{person-years}} = \frac{140}{(860)(5) + (140)(2.5)} = \frac{140}{4300 + 350} = 0.030 \text{ cases/py}$$

(Total person-years at risk comprises 860 persons followed for 5 years and 140 persons followed for 2.5 years – once the disease occurred, that subject was deemed no longer at risk. If that situation were not the case, then person-years would be computed differently.)

$ID_1 = 80 / [(320)(5) + (80)(2.5)] = 80 / 1800 = 0.044$ cases / person-year

$ID_0 = 60 / [(540)(5) + (60)(2.5)] = 60 / 2850 = 0.021$ cases / person-year

IDR = 0.044 / 0.021 = 2.095 = 2.1 (compare to CIR of 2.0)

IDD = 0.044 - 0.021 = .023 cases / person-yr OR 23 cases / 1000 person-yrs

Note that each ID is very close to the corresponding CI divided by the number of years (5). When the incidence is low, the CI approximately equals ID x (time interval).

Measures of association – non-dichotomous exposure

Ratio measures of association are suited to dichotomous (i.e., two-category) measures, such as presence of disease (yes or no) or exposure (yes or no). If the exposure has multiple categories (for example, different types of industrial solvents or several levels of exposure), a ratio measure of effect can be computed for each type or level compared to the unexposed group (if there is no unexposed group, then one exposure or level can be selected as a reference category). Consider, for example, the classic study by Wynder and Graham (1950) on lung cancer and cigarette smoking. In this case, "None (less than 1 per day)" is selected as the reference category, and the odds ratio is computed for each higher level of smoking relative to the reference level.

Cigarette smoking histories of 605 male lung cancer patients and 780 controls

Amount of cigarette smoking for 20+ years.* (percent distribution)	Lung cancer patients [N=605]	Controls [N=780]	OR
None (less than 1 per day)	1.3	14.6	1.0**
Light (1-9 per day)	2.3	11.5	2.2
Moderately heavy (10-15 per day)	10.1	19.0	6.0
Heavy (16-20 per day)	35.2	35.6	11.1
Excessive (21-34 per day)	30.9	11.5	30.2
Chain (35+ per day)	20.3	7.6	30.0

* includes pipe and cigar smokers, with a conversion formula.

** reference category.

The odds ratios (OR) are obtained by forming a 2x2 table for each exposure level relative to the reference level. For example, for "Heavy (16-20 per day)" compared to "None":

	Lung cancer	Control
Heavy	35.2	35.6
None	1.3	14.6

$$OR = \frac{35.2 \times 14.6}{35.6 \times 1.3} = 11.1$$

(As stated earlier, the OR calculation can be done just as easily from percentages as from the actual numbers of cases and controls, since the first step is to derive proportions from the numbers. The fact that these percentages are age-adjusted actually means that the ORs are age-adjusted as well.)

The odds ratios reveal quite the existence of a marked dose-response relationship.

Measures of association – non-dichotomous disease

When the disease or outcome variable is not dichotomous (e.g., body mass index) but the exposure is, the outcome variable can be categorized (e.g., "above or below 30% greater than ideal weight") to enable computation of ratio measures of association. Alternatively, a summary statistic (e.g., mean body mass) can be computed for each category of the exposure, but then we have no measure that can be interpreted as relative risk.

When both disease and exposure have multiple ordered categories (e.g., injury severity rating with several levels (an ordinal variable), parity (a count variable), or blood pressure (a continuous measure), categorization can be imposed to obtain a ratio measure of effect. Alternatively, the relationship between outcome and exposure can be plotted, and the slope used as a measure of the

strength of the relationship (e.g., a 2 mmHg increase in diastolic blood pressure for every 14 grams of alcohol consumed is stronger than a 1 mmHg increase for every 14 grams). Linear regression coefficients are used to estimate the slope of the relationship and provide a satisfactory index of strength of association for continuous variables, though one that cannot readily be compared to measures of relative risk. We will return to regression coefficients later in the course.

Correlation coefficients are often used as measures of association between ordinal or continuous variables, but as explained below, these are not regarded as epidemiologic measures of strength of association.

Other measures of association

> "When I use a word, it means precisely what I want it to, neither more nor less" (Lewis Carroll, Alice in Wonderland)

As mentioned earlier, a point of confusion for the learner is the difference between what epidemiologists mean by a measure of association and what is measured by various statistics that are also referred to as measures of association. To clarify this unsatisfactory state of affairs, we will discuss two measures that are widely used in both epidemiology and other disciplines, but which epidemiologists regard as very different from the measures of association we have discussed above.

Chi-square for association

A nearly ubiquitous statistic in epidemiology is the chi-square for association. The chi-square and its associated p-value address the question of the degree to which an association observed in a sample is likely to reflect an association in the population from which the sample was obtained, rather than simply have arisen due to sampling variability. The p-value estimates the probability that variability of random sampling can result in two variables being associated in a sample even if they are entirely independent in the population. Although there is obviously a connection between the question addressed by the chi-square and the question addressed by the relative risk, the two questions are by no means interchangeable. For example, consider the table at the very beginning of this chapter.

CHD and oral contraceptives (OC) in women age 35 years or more

	OC	$\overline{\text{OC}}$	Total
CHD	30	20	50
$\overline{\text{CHD}}$	30	70	100
Total	60	90	150

Regarding these data as having come from a hypothetical case-control study, we select the odds ratio (OR) as the appropriate measure of strength of association. Since CHD is a rare disease, the OR will estimate the CIR as well as the IDR. The OR for the above table is:

$$OR = \frac{30 \times 70}{20 \times 30} = 3.5$$

i.e., the observed association suggests that the risk of CHD in women 35 years or older who use OC is 3.5 times that of similarly aged women who do not use OC.

The chi-squared statistic for this table will yield a p-value that approximates the probability that a table with an OR of 3.5 or stronger will arise from a random draw of 50 women (who will be called "cases") from a population of 60 OC users and 90 nonusers. That chi-squared statistic is 12.4, which corresponds to a very small probability – much lower then 0.0001, or 1 in a thousand draws (the computation will be covered in a later part of the course). Suppose instead that the study that yielded the above table had been only one-fifth as large. Keeping the same proportion in each of the four cells, we would then have this table:

CHD and oral contraceptives (OC)in women age 35 years or more

	OC	$\overline{OC}$	Total
CHD	6	4	10
$\overline{CHD}$	6	14	20
Total	12	18	30

The odds ratio for this table is still 3.5, but the chi-squared statistic is now only 2.42, which corresponds to a p-value of 0.12. The greater p-value results from the fact that it is much easier to obtain an association with OR of 3.5 or greater by randomly drawing 10 "cases" from a room with 12 OC users and 18 nonusers than by randomly drawing 50 "cases" from a room with 60 OC users and 90 nonusers.

Since the OR remains identical but the chi-squared statistic and its p-value change dramatically, clearly the epidemiologic measure of association and the chi-square are measuring different features of the data. The chi-squared statistic is used to evaluate the degree of numerical evidence that the observed association was not a chance finding. The epidemiologic measure of association is used to quantify the strength of association as evidence of a causal relationship.

Correlation coefficients

Correlation coefficients are measures of linear or monotonic associations, but again not in the same sense as measures of relative risk. The linear correlation coefficient (Pearson or product-moment correlation, usually abbreviated "r") measures the degree to which the association between two variables is linear. An r of zero means that the two variables are not at all linearly related (they may nevertheless be associated in some other fashion, e.g., a U-shaped relationship). An r of +1 or -1 means that every pair of observations of the two variables corresponds to a point on a straight line

drawn on ordinary graph paper. However, knowing whether or not the relationship is linear tells us nothing about the steepness of the line, e.g., how much increase in blood pressure results from a 5% increase in body mass. Other correlation coefficients (e.g., Spearman) measure the degree to which a relationship is monotonic (i.e., the two variables covary, without regard to whether the pairs of observations correspond to a straight line or a curve).

Epidemiologists think of the relationships between variables as indications of mechanistic processes, so for an epidemiologist, strength of association means how large a change in risk or some other outcome results from a given absolute or relative change in an exposure. If the assumption is correct, the strength should not depend upon the range of exposures measured or other aspects of the distribution. In contrast, r is affected by the range and distribution of the two variables and therefore has no epidemiologic interpretation (Rothman, p.303). Standardized regression coefficients are also not recommended for epidemiologic analysis for similar reasons (see Greenland, Schlesselman, and Criqui, 1986).

Correlation coefficients between dichotomous variables — Correlation coefficients can be particularly problematic when used to quantify the relationship between two dichotomous (binary) factors, especially when one or both of them are rare. The reason is that correlation coefficients between binary variables cannot attain the theoretical minimum (-1) and maximum (+1) values except in the special case when the both factors are present half of the time and absent half of the time (Peduzzi, Peter N., Katherine M. Detre, Yick-Kwong Chan. Upper and lower bounds for correlations in 2 x 2 tables–revisited. *J Chron Dis* 1983;36:491-496). If one or both factors are rare, even if the two variables are very strongly related, the correlation coefficient may be restricted to a modest value. In such a case an apparently small correlation coefficient (e.g., 0.15) may actually be large in comparison with the maximum value obtainable for given marginal proportions.

For example, the correlation coefficient between smoking and lung cancer cannot be large when the proportion of lung cancer cases is small but that of smokers is large, as shown in the following example (Peduzzi PN, Detre KM, Chan YK. Upper and lower bounds for correlations in 2 x 2 tables–revisited. *J Chron Dis* 1983;36:491-496) based on data from Allegheny County, PA:

	Smoker	Nonsmoker	Total
Lung cancer	20	2	22
No lung cancer	14,550	9,576	24,126
Total	14,570	9,578	24,148
Lung cancer incidence			0.001
Smoking prevalence			0.60
Odds ratio			6.6
	Correlation coefficient (r)	R-square (R^2)	
Based on above data	0.190	0.00036	
If all cases were smokers	0.024	0.00058	
If no cases were smokers	-0.037	0.00157	

Here, the correlation coefficient (r) is a meagre 0.019, with a corresponding R^2 ("proportion of variance explained") of 0.000356. Even if all 22 lung cancer cases were smokers, the correlation coefficient would rise only to 0.024 (with $R^2 = 0.0006$), and if no lung cancer cases smoked r falls only to -0.037. In contrast, the OR is 6.6, indicating a strong relationship (the RR and IDR are essentially the same, since the outcome is so rare). Therefore the correlation coefficient and proportion of variance explained are not readily applicable to relationships between dichotomous variables, especially when the row or column totals are very different.

Measures of Impact

Concept

Relative risk measures compare the risk (or rate) in an exposed group to that in an unexposed group in a manner that assesses the strength of association between the exposure and outcome for the purpose of evaluating whether the association is a causal one, as we will see in the chapter on Causal Inference. But when we have decided (or assumed) that the exposure causes the outcome, we often wish to assess the individual and/or public health importance of a relationship, i.e.,

- How much of a disease can be attributed to a causative factor?
- What is the potential benefit from intervening to modify the factor?

The answers to these questions enter into public health policy-making and, in principle, individual decision-making, since they indicate the amount or proportion of the burden of a disease that can be prevented by eliminating the presumed causal factor (e.g., pollution control) or by carrying out a preventive intervention (e.g., fortification of foods). Examples of the kind of questions that prompt the use of measures of impact are:

1. Now that I am 35 years old, my CHD risk from taking oral contraceptives is twice as great as when I was 25. But *how much more risk* do I have due to taking the pill?
2. In HIV-discordant couples in which a condom is not used and one partner has a bacterial sexually transmitted disease, *how much of the risk* of heterosexual transmission of HIV is due to presence of the sexually transmitted disease and therefore might be eliminated through STD control measures?
3. *How many cases* of asthma are due to ambient sulfur dioxide?
4. *What proportion of motor vehicular deaths* can be prevented by mandatory seat belt use.
5. What proportion of perinatal HIV transmission has been prevented through the use of prenatal, intrapartum, and neonatal zidovudine?

To answer these questions we employ **attributable fractions**, which are measures of **impact** or **attributable risk**.. The concept of attributable is of central importance for public health, since it addresses the question of "so what?". Although some students find the topic of attributable risk a source of confusion, at least some of their confusion is attributable (!) as much to the terminology as to the basic concept. There are, however, a number of subtleties and legitimate sources of confusion related to attributable risk. To introduce the concept we make the simplifying assumptions that the exposure in question has either adverse or beneficial effects but not both, that the exposed and unexposed groups are identical except for the exposure, and that either no person is susceptible to getting the outcome from both the exposure and some other causal factor (i.e., a person who will experience the outcome due to some other causal factor will not experience it due to the exposure, and vice-versa). We also begin by focusing on risks and proportions, rather than on rates.

One more prefatory note: at the risk of provoking a reaction of "Duh!", I will note that questions of attributable risk arise only in situations where more than one factor can cause the outcome under consideration. When the outcome has only a single causal factor (typically, where the outcome is defined in terms of the etiologic agent, as with infectious diseases) all of the cases must be attributable to that factor. Eliminating the factor would avoid all risk. If a necessary cause ("C") requires a co-factor or susceptibility factor ("S") for the effect to occur, then all of the cases are attributable both to "C" and to "S". This last point also illustrates that attributable fractions do not sum to 1.0, even though they are often expressed as percentages.

Perspectives

There are a variety of different measures of impact, and at least twice that many names for them. (For example, the term "attributable risk" is sometimes used to refer to the risk difference, sometimes to the population attribute risk proportion described below, and sometimes to the class of measures of impact. See Rothman and Greenland for various usages, with citations) One reason for the multiplicity of measures is simply to have a measure for each of the various ways to ask a question about impact. That is, the question can be asked in absolute ("How much" risk) or relative ("What proportion" of risk) terms. It can be asked with reference specifically to persons exposed to the factor or with reference to the whole population. Also, the factor being considered may cause or prevent the outcome. Various combinations of these alternatives call for different measures. The justification for having more names than measures (and for using the same name for different measures) is unclear.

Absolute perspective

The **absolute perspective** for attributable risk is expressed by the questions, "**How much** of the risk is attributable to the factor?" and "How many cases might be avoided if the factor were absent?" The answer is obtained by estimating the **risk difference** or the difference in the number of cases for exposed and unexposed persons. The risk difference, for example, provides an estimate of the **amount of risk in exposed persons** that is "attributable" to the factor (assuming causality). If we are interested in the amount of risk in that is attributable to the exposure **in the total population** (assuming causality), we multiply the risk difference by the **exposure prevalence** in the population. If we are interested in the **actual number of cases** that are attributable, i.e., that could have been avoided by complete elimination of the exposure (before any irreversible effects have occurred), we can multiply the risk difference by the population size.

Relative perspective

The **relative perspective** for attributable risk is expressed by the question, "**What proportion** of the risk is attributable to the factor?" and "What proportion of the cases of the disease might be avoided if the factor were absent?". Here, we need to express the amount of risk attributable to the factor relative to the total risk in exposed persons or in the total population. The measure for the exposed population is sometimes referred to as the "**attributable risk proportion**" (ARP) or the "**excess fraction**" (see Rothman and Greenland). The measure for the entire population is sometimes referred to as "population attributable risk proportion" (PARP).

Attributable risk proportion

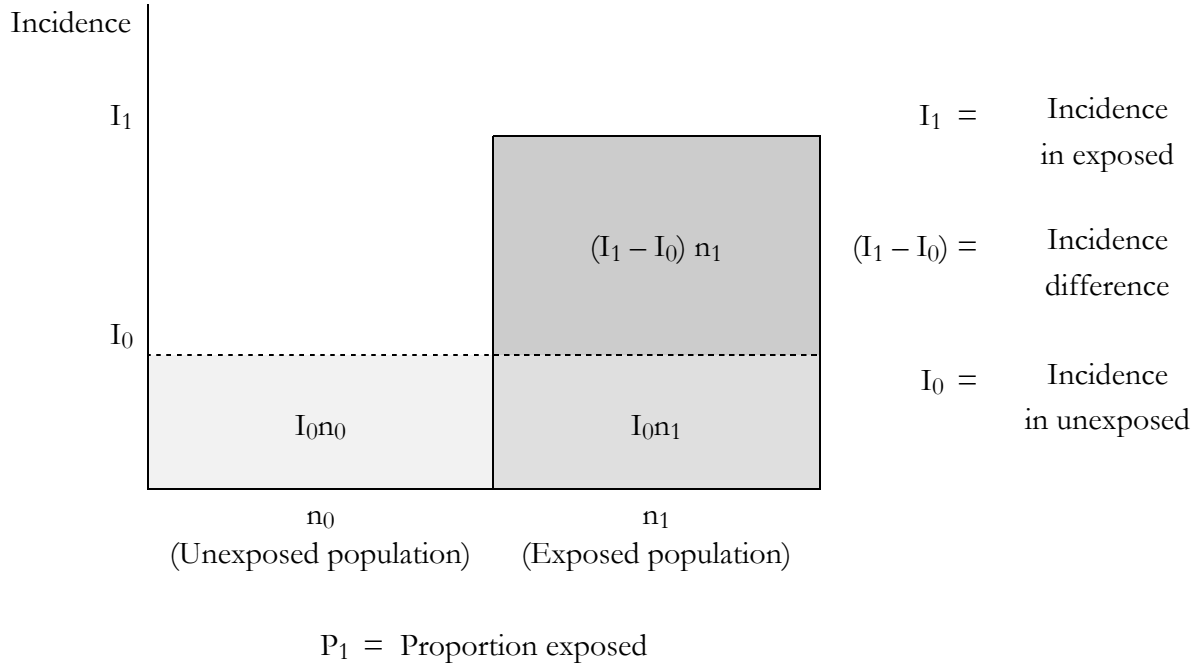
The ARP is directly related to the strength of the association between the exposure and the disease – if the exposure doubles the risk, then half of the risk is attributable to the exposure; if the exposure triples the risk, then two-thirds of the risk is attributable to the exposure; if the exposure multiples the risk fourfold, then the ARP is three-fourths, etc.

Population attributable risk proportion

The PARP reflects not only the strength of association but also the prevalence of the exposure in the population. Obviously an exposure can do more damage (have more impact) if it is experienced by many people, rather than only a few. The PARP adds this consideration to the ARP. Note that many older texts and articles refer to the PARP simply as "attributable risk".

The following diagram displays the basis for the various measures of attributable risk. The basic idea is, simply, that if we observe incidence I_1 in an exposed population and a lower incidence I_0 in a comparable unexposed population, and we make the assumption that the exposure is causing the higher incidence in the exposed population, then it is logical to suppose that the difference, $I_1–I_0$, is the amount of incidence that is due to the exposure. Then, depending on the way in which we are asking the question, this "attributable incidence" is expressed as an absolute difference or as a relative difference, and in relation to exposed persons only or to the entire population.

Diagrammatic representation of attributable risk in a population



In the above diagram:

n_0 and n_1 represent, respectively, the numbers unexposed and exposed persons, or the amounts of unexposed and exposed population-time; $n = n0 + n_1$

P_0 and P_1 represent, respectively, the proportions of unexposed and exposed persons or population time (i.e., $P_1 = n_1/n$)

I_0 is the incidence proportion (cumulative incidence) of disease in unexposed persons, so I_0n_0 is the expected number of cases among unexposed persons, i.e., the area of the unshaded rectangle.

I_1n_1 is, similarly, the expected number of cases among exposed persons, i.e., the combined area of the two shaded rectangles.

$(I_1–I_0)$ is the incidence difference or attributable risk. It gives the amount of incidence "attributable" to exposure, i.e., the amount of incidence in exposed persons over and above the incidence they would be expected to have had (I0) in the absence of exposure.

$(I_1–I_0)n1$ is the expected number of cases among exposed persons beyond those expected from their background incidence (I_0), i.e., attributable cases (the area of the cross-hatched rectangle). Attributable cases are simply the attributable risk multiplied by the number of exposed persons.

RR is the relative risk (risk ratio, CIR), I_1/I_0

The attributable risk proportion in exposed persons [ARP] is the proportion of exposed cases that is "attributable" to the exposure. This proportion is:

$$\text{ARP} = \frac{\text{"Attributable cases"}}{\text{All } \textit{exposed} \text{ cases}} = \frac{(I_1 - I_0)n_1}{I_1n_1} = \frac{I_1 - I_0}{I_1} = \frac{RR - 1}{RR}$$

(the RR's are obtained by dividing numerator and denominator by I_0).

Similarly, the population attributable risk proportion [PARP], the proportion of all cases that is attributable to exposure, is:

$$\text{PARP} = \frac{\text{"Attributable cases"}}{\text{All } \textit{cases}} = \frac{(I_1 - I_0)n_1}{I_1n_1 + I_0n_0} = \frac{I_1n_1 - I_0n_1}{I_1n_1 + I_0n_0} = \frac{P_1(RR–1)}{1 + P_1\,(RR–1)}$$

The right-hand formula (see the assignment solution for its derivation) displays the relationship of the PARP to exposure prevalence and "excess risk" (RR-1). The denominator cannot be less than 1, so if the numerator is very small (e.g., very low exposure prevalence and/or weak association), then the PARP will also be very small. Conversely, for a very prevalent exposure (e.g., P_1=0.80) and very strong association (e.g., RR=9), then the numerator [0.80 x (9-1)] will be large (6.4). The denominator will be close to this value, since the 1 will have little influence. Thus, the PARP will

show that a large proportion (i.e., close to 1.0) of the cases are attributable. As the prevalence rises, the PARP comes closer to the ARP (when P1=1, as it does in the exposed population, the PARP formula reduces to that for the ARP).

The joint influence of strength of association and prevalence of exposure on the PARP may be easier to see in the following algebraic reformulation:

$$PARP = \frac{1}{1 + 1/[P_1(RR-1)]}$$

Definitions and formulas

Attributable risk [absolute]: the amount of the risk in the exposed group that is related to their exposure. Attributable risk is estimated by the cumulative incidence difference or incidence density difference:

$$AR = I_1 - I_0$$

Population attributable risk [absolute]: the amount of risk in the population (i.e., in exposed and unexposed persons taken together) that is related to exposure. Population attributable risk is equal to the attributable risk multiplied by the prevalence of the exposure:

$$PAR = AR \times P_1 = (I_1 - I_0)P_1 = I - I_0$$

[This measure is not often used, but is helpful here to complete the pattern. "I" without a subscript refers to the total, or crude incidence. The equivalence of the middle and right-hand terms in the above expression can be seen by substituting $(I_1P_1 + I_0P_0)$ for I and $(I_0P_0 + I_0P_1)$ for I_0.]

Attributable risk proportion (or percent) [ARP]: the proportion (percent) of the risk in the exposed group that is related to their exposure.

$$ARP = \frac{I_1 - I_0}{I_1} = \frac{RR - 1}{RR} = \frac{AR}{I_1}$$

Population attributable risk proportion (or percent) [PARP]: the proportion (percent) of the risk in the population that is related to the exposure.

$$PARP = \frac{P_1 (RR - 1)}{1 + P_1 (RR - 1)} = \frac{I - I_0}{I} = \frac{PAR}{I}$$

(Derivations are shown in the assignment solutions)

Case-control studies

The absolute measures (AR and PAR) require estimates of incidence, so they cannot be estimated from the results of a case-control study without additional information on incidence. If the disease is rare, the ARP and PARP can be estimated from a case-control study by using the OR as an estimate of the RR. The ARP is then simply (OR – 1)/OR. A formula for the PARP can be derived using Bayes Theorem and algebra (see below):

$$PARP = \frac{P_{E|D}\,(RR - 1)}{RR} = (P_{E|D}) \times ARP$$

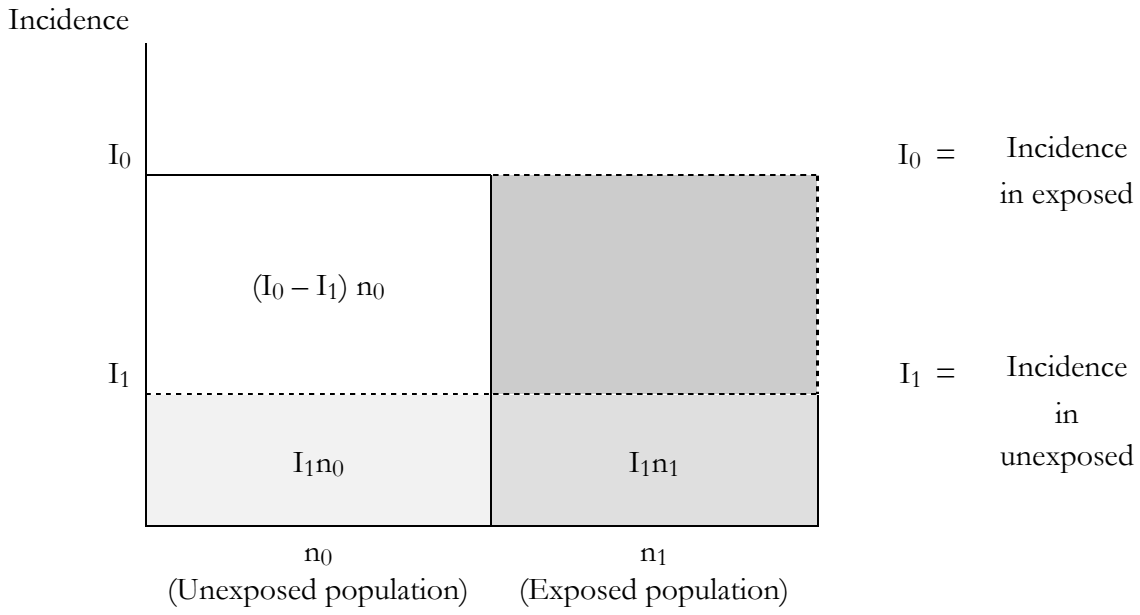
where $P_{E|D}$ is the proportion of cases who are exposed. Since the right-hand formula does not require knowledge of the exposure prevalence in the population nor the actual disease incidence, this formula can be estimated from a case-control study which gives an estimate of RR or IDR.

Preventive fraction

When $I_1 < I_0$ (e.g., for a vaccine, use of protective equipment, or pollution control devices), a "preventive" fraction is needed. Since a protective exposure (assuming causality) reduces the risk of the outcome, we cannot think in terms of "attributable cases" since the "cases" have not occurred! Instead, we define a preventive fraction as a proportion of "potential cases" that were prevented, i.e., that did **not** occur because of the protective exposure. For vaccines, this proportion is referred to as **vaccine efficacy** or effectiveness.

As with attributable risk, there are two variants, one for those exposed to the preventive intervention and one for the population as a whole (both are based on the "relative" perspective; the absolute perspective does not appear to be used). The following diagram, similar to that for attributable fractions, will be used.

Diagrammatic representation of preventive fraction in a population



P_1 = Proportion exposed

Where n_1, n_0, I_1, I_0 are as before and $(I_0 - I_1)n_1$ denotes the "prevented cases", i.e., the number of potential cases that would have occurred if the exposure were not associated with lower incidence (recall that I_0 is greater than I_1) or had not been present. $I_1 n_1$ are the cases that occurred in spite of the intervention.

Therefore, the preventive fraction in the exposed (PF_1) quantifies the prevented cases as a proportion of all potential cases in exposed persons. The preventive fraction in the population (PF) expresses the prevented cases as a proportion of all potential cases in the entire population. In each case, the "prevented cases" are cases that would have occurred but for the preventive exposure; the "potential cases" are prevented cases plus actual cases.

From the diagram:

Preventive fraction in exposed

(PF_1 - for those exposed to the preventive measure)

$$PF_1 = \frac{\text{“Prevented } \textit{potential} \text{ cases”}}{\text{All } \textit{potential exposed cases}} = \frac{(I_0 - I_1)\, n_1}{I_0 n_1} = \frac{(I_0 - I_1)}{I_0} = 1 - RR$$

(since $I_1 < I_0$, $RR < 1.0$).

Preventive fraction in the population (PF)

$$PF = \frac{\text{“Prevented } \textit{potential} \text{ cases”}}{\text{All } \textit{potential cases}} = \frac{(I_0 - I_1)\, n_1}{I_0\, n} = \frac{(I_0 - I_1)\, P_1}{I_0} = P_1\, PF_1$$

(recall that n_1/n is the proportion exposed, P_1).

The preventive fraction represents the proportion (or percent) of the potential burden of disease which is prevented by the protective factor. The following formula displays this aspect clearly:

$$PF = \frac{(I_0 - I_1)\, n_1}{I_0\, n} = \frac{(I_0 n_1 - I_1 n_1) + (-I_0 n_0 + I_0 n_0)}{I_0\, n} = \frac{(I_0 n_1 + I_0 n_0)}{I_0\, n} - \frac{(I_1 n_1 + I_0 n_0)}{I_0\, n} = \frac{I_0 - I}{I_0}$$

I_0 is the risk in people unexposed to the preventive measure. If no one received its benefits, then the risk in the entire population would also be I_0. The actual overall risk, I, represents an average of the risks for those exposed to the preventive measure and those not exposed, weighted by their respective sizes ($I_1 n_1 + I_0 n_0$). So I_0–I is the difference between the risk that *could have been* observed and the risk that *was* observed, which difference is assumed to be attributable to effectiveness of the preventive measure and its dissemination. The last formula expresses this difference as a proportion of the risk in the absence of the preventive measure.

In all of these measures, of course, the assumption is made, at least for purposes of discussion, that the relationship is causal, and in some cases, that removing the cause (or introducing the preventive factor) is fully and immediately effective. In any specific example, of course, the latter assumption can be varied.

Unified approach to attributable risk and preventive fraction

Although there on many subtleties, the basic idea of attributable risk and preventive fraction is simple. That simplicity is overshadowed by the array of formulas. The following conceptualization brings out the underlying simplicity and may be the easiest way to derive formulas when needed.

The basic objective is to quantify the impact of an exposure or preventive measure in terms of the burden of a disease. Large impacts come from:

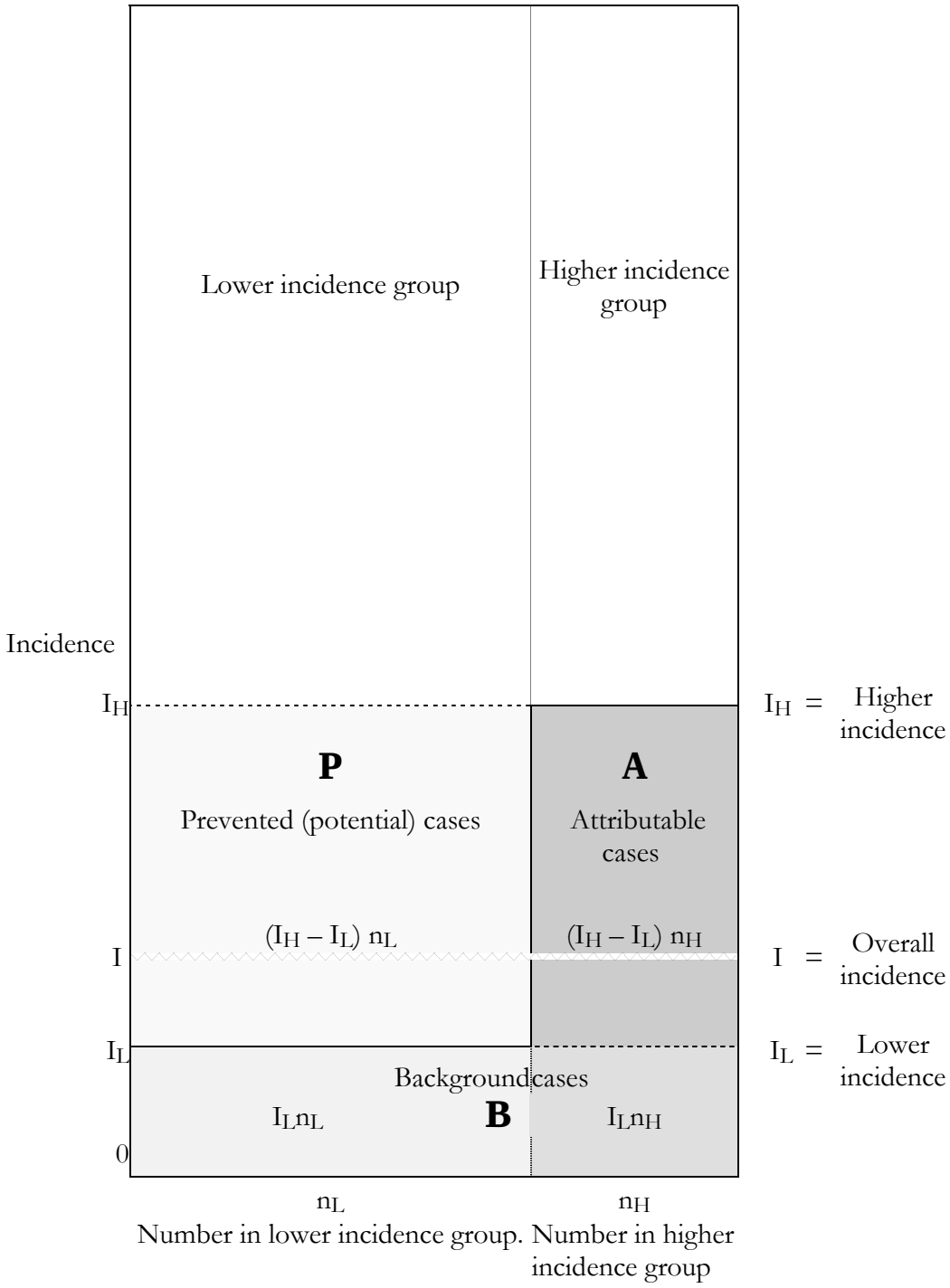
1. high frequency of disease
2. powerful risk or preventive factor
3. large proportion of people exposed to the factor

One aspect that complicates the formulas is the fact that incidence in people exposed to a risk factor is greater than in the unexposed, but incidence in people exposed to a preventive factor is lower than in the unexposed. We can side-step this difference by thinking in terms of the higher incidence and the lower incidence.

The diagram on the following page represents a population at risk in which people can be classified into two exposure groups, one with lower incidence (e.g., physically active) and the other with higher incidence (e.g., sedentary). The width of each tall rectangle indicates the number of people in the corresponding exposure group. Physical activity and sedentary lifestyle make a good example, because they will work as well for the risk factor (attributable risk) perspective and the preventive factor perspective. Let us use I_L and I_H to represent the lower and higher incidence, and N_L and N_H to represent the number of people or amount of population time in the lower and higher incidence categories, respectively.

In this diagram, rectangle **A** [$N_H (I_H - I_L)$] represents the **attributable** cases. These are the cases that would not have occurred were it not for the risk factor (or for the absence of the preventive factor). Rectangle **P** [$N_L (I_H - I_L)$] represents the **prevented** cases. These cases are only potential, since they have not occurred. They are the cases that would have occurred had the preventive factor not been present (or if the risk factor were to spread to the lower incidence group). Rectangle **B** [$N_L I_L + N_H I_L$] represents the unavoidable (background) cases. They occur despite the preventive factor or the absence of the risk factor. The total number of cases is represented by the sum of the rectangles for the two exposure groups [$N_L I_L + N_H I_H$]. If I is the overall (crude) incidence, then the total number of cases can also be written as [$(N_L + N_H) I$]. The total of potential cases (i.e., those observed plus those prevented) corresponds to [$(N_L + N_H) I_H$], the result of subjecting the entire population to the higher incidence.

Diagrammatic representation attributable and prevented cases



n_1 = Number with the risk or preventive factor.

P_1 = Proportion with the risk or preventive factor.

With this diagram and notation we can express both attributable and preventive fractions in a more parallel manner. The population attributable risk (PARP) is simply $\mathbf{A}/(\mathbf{A} + \mathbf{B})$ and the prevented fraction (PF) is simply $\mathbf{A}/(\mathbf{A} + \mathbf{B} + \mathbf{P})$. We can therefore write the formulas we derived earlier as:

$$\text{PARP} = \frac{\text{“Attributable cases”}}{\text{All cases}} = \frac{\mathbf{A}}{\text{All cases}} = \frac{N_H (I_H - I_L)}{N_L I_L + N_H I_H} = \frac{P_H (I_H - I_L)}{I}$$

The last step made use of the facts that $N_H/(N_L + N_H)$ is the prevalence of the exposure and that the overall incidence I equals the total cases divided by the total population. Since the attributable risk proportion (ARP) concerns only the exposed group, $P_H = 1$ and $ARP = (I_H - I_L)/I_H$, which also equals $(RR - 1)/RR$.

Similarly, the prevented fraction is:

$$\text{PF} = \frac{\text{“Prevented cases”}}{\text{All } \textit{potential} \text{ cases}} = \frac{\mathbf{P}}{(N_L + N_H) I_H} = \frac{N_L (I_H - I_L)}{(N_L + N_H) I_H} = \frac{P_L (I_H - I_L)}{I_H}$$

If we divide numerator and denominator by I_H, we obtain $P_L (1 - RR)$. The prevented fraction in those exposed to the preventive intervention concerns only the lower incidence group, so P_L and $PF_1 = (1 - RR)$.

With this notation we can see the essential equivalence between the PARP and the PF. They both involve the incidence difference times the number of people in the “exposed” group. They both are expressed as a proportion of all cases, except that for the PF we need to include the potential cases in the denominator – otherwise it would not be a proportion.

PARP in a case-control study

Case-control studies, as will be discussed in the following chapter, do not provide an estimate of incidence unless additional information is available. Nevertheless, we can use incidence ratios (RR) instead of incidences, and then we can use the OR to estimate the RR. If the control group is population-based, it may provide a direct estimate of exposure prevalence. If not, we can modify the formula and use the prevalence of exposure in the cases. This prevalence is simply the exposed cases ($N_H I_H$) divided by all cases ($N_L I_L + N_H I_H$), which is very similar to the PARP formula.

The result, $(N_H I_H)/(N_L I_L + N_H I_H)$, is very similar to the PARP formula. The only difference is in the numerators: $N_H I_H$ versus $N_H (I_H - I_L)$. We can get from the exposure prevalence expression to the PARP by multiplying the former by $(I_H - I_L)/I_H$ or, equivalently, by $(RR - 1)/RR$, which we can estimate by $(OR - 1)/OR$ if the disease is rare. So we can estimate PARP for a rare disease from a case-control study that can measure neither incidences nor exposure prevalence by using OR to estimate RR in:

$$PARP = \frac{N_H (I_H - I_L)}{N_L I_L + N_H I_H} = \frac{P_{H|D} (RR - 1)}{RR} = P_{H|D} (ARP)$$

This diagram and these formulas are worth becoming very familiar with, since doing so will help to develop an in-depth understanding of incidence, prevalence, relative risk, impact, and weighted averages and also to derive any of the basic attributable risk or preventive fraction formulas. Picture how the diagram will change as the prevalence of the exposure increases or decreases. How will the wavy line (the overall incidence, I) move as the other variables change? How is it related to I_L, I_H, N_L, N_H? (This is definitely a key relationship to know). What happens when everyone is exposed? When no one is exposed? What is the prevalence of exposure in cases? How will it change as incidence and/or overall exposure prevalence change?

Interpreting attributable fractions

Although the basic concept of attributable risk is intuitively meaningful, it turns out that it has many subtleties and nuances. Appreciation of the subtleties has come only relatively recently, so that much of what has been written (even by yours truly) and some of what is still being written (hopefully not by yours truly) is not completely accurate. The confusion is aggravated by the multitude of terms that have been introduced, with usages that differ from one author to another. In addition to multiple terms for the same concept (a common problem in epidemiology), there are also instances where a single term is applied to different concepts. Therefore, if you find yourself being confused by something you are reading in this area, always consider the possibility that what you are reading may be confused as well.

The complications arise when we try to interpret attributable fractions (e.g., ARP, PARP) in etiologic (causal) terms, which is of course what we were interested in at the outset. Consider the following two questions, which figure prominently in product liability litigation, where courts have held that recovery requires a finding that the plaintiff's disease was "more likely than not" a consequence of exposure to the product (e.g., asbestos, prescription drugs, silicone breast implants, tobacco).

- Among nonsmokers exposed to X, *what proportion of Y were caused by* X?
- What is the probability that person Z's case of Y *resulted from* X?

What distinguishes these two questions from the illustrative ones at the beginning of the section is the use of causal terminology ("caused by", "resulted from") instead of the more general (and

vaguer) "attributed to". Incidence and measures derived from incidence show only over, or net effects, not the causal processes that produce them. For example, even though a sedentary lifestyle increases the risk of coronary heart disease, physical exercise can acutely increase the risk of a cardiac event. When we compare the rate of cardiac events in a sedentary group to the rate in a group of people who get regular exercise, the difference in incidence rates measures the increased rate of cardiac events *associated with* a sedentary llifestyle. But if some of the incidence of cardiac events in exercisers actually *results from exercising*, then the difference in incidence between the two groups measures the *net* harm from a sedentary lifestyle, rather than the *total* effect. By comparing the incidence rates we are letting the cardiac events in the exercisers offset some of the events in the sedentary group, with the relative size of benefit and harm depending upon the kinds of people (e.g., genetic characteristics or distributions of other exposures) who are exercise and do not exercise. In general, epidemiologic data will not reveal what contributes to the net incidence difference.

Similarly, if the action of one causal factor can preempt the opportunity for another factor to cause the disease (because the disease has already occurred), then there is no way to know from epidemiologic data which factor caused the disease in a person or population exposed to both causal factors. For this reason, it is problematic to interpret attributable risk measures as **etiologic fractions**, although many writers have used the terminology interchangeable. According to Greenland (1999: 1167), the "key fallacy in much of the literature and testimony regarding the probability of causation is the use of the following generally incorrect equations: Etiologic Fraction = Rate Fraction and Probability of Causation = Rate Fraction . . .", where the etiologic fraction (EF) is "the fraction of these individuals for whom exposure was a contributory cause of the disease" (Greenland, 1999: 1166) and the rate fraction (RF) is the incidence rate difference divided by the incidence rate in the exposed (analogous to the ARP, except derived from incidence rates rather than incidence proportions) (p1167). In algebraic terms, $EF=(A_1+A_2)/A_T$, where A_1 are exposed persons who would have developed the disease at some point but whose disease was accelerated due to the exposure, A_2 are exposed persons whose disease would never have occurred without the exposure, and A_T is A_1+A_2 plus exposed persons who develop the disease completely independently of exposure. The EF estimates the **probability of causation**, since $(A_1+A_2)/A_T$ is the probability that a person randomly selected from A_T had his/her disease accelerated by (A_1) or completely caused by (A_2) the exposure. The proportion A_2/A_T is the **excess fraction**, since it gives the proportion of the total caseload that would not have occurred without the exposure (Greenland, 1999), regardless of time to occurrence. (Greenland observes that the failure to distinguish the excess fraction from the etiologic fraction is a "major problem in most of the literature", and regards the term "attributable risk" as particularly misleading even though it "dominates the American literature", both in biostatistics and epidemiology [p.1168].)

Answer to question at beginning of the chapter about the association between CHD and OC:

> The proportion of CHD cases in the sample of 40 must be somewhere between 30/60 = 0.5 (the proportion of cases among OC users) and 20/90 = 0.2222 (the proportion among nonusers). If the sample consists of 22 users and 18 nonusers, then the best estimate of the sample proportion of CHD cases is:

$$\text{Proportion with CHD} = 0.5\left(\frac{22}{40}\right) + 0.2222\left(\frac{18}{40}\right) = 0.5(0.55)+0.2222(0.45) = 0.375$$

> Therefore, the best estimate of the overall proportion with CHD is approximately 0.375 or 15 women in the sample of 40.

Summary

There are three categories of measures: Frequency/extent, association, impact

(1) Measures of frequency or extent (especially prevalence and incidence)

In epidemiology, incidence is the occurrence of any new health-related event (e.g., disease, death, recovery). Incidence is quantified as a:

PROPORTION: the proportion of a population who experience the event; also called "RISK", since it estimates the average risk per person for the period. [Risk] ODDS are simply a transformation of risk [risk/(1-risk)].

RATE: the number of health events per person per unit time; corresponds to the average risk per person per unit time.

MEASURE	EPIDEMIOLOGIC ESTIMATOR	UNITS	LIMITS
Risk	Cumulative Incidence (CI)	Dimensionless	0 to 1
Rate	Incidence Density (ID)	1/time	0 to "infinity"
$Odds_r$	CI / (1-CI)	Dimensionless	0 to "infinity"

CI (a proportion) is used to estimate an individual's risk of developing a disease. ID (a rate) is used to estimate the force intensity of occurrences. Risk and rate are related, since the greater the intensity of occurrences in a population, the greater the risk of an event to any member of the population. When CI is small (i.e., because of a low the intensity of disease or a short time interval), ID is approximately equal to CI divided by the number of years of followup. When CI is not small, the relationship is more mathematically complex.

Application

The choice of an incidence measure (either CI or ID) depends upon:

a. OBJECTIVES OF THE STUDY

CI provides a direct estimate of an individual's risk, as may be useful for making clinical and/or personal decisions;

ID is often preferred for assessment of the population impact of a health event or for testing etiologic hypotheses.

b. PRACTICAL CONSIDERATIONS

CI may be preferred:

- if the health event has a restricted risk period
- if it is difficult to ascertain time of change in health status
- for ease of comprehension.

ID may be preferred:

- if the health event has an extended risk period
- if lengths of follow-up vary
- if there is a large loss to follow-up
- if the health event can recur (e.g., bone fractures).

A ratio of two risk estimates is a "risk ratio" (RR). A ratio of two rate estimates is a "rate ratio" (RR). A ratio of two odds is an "odds ratio" (OR). All these measures are sometimes referred to as "relative risk" (RR), though strictly speaking only the first pertains to risk.

(2) Measures of association

e.g. Ratios of proportions (CIR), ratios of rates (IDR), ratios of odds (OR_r and OR_e)

$$\text{CIR} = \frac{a / (a + c)}{b / (b + d)} = \frac{\text{CI in exposed}}{\text{CI in unexposed}} = \text{"risk ratio", "relative risk"}$$

where a=exposed cases, b=unexposed cases, c=exposed noncases, d=etc.

In probability terms, CIR = Pr(D|E) / Pr(D|E)

How do we interpret the CIR?

a. If CIR = 1, then no association between exposure and disease.
b. If CIR > 1 then exposure appears to be associated with increased risk of disease, i.e., exposure may be harmful.
c. If CIR < 1 then exposure appears to be associated with decreased risk of disease, i.e., exposure may be protective.

(CIR's less than 1.0 can be awkward to think about, so it many cases it is helpful to reverse the disease or exposure category to obtain the reciprocal of the CIR. A CIR of 0.4 then becomes a CIR of 2.5)

CIR can be directly estimated if the exposure status is known before the occurrence of disease, as in a prospective followup study or a retrospective cohort study.

When a disease is rare, the OR_r approximates the CIR – a useful thing to know because logistic regression models may be employed to estimate odds ratios:

$$OR_r = \frac{\text{Odds of disease in exposed}}{\text{Odds of disease in unexposed}} = CIR$$

The OR (whether the "risk OR" or "exposure OR") is easy to calculate as the cross-product ratio: (a x d) / (b x c).

The risk and exposure OR's are calculated identically from a 2 x 2 table, but that doesn't mean they are equivalent "epidemiologically". Remember that the numbers in the 2 x 2 table are only an abstraction from the actual study experience and must be used with the design in mind (i.e., a case-control design is not equivalent to a longitudinal design). In a cohort study, we typically compute a CIR or an average IDR. In a follow-up study without a fixed cohort, we typically compute an IDR. In a case-control study we typically compute an OR. In a cross-sectional study, we typically compute a prevalence ratio or a prevalence OR. If one of the cumulative incidences is known, the OR estimate (e.g., from a logistic regression model – see chapter on Data Analysis and Interpretation) can be converted to a risk ratio estimate by the following formula (Zhang and Yu, 1998; notation changed to match that used in this chapter):

$$RR = \frac{OR}{(1 - CI_0) + (CI_0 \times OR)}$$

A prevalence ratio can be estimated from a prevalence odds ratio in the same manner, if the prevalence in the unexposed is known.

3) Measures of impact

"How much" of a disease can be attributed to an exposure can be considered as:

- an amount of the risk or incidence in the exposed (CID) or in the total population (usually presented as a number of cases)
- a proportion of the risk or incidence in the exposed (ARP) or in the total population (PARP).

The contributors to impact measures are:

1. Strength of association – affects all measures of impact.
2. Level of background incidence – affects only amount of incidence (CID, IDD)
3. Population prevalence of the exposure – affects only impact in the population (e.g., PARP).

Appendix — Relating risk factors to health outcomes

WARNING: this chapter has recently been converted to MS Word, so some algebraic errors may have been introduced.

Estimating exposure-specific incidence and attributable risk from a case-control when the crude incidence is known

This procedure makes use of the fact that the crude incidence can be expressed as a weighted average of exposure-specific incidences:

$$I = P_1 I_1 + P_0 I_0$$

where:

I = crude incidence

I_1 = incidence in exposed population

I_0 = incidence in unexposed population

P_1 = proportion of the population that is exposed

P_0 = proportion of the population that is unexposed

Since the RR (relative risk, CIR, IDR) = I_1/I_0, it is possible to substitute RR x I_0 for I_1 in the above expression:

$$I = P_1 \times RR \times I_0 + P_0 I_0$$

Similarly, since $P_1 + P_0 = 1$, we can substitute for $1 - P_1$ for P_0:

$$I = P_1 \times RR \times I_0 + (1 - P_1) \times I_0$$

Solving for I_0 yields:

$$I_0 = \frac{I}{P_1 \times RR + (1 - P_1)} = \frac{I}{1 + P_1 (RR - 1)}$$

Since for a rare disease we can estimate RR by using OR, the final formula is:

$$I_0 = \frac{I}{1 + P_1 (OR - 1)}$$

This formula can be used for a case-control study where:

1. The control group has been selected in such a way that the proportion exposed estimates P_1 in the population;
2. There is information available to estimate the crude incidence of the disease;
3. The disease is sufficiently rare (e.g., incidence less than 10% by the end of follow-up) so that the OR estimates the RR fairly well.

Once we have estimated I_0, we estimate I_1 by multiplying by the OR. From I_1 and I_0 we can estimate attributable risk.

Demonstration that OR estimates CIR when CI's are small

For this demonstration and the following one, a different notation will simplify the presentation. We will use D and E for disease and exposure, so that we can upper case for presence and lower case for absence. Thus, subscript E will refer to the presence of exposure, subscript e will refer to the absence of exposure. Similarly, subscript D refers to cases, subscript d to noncases. P stands for probability, which can also be interpreted as prevalence (when applied to exposure) or risk (when applied to disease). The vertical bar means "given" or "conditional on". Thus:

P_E = Probability of being exposed (i.e., prevalence of exposure)

P_e = Probability of being unexposed (1 - P_E)

P_D = Probability of disease (risk)

P_d = Probability of nondisease (1 - P_D)

$P_{E|D}$ = Probability of exposure conditional on disease (i.e., prevalence of exposure in cases)

$P_{e|D}$ = Probability of nonexposure conditional on disease (i.e., 1 - $P_{E|D}$)

$P_{E|d}$ = Probability of exposure conditional on non-disease (i.e., prevalence of exposure in noncases)

$P_{e|d}$ = Probability of nonexposure conditional on non-disease (i.e., 1 - $P_{E|d}$)

$P_{D|E}$ = Probability of disease (risk) conditional on exposure (i.e., risk of disease in the exposed)

$P_{d|E}$ = Probability of nondisease conditional on exposure (i.e., 1 - $P_{D|E}$)

$P_{D|e}$ = Probability of disease conditional on non-exposure (i.e., risk of disease in the unexposed)

$P_{d|e}$ = Probability of nondisease conditional on non-exposure (i.e., 1 - $P_{D|e}$)

By definition, odds_r = risk / (1 - risk). The risk odds ratio (OR_r) is the ratio of odds for exposed persons to odds for unexposed persons. In probability notation:

$$OR_r = \frac{P_{D|E}/(1 - P_{D|E})}{P_{D|e}/(1 - P_{D|e})} = \frac{P_{D|E}}{P_{D|e}} \times \frac{(1 - P_{D|e})}{(1 - P_{D|E})} = RR \times \frac{(1 - P_{D|e})}{(1 - P_{D|E})}$$

since $P_{D|E} / P_{D|e} = RR$. When a disease is rare, $P_{D|E}$ and $P_{D|e}$ are both small, so $OR_r \approx RR$. Using CI to estimate risk and CIR to estimate RR, we have $OR_r \approx CIR$ when CI is small in both exposed and unexposed groups. To illustrate, make up some 2 x 2 tables to reflect various disease risks and CIR's, and compute the OR's. You can verify that the OR is always farther from 1.0 than the CIR but when the incidence is below about 10%, the OR deviates little from the CIR. The OR can also be expressed as $OR_r = CIR + (OR_r -)P_{D|E}$ (when OR > 1; see Hogue, Gaylor, and Schultz, 1983), which demonstrates that the absolute difference between OR and CIR is related to the size of the OR and the disease risk in exposed persons.

Estimating PARP from a case-control study without additional information

In a case control study where the OR estimates the CIR, we can use the OR and the proportion of exposure in cases to estimate the PARP.

Begin with the formula presented earlier:

$$PARP = \frac{P_1 (RR - 1)}{1 + P_1 (RR - 1)}$$

Translating to the new notation (P_E for P_1):

$$PARP = \frac{P_E (RR - 1)}{1 + P_E (RR - 1)}$$

Removing the parentheses in the denominator and substituting ($P_{D|E}/P_{D|e}$) for RR in the denominator (only):

$$PARP = \frac{P_E (RR - 1)}{1 + P_E (RR - 1)} = \frac{P_E (RR - 1)}{1 + P_E RR - P_E} = \frac{P_E (RR - 1)}{1 + P_E (P_{D|E} / P_{D|e}) - P_E}$$

Multiplying numerator and denominator by $P_{D|e}$:

$$PARP = \frac{P_{D|e}\ P_E\ (RR - 1)}{P_{D|e}\ (1 + P_E\ (P_{D|E} / P_{D|e}) - P_E)} = \frac{P_{D|e}\ P_E\ (RR - 1)}{P_{D|e} + P_E\ P_{D|E} - P_{D|e}\ P_E}$$

Substituting $(1-P_e)$ for the second P_E in the denominator (only) and re-arranging terms:

$$PARP = \frac{P_{D|e}\ P_E\ (RR - 1)}{P_{D|e} + P_E\ P_{D|E} - P_{D|e}\ (1-P_e)} = \frac{P_{D|e}\ P_E\ (RR - 1)}{P_{D|e} + P_E\ P_{D|E} - P_{D|e} + P_e\ P_{D|e}}$$

Removing $+ P_{D|e}$ and $- P_{D|e}$ from the denominator, we then have $P_E P_{D|E} + P_e P_{D|e}$.

Since this is a weighted average of risk in the exposed (weighted by the probability of exposure) plus risk in the unexposed (weighted by the probability of nonexposure), the denominator simplifies to P_D (this was presented a couple of pages earlier as: $I = P_1 I_1 + P_0 I_0$). Therefore we have:

$$PARP = \frac{P_{D|e}\ P_E\ (RR - 1)}{P_E\ P_{D|E} + P_e\ P_{D|e}} = \frac{P_{D|e}\ P_E\ (RR - 1)}{P_D} = \frac{P_E\ P_{D|e}\ (RR - 1)}{P_D}$$

From Bayes Theorem it can be shown that:

$$\frac{P_E}{P_D} = \frac{P_{E|D}}{P_{D|E}}$$

so that the preceding formula can be written:

$$PARP = \frac{P_{E|D}\ P_{D|e}\ (RR-1)}{P_{D|E}} = \frac{P_{E|D}\ (RR-1)}{P_{D|E} / P_{D|e}} = \frac{P_{E|D}\ (RR-1)}{RR}$$

which can also be written as $P_{E|D}$ ARP.

In a case-control study, we know $P_{E|D}$, the probability (proportion) of exposure among cases and use the OR to estimate the RR (assuming that the disease is rare).

Bibliography

Textbook chapters (see preceding listing under Measuring Disease).

Davies, Huw Talfryn Oakley; Iain Kinloch Crombie, Manouche Tavakoli. When can odds ratios mislead? BMJ 1998;316:989-991.

Deubner, David C., Herman A. Tyroler, John C. Cassel, Curtis G. Hames, and Caroline Becker. Attributable risk, population attribution risk, and population attributable fraction of death associated with hypertension in a biracial population. *Circulation* 1975;52:901-908

Freeman, Jonathan; George B. Hutchison. Duration of disease, duration indicators, and estimation of the risk ratio. *Am J Epidemiol* 1986; 124:134-49. (Advanced)

Gladen, Beth C. On graphing rate ratios. *Am J Epidemiol* 1983; 118:905-908.

Greenland, Sander. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *Am J Public Health* 1999 (August)

Greenland, Sander. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987; 125:761-768 and correspondence in *Am J Epidemiol* 1988; 128:1181-1184.

Greenland, Sander; James M. Robins. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988; 128:1185-1197. (Intermediate)

Greenland, Sander; Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol* 1986; 123:203-208. (Advanced)

Hebert, James R.; Donald R. Miller. Plotting and discussion of rate ratios and relative risk estimates. Letter. *J Clin Epidemiol* 1989; 42(3); 289-290.

Hogue, Carol J.R.; David W. Gaylor, and Kenneth F. Schulz. Estimators of relative risk for case-control studies. *Am J Epidemiol* 1983;118:396-407.

Koopman JS, Longini IM, Jacquez JA, Simon CP, et al. Assessing risk factors for transmission of infection. *Am J Epidemiol* 1991 (June); 133(12):1199-1209.

Lee, James. Odds ratio or relative risk for cross-sectional data? *Intl J Epidemiol* 1994;23:201-203.

Lee, James; KS Chia. Estimation of prevalence rate ratios for cross sectional data an example in occupational epidemiology. *Br J Ind Med* 1993;50:861-2.

Thompson, Mary Lou; J.E. Myers, D. Kriebel. Prevalence odds ratio or prevalence ratio in the analysis of cross-sectional data: what is to be done. *Occup Environ Med* 1998;55:272-277.

Walter, Stephen D. Calculation of attributable risks from epidemiological data. *Intl J of Epidemiol* 7:175-182, 1978.

Walter, Stephen D. Choice of effect measure for epidemiological data. *J Clinical Epidemiology* 2000;53:931-939.

Zhang, Jun; Kai F. Yu. What's the relative risk: a method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998;280(19):1690-1691.

Relating risk factors to health - Assignment

1. Give a one-sentence definition, in terms that you might employ in an article for the educated but non-professional public, of:

   a. Cumulative incidence ratio

   b. Incidence density ratio

   c. Odds ratio

2. The following data come from a study conducted at Group Health Cooperative of Puget Sound (Orleans CT, Schoenbach VJ, Wagner EH, Quade D, Salmon MA, Pearson DC, et al. Self-help quit smoking interventions. *J Cons Clin Psychol* 1991:59:439-448). Smokers wanting to quit were enrolled into a self-help, quit-smoking trial and were randomized into one of four groups (M=quitting manual, MS=M plus social support brochure, MST=MS + telephone counseling calls, and C[control]=annotated guide to existing quitting resources). Interventions were then mailed to participants, and abstinence from tobacco use (not even a puff for the past 7 days and no use of other tobacco in the past month) was measured by mailed questionnaire and/or telephone interview at approximately 8-, 16-, and 24-months after recruitment. The 16-month follow-up obtained smoking status information for 1877 participants; salivary cotinine was measured on a geographically-selected sample of self-reported abstainers.

```
GHFCA3C:  Free & Clear (Schoenbach/Orleans) - Live data
Quit rates (SFQUIT7) by randomization group(s)
2nd follow-up respondents           16:17 Wednesday, July 26, 1989
All subjects

TABLE OF SFQUIT7 BY RGP
SFQUIT7(Quit 7 days at FU2)     RGP(Randomization Group)
```

Frequency / Percent / Row Pct / Col Pct	C	M	MS	MST	Total
0:Quit	84	71	67	109	331
	4.48	3.78	3.57	5.81	17.63
	25.38	21.45	20.24	32.93	
	18.06	15.20	14.23	23.00	
1:Smoking	381	396	404	365	1546
	20.30	21.10	21.52	19.45	82.37
	24.64	25.61	26.13	23.61	
	81.94	84.80	85.77	77.00	
Total	465	467	471	474	1877
	24.77	24.88	25.09	25.25	100.00

a. Quit rates were measured as the proportion abstinent at the time of follow-up. What was the overall quit rate for the 1877 smokers?

b. Is this "quit rate" a cumulative incidence-type measure or an incidence density-type measure? Briefly explain the basis for your answer.

c. Give one or more reasons for which type of incidence measure (i.e., a cumulative incidence type or an incidence density type) is preferable given the study design.

d. Briefly describe the 16-month results of the study.

e. The MS and MST conditions received identical interventions except that the MST condition included the availability of a toll-free telephone "Quitline" and four counselor-initiated telephone calls during the first year of follow-up. Compare the quit rates for the MS and MST groups, and compute a CIR and an OR. Compare your calculations of the CIR and OR and briefly indicate the reason for the difference in them and which measure is preferred.

f. Compute and interpret an appropriate measure of impact of the telephone component.

3. Hepatocellular adenoma (HCA), a rare type of benign though potentially fatal liver tumor, is associated with long term oral contraceptive (OC) use, especially in older women. A case-comparison study showed that the effect of duration of OC use on the risk of developing HCA is marked:

Duration	Rate ratio
1 or less	1*
4 years	9
4-7 years	120
8+ years	500

* Reference level (includes none)

(Source: Armed Forces Institute of Pathology and Center for Disease Control. Increased risk of hepatocellular adenoma in women with long term use of oral contraceptive. *Morbidity and Mortality Weekly Report* 26 (36):293-294, September 9, 1977, cited in *Oral Contraceptives*, Population Reports Series A, Number 5, January 1979.)

Assuming that the incidence density (ID) of HCA for one year or less use of OC is 0.06/100,000 per year (i.e., 6 per 10,000,000 women-years), what are the attributable rate (rate difference) over baseline and the attributable rate proportion associated with each duration category of OC use? Interpret these measures and state what implications you might draw. (For this question, use the attributable risk formulas from the chapter even though the data are for rates.)

4. In a study of adverse effects of radiotherapy among immigrant children in Israel (Ron E, Modan B, and Boice JD. Mortality after radiotherapy for ringworm of the scalp. *Am J Epidemiol* 1988:127:713-25), 10,834 irradiated children were identified from original treatment records and matched to 10,834 nonirradiated, tinea-free comparison subjects selected from the general population. Follow-up was accomplished using the Israeli Central Population Registry, which enabled nearly all subjects to be followed forward in time (retrospectively) for a mean of 26 years following age at irradiation. Computation of person-years of observation began at the date of treatment for tinea capitis, or the equivalent date for the matched comparison subjects, and ended at the date of death for those who died or May 31, 1982 for those not known to have died. Person-years of observation were: irradiated subjects, 279,901 years; comparison subjects, 280,561 years. During the follow-up there were 49 deaths from cancer in irradiated subjects, and 44 in the nonirradiated population comparison subjects (data from table 3 in Ron et al.). (For these questions, use the attributable risk formulas from the chapter even though the data are for rates.)

   a. What are the rates of cancer death in the two groups?

   b. Calculate and describe in one sentence the incidence density ratio for cancer death comparing irradiated and nonirradiated subjects?

   c. Assuming causality, estimate how many cancer deaths per 100,000 person years of follow-up of irradiated subjects were attributable to radiotherapy.

   d. Again assuming causality, what proportion of cancer deaths in irradiated subjects were due to radiation therapy?

   e. If 10% of this population had received radiotherapy for tinea capitis, what proportion of all cancer deaths within the relevant age span (mean age 7 to 33 years) would be due to radiation therapy?

5. Algebraic calisthenics: There are various formulas for the population attributable risk proportion (PARP), including several given in the lecture handout. Demonstrate the algebraic equivalence of the PARP formulas in the text, i.e., derive each of the subsequent formulas from the one derived from the attributable risk diagram:

$$\frac{\text{"Attributable cases"}}{\text{All cases}} = \frac{(ID_1 - ID_0)n_1}{I_1n_0 + I_1n_1}$$

a. $\dfrac{I_{crude} - I_0}{I_{crude}}$ b. $\dfrac{p_1(RR - 1)}{1 + p_1(RR - 1)}$ c. $\dfrac{1}{1 + 1/[\, p_1(RR-1)]}$

Relating risk factors to health - Assignment solutions

1. Definitions:

   a. Cumulative incidence ratio (CIR) - a measure of association equal to the ratio of two cumulative incidences, or the proportion of one group who experience an event relative to the proportion of another group who experience the event.

   b. Incidence density ratio (IDR) - a measure of association equal to the ratio of two incidence densities, or the magnitude of the incidence rate in one group relative to the incidence rate in another.

   c. Odds ratio - a measure of association based on the odds of disease or exposure in two groups; the odds ratio often estimates or approximates the IDR and/or CIR

2. a. The overall quit rate was 17.6% (331/1887).

   b. The quit rate is the proportion of abstainers among participants who provided data at their 16-months follow-up. In this sense, the quit rate is the prevalence of abstinence at a point in time, with time being expressed relative study enrollment. In fact, the smoking cessation literature sometimes refers to this type of quit rate as "abstinence prevalence". Since all participants were smokers at baseline, the quit rate can also be regarded as the cumulative incidence for becoming a nonsmoker during 16 months of follow-up. The problem with using cumulative incidence to measure quitting smoking is that abstinence is a reversible state, so the "cases" (quitters) in this study may shortly thereafter revert to "noncases" (smokers). The proportion of participants who quit for 24 hours at some time during the 16-months of follow-up is more clearly a cumulative incidence, but it does not quite tell us what we want to know.

   c. Although quitting smoking is an "event", or at least a change of status, it is difficult to translate into a conventional incidence measure. It would be possible to compute an incidence rate based on the number of quits divided by the number of participant-months of follow-up. However, such an incidence rate has no useful interpretation, since a low incidence rate could mean few participants quit or that participants quit and stayed quit. A high rate could mean that many participants quit or that participants kept quitting and relapsing.

      Although it's difficult to know when permanent nonsmoking status has been achieved, the longer the period of continuous abstinence the greater the probability of remaining smoke-free. Since quitting smoking for good has an "extended risk period", an incidence rate of

number of "permanent" quits (defined on the basis of some duration of nonsmoking) per participant-year of observation might be appropriate for measuring the effect of a continuing quit-smoking intervention (e.g., a prolonged advertising campaign). For the most part, though, experimental interventions take place over a fairly short period of time, and their effect is assumed to take place either during the intervention or shortly afterward, a situation argues for a cumulative incidence quit rate during the expected period of effect. Given the conceptual complexities as well as the limitations of biochemical verification, continuous abstinence from the completion of an intervention and abstinence prevalence appear to be the measures most commonly used.

d. Quit rates ranged from 14.2% to 23.0%, with the highest rate in the MST group and the lowest rates in the M and MS groups. The control group had an intermediate quit rate. Although the differences in absolute quit rates were modest, the MST group was clearly the highest. On the assumption that the control group received the least intervention, it is surprising that its quit rate appeared to be higher than the two mail-only intervention groups. (Indeed, one can speculate whether the quitting manual and/or social support brochures by themselves actually depressed quitting below what would have happened; conversely, the controls may have been more likely to obtain quitting assistance from other sources. (Note: the quit rates can be read from the bottom line (Col Pct) of the upper row or computed by dividing the number of quitters in each condition by the total for that condition.)

e. The CIR for quitting for MST vs. MS groups is 0.230/0.142 = 1.62; i.e., the MST group quit rate was 60% higher than or 1.6 times the rate for the MS group. The OR for quitting for MST vs. MS groups is (0.230/0.770 divided by 0.142/0.858) = (.230*.858)/(.142*.770) = 1.8. As always, the OR is farther from 1.0 than the CIR The OR approximates the CIR when the outcome is rare, which is not quite the case here (quit rates of 14%-23%). However, when the CIR is not far from 1.0, as is the case here, the OR will be only modestly larger.

f. The "attributable risk" (quit rate difference) is 23.0% - 14.2% = 8.8% (absolute). As a percentage of the quit rate in the "exposed" (ARP or EF_1), the impact of the telephone component would be AR/I_1 or 8.8/23.0 = 38%. Thus, the telephone component appeared to account for nine percentage points or 38% of the quit rate in the MST group.

3. (For this question, we are ignoring the distinction between rates and risks)

Duration	Relative risk	Attributable risk**	Attributable risk proportion
1 year or less	1*		
4 years	9	(9-1) (0.06) = 0.48	(9-1)/9 = 0.89
4-7 years	120	(120-1) (0.06) = 7.1	(120-1)/120 = 0.99
8+ years	500	(500-1) (0.06) = 30.0	(500-1)/500 = ~1.00

* Reference level (includes none) ** per 100,000

There is an extremely strong association between OC use (4 years or longer) and hepatocellular adenoma, and the attributable risk proportion is nearly 1.0 for OC use above 4 years. The excess risk incurred by OC users, however, is miniscule at 4 years of OC use, and quite modest until 8 or more years. The implication is that the association is likely to be causal (due to the strength of the ratio measure of effect) but the resulting increase in risk does not become important until more than 4 years of OC use.

4. a. Irradiated subjects: ID = 49/279,901 person-years

= 0.000175, or 17.5 cancer deaths per 105 person-years

Comparison subjects: ID = 44/280,561 person-years

= 0.000157, or 15.7 cancer deaths per 105 person-years

b. IDR = ID_1/ID_0 = (17.5 per 105 person-years) / (15.7 per 105 person-years) = 1.1. The rate of cancer deaths in the exposed population is 1.1 times that in the non-exposed comparisons.

c. Rate difference = $ID_1 - ID_0$ = 17.5 - 15.7 = 1.8 cancer deaths per 100,000 person-years.

d. Rate fraction = $(ID_1 - ID_0)/ ID_1$ = (17.5 - 15.7)/17.5 = 0.10

or 10% of the cancer deaths in the exposed group are due to radiotherapy.

e. Population attributable risk proportion = $p_1(IDR - 1) / [1 + p_1(IDR - 1)]$

= $0.10(1.11 - 1) / [1 + 0.10(1.11 - 1)]$ = 1.08% of cancer deaths (ignoring the distinction between risk and rate)

5.

a. Begin with: $$\frac{\text{"Attributable cases"}}{\text{All cases}} = \frac{(I_1 - I_0)n_1}{I_1n_0 + I_1n_1}$$

Then remove the parenthesis in the numerator, add and subtract I_0n_0, and rearrange terms:

$$= \frac{I_1 n_1 - I_0 n_1 + I_0 n_0 - I_0 n_0}{I_0 n_0 + I_1 n_1} = \frac{I_1 n_1 + I_0 n_0 - I_0 n_1 - I_0 n_0}{I_0 n_0 + I_1 n_1}$$

The crude rate (I) is a weighted average of the stratum-specific rates (I_1, I_0), weighted by the proportions in each stratum, so $I_1 n_1 + I_0 n_0 = I\ n$. Divide numerator & denominator by $n(=n_1 + n_0)$:

$$= \frac{(I\ n) - I_0\ (n_1 + n_0)}{(I \times n)} = \frac{I - I_0}{I}$$

b. Begin again with: $$\frac{\text{"Attributable cases"}}{\text{All cases}} = \frac{(I_1 - I_0) n_1}{I_1 n_0 + I_1 n_1}$$

(i) Add and subtract $I_0 n_1$ in the denominator,

(ii) rearrange numerator and denominator, and

(iii) divide by $n\ I_0$, recalling that $n=(n_1 + n_0)$, $RR= I_1 / I_0$, and $p_1 = n_1 / n$:

$$\frac{I}{\frac{(I_1 - I_0)\ n_1}{I_0 n_0 + I_1\ n_1 - I_0\ n_1 + I_0\ n_1}} \overset{ii}{=} \frac{n_1 (I_1 - I_0)}{(n_0 + n_1)\ I_0 + n_1\ (I_1 - I_0)} \overset{iii}{=} \frac{p_1 (RR - 1)}{1 + p_1 (RR - 1)}$$

c. The formula: $$\frac{1}{1 + 1/[\ p_1 (RR - 1)]}$$ is obtained by dividing numerator and denominator in the preceding formula by the numerator, $p_1(RR - 1)$.

8. Analytic study designs

The architecture of the various strategies for testing hypotheses through epidemiologic studies, a comparison of their relative strengths and weaknesses, and an in-depth investigation of major designs.

Epidemiologic study designs

In previous topics we investigated issues in defining disease and other health-related outcomes, in quantitating disease occurrence in populations, in relating disease rates to factors of interest, and in exploring and monitoring disease rates and relationships in populations. We have referred to cohort studies, cross-sectional, and case-control studies as the sources of the measures we examined, but the study designs themselves were secondary to our interest. In the present chapter we will define and compare various study designs and their usefulness for investigating relationships between an outcome and an exposure or study factor. We will then examine two designs – intervention trials and case-control studies – in greater depth.

The study designs discussed in this chapter are called analytic because they are generally (not always) employed to test one or more specific hypotheses, typically whether an exposure is a risk factor for a disease or an intervention is effective in preventing or curing disease (or any other occurrence or condition of interest). Of course, data obtained in an analytic study can also be explored in a descriptive mode, and data obtained in a descriptive study can be analyzed to test hypotheses. Thus, the distinction between "descriptive" and "analytic" studies is one of intent, objective, and approach, rather than one of design. Moreover, the usefulness of the distinction is being eroded by a broad consensus (dogma?) in favor of testing hypotheses. Since to characterize a study as "only descriptive" tends to devalue it, investigators understandably try to portray their studies as "analytic" and "hypothesis-driven" in order to make a better impression and to improve their chances for funding and journal space. (Opinions expressed herein are not necessarily those of the sponsor!)

Whether the study is "descriptive" or "analytic", it is important to clearly identify the objectives of the study (preferably identifying the specific parameters to be measured – see Rothman and Greenland) and the rationale (i.e., the case for conducting the research). There are innumerable decisions, judgments, and compromises that must be made during the design, conduct, analysis, and interpretation of a study, and the principal guideposts for making them are the study objectives and rationale. For example, if the objective is to test hypotheses, then the investigator designs and conducts the study so as to maximize the usefulness of the data for testing these hypotheses. Failure to keep the study objectives prominently in one's mind increases the advantage of hindsight over foresight.

Epidemiologic investigations of disease etiology encounter many challenges, especially when they must contend with one or more of the following:

1. Difficulties in defining and measuring the disease;

2. Imprecision in determining the time of onset of the disease;
3. Prolonged intervals between exposure to a causative agent and disease onset (induction period) and between disease onset and detection (latency);
4. Multifactorial disease etiology; and
5. Differential effect of factors of interest on incidence and course of the disease.

[See Mausner and Kramer, Chapter 7, pp 178 *et seq*.]

Even more daunting can be studies of phenomena other than clinical diseases, where less assistance is available from the biomedical armamentarium.

In view of these and other challenges, including the logistical and practical ones of obtaining access to subjects, measuring variables of interest, protecting subjects' rights, and assembling sufficient cases for rare diseases, the basic epidemiologic analytic strategy may be characterized as "by any (ethical) means necessary", along with "try to get the best but if you have to, make do with what's available". For this reason there are innumerable variations in the details of study design. But in terms of the basic architecture - how the principal components of a study are assembled - there are certain basic designs.

Traditional classification of epidemiologic study designs

A logical sequence of study designs encountered in epidemiology is:

1. Case reports
2. Case series
3. Ecologic (also called correlational)
4. Cross-sectional
5. Case-control
6. Follow-up/cohort
7. Intervention trials/controlled trials

The first two of these designs are employed in clinical, rather than epidemiologic, studies, but often are precursors to epidemiologic studies. The next two designs are regarded as primarily descriptive, the last design is primarily analytic, and designs 5 and 6 can be employed in analytic (hypothesis testing) or descriptive modes, depending upon the extent to which the study is oriented towards a pre-existing specific hypothesis. Of course, it may be difficult to obtain resources for a lengthy or expensive study without *a priori* hypotheses, but there are exceptions. Of course, once the data have been collected for whatever purpose, they will often be subject to a search ("search and destroy", as some would have it; "seek and ye shall find" in the view of others) for other associations and insights.

Progression of types of studies

In the classic or ideal scenario, studies of disease etiology unfold from simple, inexpensive, and rapid investigations that identify hypotheses to complex, costly, and lengthy ones to evaluate these hypotheses. General, exploratory studies typically take place before highly focused studies.

New syndrome or outbreak

The stimulus to investigating disease etiology may be prompted by the appearance of a new or previously unrecognized syndrome. In this case the initial efforts will be aimed at characterizing the syndrome, developing a case definition, and searching for characteristics that differentiate people with the disease from persons without the disease. Or, a previously recognized disease may occur in a population group or geographical area where it has not been thought to occur. Such nonroutine situations then prompt a case report in a medical journal, notification of public health officials, or other actions that lead to initial studies – typically case series's and outbreak investigations – to define the nature of the situation and to look for leads to its cause.

As we recounted in an earlier chapter, the history of AIDS epidemiology followed this classic pattern. Recognition of AIDS began with case reports and case series's describing cases of young otherwise healthy men in California and New York City with *Pneumocystis carinii* pneumonia (PCP) and Kaposi's Sarcoma (*MMWR* 1981;30:250-2 and 305-8). Before that time, PCP had been seen only in persons who had been medically immunosuppressed in connection with a transplant operation. Kaposi's Sarcoma had been known as a disease of Africans and elderly men of Mediterranean origin. The initial case series's described common and variable features of the syndrome. For example, all of the patients were men who had sex with men, most had a large number of male sex partners, and many used inhalants, a type of recreational drugs.

The case series's led to an initial AIDS case definition for the purposes of identifying additional cases and inaugurating surveillance. With a case definition in hand, it was also possible to conduct case-control studies in which persons with the disease could be compared with persons without the disease and characteristics associated with the condition identified. Comparisons of AIDS cases to apparently healthy male homosexual controls indicated that the cases had higher numbers of partners, had greater involvement in certain sexual practices (anal intercourse, fisting), and more exposure to drugs used to enhance sexual pleasure. These findings led to analytic studies to test these and other exposure hypotheses.

Case reports and case series's are the clinical route to definition and recognition of disease entities and to the formulation of hypotheses. These studies are not "epidemiologic" in the sense that they have no explicit comparison group or population reference. On the other hand, one can think of an implicit comparison with "common knowledge", "general experience", etc., when the characteristics of cases are striking. An example is history of maternal exposure to diethylstilbesterol (DES) in teenage women with vaginal adenocarcinoma. Other diseases where the clinical route to hypothesis development was prominent are dental caries and fluoride, congenital malformations later linked to maternal rubella infection and retrolental fibroplasia in premature newborns later linked to oxygen exposure.

Sometimes the appearance of a new syndrome is sufficiently alarming that public health authorities are notified and involved at the outset. For example, toxic shock syndrome, with its rapid and malignant clinical course, Legionnaire's disease, where a group of conventioneers became severely ill within hours of one another, and rapidly fatal Hanta virus infections among American Indians living in the Southwestern United States in 1994 prompted investigations by public health authorities thereby prompting a much more intensive investigation of microbiologic and environmental factors.

Descriptive studies and surveillance

An alternate stimulus to investigation may come from a surveillance activity or descriptive study. The descriptive study might be a re-analysis of data collected for some other purpose (e.g., from a national population survey or possibly from an analytic study of another hypothesis or even another disease), a mapping study in which disease rates are plotted geographically, or an "ecological" study that uses data on populations rather than on individuals. For example, Warren Winklestein's observation that in the Third National Cancer Survey (US) geographical areas with high rates for cervical cancer tended to have high rates for lung cancer led him to the hypothesis that cigarette smoking might be a risk factor for cervical cancer.

Observations made from population-level data require additional caution in their interpretation, however. For example, colon cancer rates are higher in U.S. counties that use mostly surface water and in countries with high per capita meat consumption. These relationships suggest that something about surface water, e.g., chlorination, and something about meat consumption, e.g., saturated fat intake, might be factors in the development of colon cancer. However, since exposure is not known at the individual level, it is possible that the cases of colon cancer are not themselves people who drink chlorinated water or eat meat. The attempt to infer individual characteristics or relationships from group-level measures is called the "ecologic fallacy". Ecologic, or group-level, studies can nevertheless contribute important information, though, and not only in an exploratory mode.

Once the hypothesis has been advanced, analytic studies are the next epidemiologic recourse. The progression of designs at this point depends on the nature of the disease and exposure - the rarity of the disease, the length of its natural history, the problems in measuring disease and exposure, and other factors. For many diseases, especially rare ones, the usual sequence is to begin with case-control studies (since these are generally the most efficient and logistically practical design) and, unless negative results occur and are accepted, move towards follow-up studies and possibly intervention studies.

Individual-level studies

Although an "epidemiologic transition" appears to be underway, most analytic studies have the person as the unit of data collection and analysis. Thus, the four classic analytic study designs are generally thought of in relation to individual-level studies, though as we shall see they can also be employed for studies where the group is the unit of analysis. These four primary designs are:

Cross-sectional

A cross-sectional study is one in which subjects are sampled without respect to disease status and are studied at a particular point in time, as in a random-sample health survey. The term "cross-sectional study" (or "prevalence study") usually refers to studies at the individual level, even though ecologic studies are typically (though not necessarily) cross-sectional, also. The target population is generally one whose identity is of some wider interest (e.g., a political or geographical entity, a profession or workforce, or a major organization (union, HMO, student body), but may not necessarily be so.

In a cross-sectional study, the current or historical status of individuals is assessed and may be examined in relation to some current or past exposure. These studies are obviously most useful for conditions that are not rapidly fatal, not terribly rare, and/or not routinely brought to medical attention (e.g., elevated blood pressure, elevated blood cholesterol, many psychiatric disorders, diet, subclinical infection, and serologic markers of previous infections).

Since participants for a cross-sectional study are generally chosen without previous knowledge of their disease or exposure status, such studies can be used to estimate prevalence of both diseases and exposures and therefore to compute prevalence ratios and prevalence odds ratios.

Among the more widely known cross-sectional studies are the periodic national household (interview) surveys by the U.S. National Center for Health Statistics (NCHS), the annual (telephone) Behavioral Risk Factor Survey conducted by the U.S. Centers for Disease Control and Prevention (CDC), and HIV seroprevalence studies. Sometimes the process of recruiting subjects to a follow-up study (e.g., the Lipids Research Clinics Coronary Primary Prevention Trial prevalence study) serves as a cross-sectional study. The cross-sectional NCHS NHANES (National Health and Nutrition Examination Survey) study became a follow-up study when respondents were re-examined ten years later, creating the NHANES Follow-up Study.

Strengths

- Can study entire populations or a representative sample.
- Provide estimates of prevalence of all factors measured.
- Greater generalizability.

Weaknesses

- Susceptible to selection bias (e.g. selective survival)
- Susceptible to misclassification (e.g. recall)
- Information on all factors is collected simultaneously, so it can be difficult to establish a putative "cause' antedated the "effect'.
- Not good for rare diseases or rare exposures

Case-control (case-referent, etc.) studies

A case-control study is one in which persons with a condition ("cases") are identified, suitable comparison subjects ("controls") are identified, and the two groups are compared with respect to prior exposure. Thus, subjects are sampled by disease status. Case-control studies are used in infectious disease epidemiology, but they have become the primary strategy in chronic disease epidemiology. The investigation and refinement of the case-control design, a process which began in about the middle of the 20th century (see classic articles by Cornfield, 1951 and Mantel and Haenszel in 1959) constitutes a significant innovation in population-based research. (Note: The analogy that presumably led case-control theorists to adopt the term "control" from experimental designs is accurate only in a general sense, i.e., in both cases the control group serves as a point of reference of comparison for the group of primary concern. However, because of the fundamentally different architecture of experimental and case-control designs, the analogy ends there and has probably been a source of confusion in earlier writings about the case-control design. See the end of the section on selection bias in the next chapter.)

Because subjects are identified after the disease has developed, and inquiry then investigates prior exposure, the case-control study is sometimes referred to as a "retrospective" or "backwards" design. The "backwards" design poses greater demands in terms of methodological and analytic sophistication. However, by ensuring a greater balance between the numbers of cases and noncases, the case-control design generally offers much greater statistical efficiency than other designs, giving it a crucial advantage for studying rare diseases.

Case-control studies can use prevalent cases (i.e., existing at the time the study begins) or incident cases (i.e., newly diagnosed during the period of the study). In the former instance, the distinction between a case-control study and a cross-sectional study can become very blurred. In addition, data collected through other kinds of studies can be analyzed as if data had come from a case-control study, thereby providing another source of confusion.

Because case-control studies select participants on the basis of whether or not they have the disease, the case-control design does not provide an estimate of incidence or prevalence of the disease, unless data about the population size are available. But as long as the participants are chosen without regard to their exposures, the study can estimate the prevalence of one or more exposures. With these prevalences, in turn, we can estimate an exposure odds ratio which we then use to estimate the IDR or CIR in the base population.

Strengths

- Good for rare diseases
- Efficient in resources and time

Weaknesses

- Susceptible to selection bias (e.g., cases or controls may not be appropriately "representative")
- Susceptible to misclassification bias (e.g. selective recall)
- May be difficult to establish that "cause" preceded "effect".

Follow-up studies

Along with case-control studies, follow-up studies constitute the other basic observational strategy for testing hypotheses. In a follow-up study, people without the disease are followed up to see who develops it, and disease incidence in persons with a characteristic is compared with incidence in persons without the characteristic. If the population followed is a defined group of people (a "cohort"), then the study is referred to as a cohort study. Alternatively, the population under study may be dynamic (e.g., the population of a geographical region).

Follow-up studies may be done "retrospectively", where the population at risk can be defined at some time in the past and traced forward in time, or "prospectively", where the population is identified or assembled by the investigator and then followed forward in time.

Since the study population for a follow-up study is selected from among people who are free of the disease, this study design can estimate incidence based on new cases that develop during the follow-up period. Be cause the investigator can estimate incidence separately for exposed and unexposed participants, the IDR and/or CIR can be directly obtained from the incidence estimates. In some cases, the study population is gathered on the basis of an initial cross-sectional study (e.g., the Framingham and Evans County cohorts). In such cases, exposure prevalences in the base population can also be directly estimated, though this ability comes from the cross-sectional component, not from the follow-up component.

Strengths

- Better for rare exposures
- Less confusion over relative timing of exposure and disease than with other observational designs.

Weaknesses

- Costly and time consuming if disease is rare and/or slow to develop.
- Loss to follow-up (attrition) may lead to selection bias.
- Relatively statistically inefficient unless disease is common.

Intervention trials (controlled trials)

An intervention trial is a follow-up study in which the primary exposure under study is applied by the investigator. These are the only experimental form of epidemiologic studies, though they are also observational in that subjects remain in their ordinary habitats. In an intervention trial, the investigator decides which subjects are to be "exposed" and which are not (in contrast to naturalistic studies in which the subjects "choose" their exposure group by "deciding" whether to smoke, drink, exercise, work in a hazardous environment, be exposed to toxic wastes, breathe polluted air, develop elevated blood pressure, develop diabetes, etc.).

The term "clinical trial" emphasizes the controlled aspect of the intervention, at the expense of the generalizability of the results; the term "community trial" emphasizes that the trial is carried out in a

realistic setting and results may therefore be more generalizable (at the expense of having control over what subjects actually do). A community trial can involve an individual-level intervention (e.g., breast cancer screening), a community-level intervention (e.g., gun control), or interventions with elements of both levels (e.g., mass media promotion of physical exercise).

In the United States, the National Heart Lung and Blood Institute (NHLBI) sponsored and led several major (thousands of subjects, multiple expensive follow-up examinations, many millions of dollars) individual-level randomized intervention trials to confirm the value of modifying coronary heart disease and cardiovascular disease risk factors: the Hypertension Detection and Follow-up Program (HDFP), Multiple Risk Factor Intervention Trial (MRFIT), and the Lipids Research Clinics Coronary Primary Prevention Trial (LRC CPPT). More recently, the National Cancer Institute (NCI) began large-scale trials to assess effectiveness of screening techniques for cancers at a number of sites (colon, prostate). Probably the largest individual-level randomized trial in the U.S. is the Women's Health Initiative (WHI) which is funded through the National Institutes of Health (NIH, of which both NHLBI and NCI are subdivisions). Large trials of this type have also been conducted in Australia, Canada, Europe, and probably elsewhere that I am not yet aware of.

Intermediate between a formal intervention trial and a follow-up study are follow-up studies in which the intervention is applied by an outside agency (e.g., a health care provider or organization) but is not being manipulated in response to an experimental design.

Strengths

- Most like an experiment
- Provides strongest evidence for causality in relation to temporality and control for unknown "confounders"
- Fulfills the basic assumption of statistical hypothesis tests

Weaknesses

- Expensive, time consuming, sometimes ethically questionable.
- Subjects are often a highly selected group (selected for willingness to comply with treatment regimen, level of health, etc.) and may not be representative of all people who might be put on the treatment (i.e., generalizability may suffer).

Group-level (ecologic) studies or measures

Group-level studies (also called ecologic studies, correlational studies, or aggregate studies) obtain data at the level of a group, community, or political entity (county, state, country), often by making use of routinely collected data. When they use data that are already available and usually already summarized as well, these studies can be carried out much more quickly and at much less expense than individual-level studies. Group-level studies may also be the only way to study the effects of group-level constructs, for example, laws (e.g., impact of a seatbelt law), services (availability of a suicide prevention hotline), or community functioning. Multi-level studies can include both individual-level (e.g., disease, individual exposure) and group-level (e.g., median family income) variables at the same time. The popularity of multi-level studies is growing rapidly, due to the return

of interest community-level influences and the increasing availability of statistical algorithms and software to analyze multilevel data.

Each of the four classical study designs discussed above (cross-sectional, case-control, follow-up, intervention) can also be carried out with group-level variables. Thus, a set of counties, states, or countries can be analyzed in a cross-sectional manner to look at the variation in a health variable (e.g., mean blood pressure, hospitalizations for asthma, homicide rates, imprisonment rates) and its relationship to country characteristics (e.g., salt intake, air pollution, handgun laws or possession, drug policies). Many group-level studies are of this type. (Studies of homicide rates, new hospitalizations, and other phenomena that represent events, rather than conditions, should perhaps be regarded as follow-up studies, rather than cross-sectional. When only a single year's data are being analyzed or when data for several years are combined into an annual average, the traditional perspective has been cross-sectional.)

Similarly, an investigator can assemble a set of groups (e.g., animal herds, states) with high rates of some health outcome and compare their characteristics with those of states with low rates, as in a case-control study, or can monitor aggregate populations as in a follow-up study to see if differences in baseline variables (e.g., restrictions on cigarette advertising, higher cigarette taxes) are reflected in the development of outcomes (smoking initiation by adolescents). Finally, a group-level intervention trial can be conducted in which schools, worksites, neighborhoods, or political subdivisions are assigned to receive an interventions (school health clinics, curricula, media messages, or lay health advisor programs) and outcomes are monitored over time. Among the more widely-known community intervention trials are the National Cancer Institute COMMIT trial (for smoking cessation and prevention), the Stanford Three Community and Five-City Studies (cardiovascular disease), the North Karelia Study (cardiovascular disease), and recent HIV prevention trials using mass treatment for curable sexually transmitted diseases.

One situation where ecologic data are particularly useful is that where a powerful relationship that has been established at the individual level is assessed at the ecological level in order to confirm its public health impact. If a risk factor is a major cause of a condition (in terms of population attributable fraction as well as strength of association), then a lower presence of that factor in a population should presumably be linked to a lower rate of the associated outcome. Examples of studies where this approach has been taken include studies of oral contraceptive sales and CVD in women (Valerie Beral), incidence of endometrial cancer and prescription data for replacement estrogens, and motor vehicular fatalities and occupant restraint legislation or enforcement.

Ecologic measures as surrogates for individual measures

Recent articles have clarified discussions about ecologic studies by noting that there are in fact two basically different types of group-level studies, or, equivalently, two different ways in which a study can be "ecologic" (Charles Poole, Ecologic analysis as outlook and method, 1994). In the first type, a study may be "ecologic" in that the exposure status (fat intake for individuals) is estimated from the group average (per capita fat intake). In this case the group-level variable serves as a proxy for the values for individuals. The group-level average is an inferior measure of the values of individuals, but it is often much easier and economical to obtain. In addition to the loss of precision that results from using the group average as the data for individuals, there is also the danger of the

"ecologic fallacy", the erroneous inference that specific individuals in a group share the characteristics of the group.

Of course, most individuals in a group must share the characteristics of the groups which they comprise. But groups are heterogenous, and a subgroup of individuals can easily differ greatly from the group mean. For example, data showing that areas with higher average residential radon levels had higher lung cancer rates than areas with lower levels do not logically imply that the higher lung cancer rate are due to the higher radon levels. Such an inference is based on the ecologic fallacy, because it is possible that the excess lung cancers occurred to people in houses with low radon levels. In that case the group-level average would be an invalid surrogate for individual-level measurements. But even though it is not valid to infer from these data that radon exposure contributes to the elevated lung cancer rates, that may nevertheless be a correct characterization of the phenomenon. Other data are needed to draw the inference; in the meantime, these ecologic data provide the rationale for more in-depth study.

Ecologic measures as the relevant constructs

A second way in which a study can be "ecologic" is if the population, rather than the individual, is the real unit of study. In this case, a group-level factor is itself the exposure (e.g., an anti-smoking ordinance, crime rate, population density) or, occasionally, the disease (e.g., homicide rate). Although epidemiology has a long tradition of using population-level data for descriptive purposes, the use of group-level data for hypothesis testing has been out of favor because of the problem of the ecologic fallacy (even though it applies primarily to the other type of ecologic study), major limitations in the ability to control for the effects of known determinants of the outcome under study, and the ascendency of the biomedical paradigm in conjunction with the enormous expansion in capabilities for biochemical measurement and analysis.

How one regards ecologic studies depends to a certain extent on which type of studies are being considered - studies in which group-level variables are measured as economic and convenient, but inferior, measures of diseases and exposures at the individual level or studies in which the phenomena under study operate at the level of the group, rather than (or as well as) the individual. A major modifying influence, though, is one's perspective on epidemiology and public health (see chapter "The role of epidemiology in public health"). In Charlie Poole's (*AJPH*, May 1994) formulation, epidemiologists who regard the health of a community as more than the summation of the health of its individual members, regard ecologic studies (of the second variety) as critical to conduct. In contrast, epidemiologists who regard the health of a community as the summation of the health of its members regard individual-level studies as the superior form of investigation.

Although the latter view remains the dominant one in the U.S. epidemiology profession and government funding for epidemiologic research, the former has been gaining renewed attention, as evidenced by the series of articles in the May 1994 *American Journal of Public Health* from which this section draws heavily. For Susser (who at that time was editor of *AJPH*, though not for his articles), the prime justification for the ecological approach in epidemiology is the study of health in an environmental context: pairings, families, peer groups, schools, communities, cultures – contexts that alter outcomes in ways not explicable by studies that focus solely on individuals (The logic in ecological: I. The logic of analysis. *AJPH* 1994). And where group-level constructs are involved, the

ecological approach may be the appropriate level of study (Schwartz *AJPH* 1994; Susser *AJPH* 1994 [both his articles]).

Multi-level studies

Multi-level studies provide an area of agreement in this debate, since they potentially combine the advantages of both individual- and group-level studies. By using sophisticated methods of analysis – which are only now starting to become readily available thanks to the computer revolution and the development of statistical software – investigators can create mathematical models that include both group-level and individual-level variables. In principle, then, the investigator can take advantage of the ability to control for individual variability and the measurement power and precision offered by biochemical technology while at the same time addressing social, economic, and institutional influences at the community-level.

But such advantages come at a cost. Studying the effects of a group-level variable requires data for a large enough number of groups to enable comparison among them. Routinely collected data (e.g., census data) make such studies economical and relatively easy to conduct. A multi-level study, however, requires individual-level data as well, which typically means primary data collection with its attendant costs, challenges, and time. Moreover, the individual-level data must now be obtained from individuals in a larger number of groups (e.g., worksites, counties) than might be necessary if the objective of the study focused on individual-level variables.

Types of group-level variables

Group-level variables do not all possess the same degree of "groupness". One variety of group-level variable is are summaries of individual characteristics, such as per capita income. Such a variable has been termed contextual (Mervyn Susser, The logic in ecological: I. The logic of analysis. AJPH 1994) or aggregate (Hal Morgenstern, chapter 23 in Rothman and Greenland). Such variables illustrate the distinction between individual-level and group-level perspectives, since the aggregate variable measures a different construct from its name-sake at the individual level (Schwartz AJPH 1994). Thus, per capita income may be used as a surrogate measure of individual or family socioeconomic status, in which case it is inferior to the individual-level measure, or may instead directly measure income at the community-level, in which case it is a group-level measure with implications for availability of goods, services, facilities, and opportunities of all kinds education, commercial vitality, neighborhood safety, and many other aspects of the social and institutional, and physical environment.

Variables that are not summary measures of individual-level variables include factors like climate, air pollution, disasters, and laws. Susser uses the term integral variable for a variable that does not have a corresponding individual-level value. Integral variables, according to Susser, cannot be analyzed at the individual level.

Morgenstern differentiates between environmental measures and global measures. Environmental measures are "physical characteristics of the place in which members of each group live or work (e.g., air-pollution level and hours of sunlight)" and which have individual-level analogs whose value

can vary substantially among individuals. In contrast, global measure have "no distinct analogue at the individual level . . . (e.g., population density, level of social disorganization, the existence of a specific law, or type of health-care)" (p460).

One can imagine, though, that global measures may also affect individuals differently. Thus population density affects people in different ways depending upon their occupation, preferred activities, transportation requirements, needs for services, and economic resources. Social disorganization affects people more or less depending upon their age, personal social networks, occupational affiliations, need for social services, and, of course, economic resources. Anatole France's aphorism that the law forbids both the poor and the rich alike from sleeping under a bridge or stealing a loaf of bread reminds us that the law does not affect all individuals in the same way. The individual-level effects of the type of health care system depends upon the individual's need for health services, mobility, and, of course, economic resources. Even climate presumably has weaker effects on people with good climate control in their home, workplace, and automobile and who can take extended vacations.

Dependent happenings

An important category of contextual variable is "dependent happenings", where a phenomenon propagates from one person to others. Dependent happenings arise most obviously in the case of contagious diseases, where the prevalence is both a summary of individual infection status but also greatly affects the risk of infection for exposed, nonimmune persons. As an example of the inability of individual-level analysis to analyze a situation with dependent happenings, Koopman and Longini (*AJPH* May 1994;84:836-842) present a study of dengue fever in Mexican villages. The study, carried out following a multi-year epidemic, examined the association between history of infection (measured by antibody test) and presence of *Aedes aegypti* larvae in a household. The odds ratio for an analysis at the individual level was 1.1, i.e., presence of larvae was not related to a positive antibody test. By contrast, the ecological (village-level) analysis yielded an OR of 12.7.

The authors' explanation for this difference is that transmission (i.e., dependent happenings) decreases individual-level effects and increases ecological effects. With a sufficient number of infected persons in a village, the mosquitoes carry the infection to others in that village, even those whose household has not been a breeding ground for mosquitoes. In a village with few infected persons, the mosquitoes are less likely to acquire the virus so households with larvae are not in fact at elevated risk. In this scenario, higher infection prevalence in a village contributes to the ecological relationship directly (because infection prevalence is the outcome variable) and indirectly (in that mosquitoes in high prevalence villages are more likely to get infected).

Other phenomena and situations can also obscure effects of risk factors for transmission in individual-level studies (Koopman and Longini, citing Koopman et al. 1991). In fact, when a risk factor affects transmission, neither individual-level analysis nor ecological analysis works. Although infectous diseases have received the greatest attention in such work, psychosocial and behavioral phenomena (e.g., drug use including smoking and alcohol, racism) probably also constitute dependent happenings in some regards.

What measures can be estimated from basic epidemiologic study designs?

The beauty of a follow-up study is that the investigator gets to watch what is happening and to summarize the experience by calculating simple measures like the proportion of exposed subjects who develop the disease ("the incidence of the disease in the exposed") or the rate at which the disease develops in the exposed. This is often not the situation in a case-control study, in which the investigator typically assembles cases without identifying the entire exposed and unexposed populations from which the cases arise.

It is said that a follow-up study "samples by exposure status" and a case-control study "samples by disease status". This is certainly true for a case-control study, but not necessarily so for a follow-up study, which can sample without regard to exposure status. A cross-sectional study can sample by either disease or exposure or neither (i.e., a true "cross-section"). When a cross-sectional study samples by existing disease, it is essentially the same as a case-control study with prevalent cases. However, many of these concepts remain the subject of debate (if interested, see references in the first section of the bibliography).

Multiaxial classification of study designs

There have been various attempts to classify study designs in a more analytic fashion than the conventional taxonomy presented in this chapter. One approach, presented in Kleinbaum, Kupper, and Morgenstern's textbook *Epidemiologic research: principles and quantitative methods*, analyzes major designs in respect to "**directionality**" (cohort studies are said to involve "forward directionality", case-control studies to involve "backward directionality", and cross-sectional studies neither), "**timing**" (the chronological relationship between the most recent data gathering and the occurrence of the study factor and disease – if both study factor and disease were established and measured before the study began, then the study was completely "retrospective"; if both study factor and disease have not yet occurred when the study begins, then the study is completely "prospective" so that measurements can be tailored to the study requirements; studies with exposure data collected both before and disease onset studied after the start of the study were "ambispective"), "**type of population**" (cross-sectional or longitudinal, fixed cohort or dynamic population), and "**unit of observation**" (individual-level data, group-level data).

Various other conceptualizations are also in use. For example, sometimes case-control studies are said to involve "sampling on disease", because cases and controls are sampled separately (as in stratified random sampling). From this perspective, cohort studies are said to involve "sampling on exposure" – exposed and unexposed persons are sampled separately. However, though separate sampling may be necessary in order to obtain a large enough number of participants with a rare exposure, if the exposure is not rare then participants can be selected without regard to exposure status.

Participants for a cross-sectional study can be selected without regard to exposure or disease status, separately by exposure status, or separately by disease status. In the last case, a cross-sectional study is equivalent to a case-control study using prevalent cases. A basic point but one worth noting is that a study cannot estimate a dimension that has been set by its design. That is, if participants are selected separately according to their exposure status, than the proportion who are exposed cannot

be estimated from that study since that proportion is determined by the study design (and its success in recruitment), rather than from the sampling process. If participants are selected according to disease status, then exposure proportions (and odds) can be estimated but not disease prevalence (or odds). That is the reason that one cannot directly estimate risk in a case-control study. (Rothman and Greenland use the term "pseudo-risks" to refer to the proportion of cases among exposed and unexposed case-control study participants.)

Design attributes

As can be seen in the bibliography for this chapter, classification of study designs has been the subject of vigorous debate. Nevertheless, there are various important design attributes that should be noted for any given study. These attributes are:

Subject selection

Under this heading come the various considerations used in selecting participants for the study (e.g., restriction to certain age groups, enforced comparability between groups being compared (matching), natural comparability [twins, siblings], random sampling).

Method of data collection

Data can either be primary data, collected for the purposes of the study at hand or secondary data, collected for purposes other than the study at hand, such as from medical records, death certificates, billing records, or other administrative files. Data may have been collected in the distant past.

Unit of observation

As noted above, data can be collected at the individual level or only at the group level.

Evaluation of a study design

The primary dimensions for evaluating the design of a particular study are:

Quality of information: How accurate, relevant, and timely for the purposes of the study are the data?

Cost-effectiveness: How much information was obtained for how much expenditure of time, effort, resources, discomfort, etc.?

[For more on the above, see Kleinbaum, Kupper, and Morgenstern, *Epidemiologic research: principles and quantitative methods*, ch 4-5.]

The following layout may be useful for reflection or discussion, but cannot be completed unambiguously since in many cases the relative strengths of different designs depend upon the particular study question and circumstances.)

Strengths and weaknesses of the classic study designs

	Cohort		Case-control	
	(prospective)	(historical)	(incident)	(prevalent)
Ability to estimate risk				
Ascertainment of cases				
(access to care, diagnostic criteria, selective survival)				
Measurement of exposure				
Reliance on historical data				
Selective recall				
Disease may affect characteristic				
Control of all relevant variables				
Study affects subject behavior				
Temporality established				
Feasibility and logistics				
Rare exposures				
Rare diseases				
Statistical power and efficiency				
Attrition				
Time and effort				
Ethical concerns				
Cost				

Individual-level interpretations of measures of association

The individual-level follow-up study, cross-sectional study, and case-control study are fundamental designs in epidemiologic research. Data collected using any of these designs allow one to estimate an individual-level measure of association or effect, i.e., a measure of the strength or magnitude of the quantitative relationship of a study factor (i.e., exposure of interest) with a disease. We learned about these measures in a previous chapter. We revisit them here to reinforce the relationship between which measures can be estimated with which study designs.

One way of conceptualizing study designs is to regard the objective of an etiologic individual-level study as the estimation of a measure of effect relating an exposure to a disease outcome, specifically a risk ratio (CIR) or rate ratio (IDR). The preference for these measures is that, as Greenland (1987) demonstrates, they are interpretable at the level of the individual's risk or hazard function so that under certain assumptions an RR of two means that an exposed individual's risk or hazard is twice that of an unexposed individual. (Although the odds ratio does not possess an interpretation in terms of an individual's odds, it is useful through its ability to estimate a risk ratio or rate ratio. Similarly, the prevalence odds ratio is of interest primarily because under certain assumptions it estimates the incidence density ratio (rate ratio) [Greenland, 1987]).

Risk ratio

Consider the example of a pregnant woman who drinks three or more alcoholic drinks per day during pregnancy. Suppose that that drinking that amount of alcohol is associated with a 20% chance of bearing a malformed baby. If that chance is 2% for a pregnant woman who does not drink, the ratio of fetal malformations in relation to drinking three drinks/day is 10 (20%/2%). A risk ratio of 10 indicates a very strong association and therefore one that is more likely to be causal. Also, the relative risk conveys a clear, intuitive meaning about the degree by which the exposure increases risk.

We can also interpret this risk ratio at the individual level: the risk for an individual woman who drinks 3+ alcohol drinks/day during pregnancy was 10-times (or 900% greater than) that for a woman who does not drink. Such an interpretation, of course, involves a number of assumptions, i.e., that apart from the effect of drinking, the women in the exposed group have the same risk as women in the unexposed group and that the individual woman to whom the group-level association is being imputed has risk-related characteristics close to the group average. But mathematically there is no problem. [Aside: Birth outcomes such as fetal malformations are generally regarded as prevalences among babies born, since the denominator for births is generally unknowable; for simplicity the above example assumes that all pregnancies result in a live birth.]

Rate ratio

Often we estimate disease rates, rather than risks, in which case the measure of effect of interest is a rate ratio. For example, in a study of breast cancer in relation to early use of oral contraceptives, we may have anywhere from 10 to 20 years of follow-up on subjects. To accomodate these differing lengths of follow-up, we can calculate the rate of breast cancer cases per woman-year, rather than per woman. In that case a two-fold elevation would mean that the rate at which breast cancer cases are observed in women with early use of oral contraceptives was twice that in women without early use of oral contraceptives. Again, the rate ratio has an interpretation at the individual level (Greenland, 1987) and can be mathematically converted into an estimate of relative risk over a given time interval. It can also be interpreted in terms of the expected time until the event occurs in the average woman.

Incidence odds ratio

The incidence odds ratio is the ratio of odds of disease in exposed persons to the odds of disease in unexposed persons. Odds are ratios of risks. If the risk is r, the odds are r/(1–r). When the risk is small, risk and odds are nearly equal, and the odds ratio approximates the rate ratio and risk ratio.

Since the odds ratio can be estimated in a case-control study even where no other measure of relative risk is directly available, the odds ratio is of great practical importance for epidemiolgists. The prevalence odds ratio (from a cross-sectional study) also approximates the rate ratio when the duration of the condition is unrelated to exposure status. The prevalence ratio can also be the measure of primary interest, when duration is itself the outcome, such as in the treatment of depressive disorder. However, mathematically (see Greenland) there is no direct individual-level interpretation for the odds ratio (whereas the incidence proportion is the sum of the risks across all

individuals, this relationship does not hold for the incidence odds and individual odds). For this reason, Greenland argues, the CIR and IDR are preferred.

Preferred measures of association

So, our primary interest for etiologic purposes is generally the risk ratio (CIR) or rate ratio (IDR). Where we cannot estimate either of those directly, then we usually try to design the study so that we can estimate the odds ratio and use it to estimate the rate ratio or risk ratio. We may also want to estimate a measure of impact, to quantify the importance of the relationship we are studying should it turn out to be causal. In the table below are listed the kinds of measures of association and impact that can be derived from the basic epidemiologic study designs:

Measures of association for basic epidemiologic study designs

Type of study design	Measure of association	Measure of impact
Follow-up, person denominator	Risk ratio	Absolute Relative
Follow-up, person-time denominator	Rate ratio	Absolute Relative
Case-control	Odds ratio	Relative
Cross-sectional	Prevalence odds ratio or prevalence ratio	Relative

Formulas for and examples of computation

Construct (2x2, four-fold) table:

	Exposure			
Disease	Yes	No	Total	
Yes	a	b	m_1	(a + b)
No	c	d	m_2	(c + d)
Total	n_1	n_0	n	
	(a + c)	(b + d)		

Example: The following are hypothetical data involving subjects who have been determined to be either hypertensive (diastolic blood pressure >90 mmHg) or normotensive (diastolic blood pressure <=90 mmHg) and were classified into one of two categories of dietary salt intake, high or low.

Dietary Salt Intake and hypertension(Hypothetical)

	Dietary salt intake		
Hypertension	High	Low	Total
Yes	135	160	295
No	180	420	600
Total	315	580	895

If these data came from a follow-up study, then the risk of disease in exposed subjects would be a / n1, the risk in unexposed subjects would be b / n0, and the risk ratio or relative risk would be:

$$RR = \frac{a / n_1}{b / n_0} = \frac{135 / 315}{160 / 580} = 1.55$$

If these data came from a cross-sectional study, the calculations would be identical except that the data would yield measures of prevalence and a prevalence ratio instead of risk and risk ratio. However, the prevalence odds ratio (see below) would generally be preferred as a measure of association, since under the assumption of no difference in duration of hypertension between high- and low-salt people, the prevalence odds ratio estimates the incidence density ratio in the population.

If these data came from a case-control study, the above calculations would not be meaningful. Since a case-control study samples subjects on the basis of their disease status, proportion of exposed who are cases does not estimate anything. Rather, we need to compute the odds of cases and controls who are exposed! Thanks to the odds ratio, we can estimate the rate ratio in the population from which the cases arose:

Odds of Exposure in Cases (D):

$$\text{Odds} = \frac{\text{Proportion of cases exposed}}{\text{Proportion of cases not exposed}} = \frac{a / (a + b)}{b / (a + b)} = \frac{a}{b}$$

Odds of Exposure in Controls (D):

$$\text{Odds} = \frac{\text{Proportion of controls exposed}}{\text{Proportion of controls not exposed}} = \frac{c / (c + d)}{d / (c + d)} = \frac{c}{d}$$

Odds Ratio (OR)

$$OR_e = \text{Exposure odds ratio} = \frac{\text{Exposure odds in cases}}{\text{Exposure in noncases}} = \frac{a / b}{c / d} = \frac{ad}{bc}$$

$$OR_e = \frac{ad}{bc} = \frac{135 \times 420}{160 \times 180} = 1.97$$

Intervention Trials

(An earlier version of this section was written by Joellen Schildkraut, Ph.D.)

In an experiment, a set of observations are conducted under controlled circumstances. In contrast to nonexperimental, observational epidemiologic studies, experimental studies permit the scientist to manipulate conditions to ascertain what effect such manipulations have on the outcome. The objective of an experiment is the creation of duplicate sets of circumstances in which only one factor that affects the outcome varies. An example is laboratory animal studies such as those which evaluate potential carcinogens.

In such studies, the investigator has a great deal of control over the experimental units, their environment, measurements taken, and exposure to the study factors. Even genetic factors can be controlled by using inbred strains of mice. Experiments provide a means to disentangle complex problems in stepwise fashion, to reduce macro-level phenomena into collections of low-level mechanisms. This reductionist approach, made possible by laboratory experimentation, has made possible the remarkable advances in knowledge and technology of the past few centuries. The rub is that not all phenomena are amenable to dissection in this way. Laboratory experimentation on humans is greatly constrained, and extrapolation from animals to humans often problematic. Also, many conditions of interest cannot be manipulated and it is generally impossible to recreate real-life situations in the laboratory.

In epidemiology, intervention trials are the closest analog of a laboratory experiment. What distinguishes intervention trials from other types of epidemiologic studies is the manipulation of the study factor. This manipulation may be governed by random assignment, creating a true experiment, or if not, a quasi-experiment. Randomization offers the greatest opportunity to create groups that are equivalent in all regards, with the corresponding opportunity to isolate the effect of the intervention. The potential for achieving such isolation in a study with nonrandom assignment depends on the ability to adjust for differences in the analysis. Even with good data on all relevant factors adjustment may not be possible. For example, no analytic technique could correct a study where all patients with a better prognosis were assigned a new drug instead of an old drug.

Intervention trials can include testing therapeutic or preventative hypotheses, the estimation of long term health effects, and identification of persons at high risk. Types of interventions include:

- Prophylactic - focus on prevention (e.g. vaccines, cholesterol lowering
- Diagnostic - focus in evaluation of new diagnostic procedure (e.g. comparison of a less invasive diagnostic procedure to a gold standard, etc.)
- Therapeutic - focus on treatment (e.g. drug testing, evaluation of new surgical technique, etc.)

A randomized clinical trial (RCT) is defined as a prospective study that estimates the effect of an intervention by comparing participant outcomes between randomly assigned treatment and control groups. The major RCTs are multi-center studies in two or more hospitals with a common protocol. The strengths of multi-center studies are more representative patient populations, larger sample size, and shorter study period (or duration of patient intake). Finally multi-center studies enable research on rare diseases.

Drug trials go through several levels of study:

Phase I - early study to determine dose level that is not too toxic (animal studies)

Phase II - efficacy trial to estimate the effectiveness of an agent with specified precision.

Phase III - comparative trial to test whether the new agent is better than the standard or control agent.

Phase IV - for the detection of rare side effects by way of epidemiologic studies or prospective monitoring

Steps of a Clinical Trial

There are three phases in a clinical trial: 1) planning, 2) the trial (data collection), and 3) concluding phase:

1. Planning phase

Study Design

Clinical trials can be randomized controlled studies or nonrandomized studies (quasi-experiments). If the latter they can have concurrent controls (a group or groups that are regarded as similar to the experimental group and whose experience is observed during the same period of time as that of the experimental group), historical controls (a group regarded as similar and for which data are already available), or sometimes no controls.

Randomization is a method for allocation of subjects to intervention and control groups where each subject is equally likely to be assigned to one or the other. Various randomization procedures have been proposed for use in clinical trials. The most frequently used techniques are:

Simple randomization - assignment of people to treatment groups is random, not concerned with other variables

Balanced Block randomization - ensures balance in the proportion of patients assigned to each treatment with in each group or blocks of patients entered. (e.g. hospitals in a multicenter study).

Stratified randomization - is used when there are specific factors known to have a significant effect in the outcome of the trial. Separate balanced block randomization schemes are established within each level of the stratified variable or variables.

In a multicenter study, randomization can be stratified by institution because of institutional differences in the patient population, in the overall level of patient care, or in the treatment effect of the institution.

Blindness or masking

Non-blinded	- common in community trials
Blinded	- the observer is aware but the subject is not aware of treatment assignment
Double blinded	- Neither the observer or the subject is aware of treatment assignment
Triple blinded assignment	- The observer, subject, and data analyst are not aware of treatment

Concurrent and non-randomized controls can result in systematic assignment bias and uninterpretable results. Historical controls may not be comparable in terms of patient selection, external environment (even if it is the same hospital), improved diagnostic tests, and unknown factors, but the cost is cheaper and the length of the time to complete the trial is shortened. Evidence for bias in treatment assignment of controlled clinical trials was illustrated in a study by Chalmers et al. (*N Engl J Med* 1983; 309:1358-61):

Type of Study	No. of studies	>=1 significant prognostic variable	Significant Difference in fatality rate
Blinded Randomized	57	14.0 %	8.8%
Unblinded randomized	45	26.7 %	24.4%
Non-randomized	43	58.1 %	58.7%

Sample size estimates are vital to planning the study. The estimated difference in the response variable (outcome of interest), significance level, and noncompliance rate must be factored into the calculation of sample size.

2. Trial phase (data collection)

Screening can be applied to those already admitted to the hospital or those who can be contacted from outpatient services. Patients should be those who are likely to benefit from the intervention and those who are likely to comply with the intervention schedule.

Treatment allocation can be 1) fixed in the beginning, optimally in one to one ratio, 2) it can be adaptive allocation where results of an ongoing trial influences allocation so that the proportion of patients with beneficial treatment is maximized or 3) crossover design which helps to eliminate the variation between patients.

Study monitoring can be implemented so that if trends demonstrate that one treatment was significantly better or worse than the other with respect to any study endpoints (mortality, morbidity, side effects) it would be the responsibility of a special committee to determine whether the study should be terminated.

Long term follow-up is important in clinical trials since patients sometimes do not adhere to the originally assigned therapy.

3. Analysis and publication phase

Some issues of relevance to the analysis of data from randomized clinical trials include: baseline comparability of treatment groups, selection of prognostic factors, methods for evaluating treatment differences, non-adherence to assigned therapy, and post-stratification. Survival analysis is often the method of choice. A major consideration is how to analyze data when 1) persons are discovered, after randomization, who do not meet entry criteria, 2) withdrawals, 3) noncompliant subjects, 4) subjects who switch treatments. Exclusion of any groups will "undo" the pure randomization scheme and could result in a biased estimate of effect.

Advantages and disadvantages of RCTs:

Advantages	Disadvantages
1. Prospective	1. Contrived situation
2. Randomization	2. Human behavior may be difficult to control
3. Clear temporal sequence	3. Ethical constraints
4. Best evidence for causation	4. Exclusions may limit generalizability
	5. Expensive in time, personnel, facilities, and budget

Case-control studies

Of the three remaining classic epidemiologic study designs – cross-sectional, cohort or follow-up and case-control – the case-control is the least straightforward. We will therefore devote the following section to examining the "anatomy" and "physiology" of case-control studies.

Definition of a case-control study

A study that starts with the identification of persons with the disease (or other outcome variable) of interest, and a suitable control (comparison, reference) group of persons without the disease.

> The relationship of an attribute to the disease is examined by comparing the diseased and nondiseased with regard to how frequently the attribute is present or, if quantitative, the levels of the attribute in each of the groups. — Last JM, A dictionary of epidemiology. 2nd edition, NY, Oxford, 1988
>
> Synonyms: case comparison study, case compeer study, case history study, case referent study, retrospective study

Defining characteristic

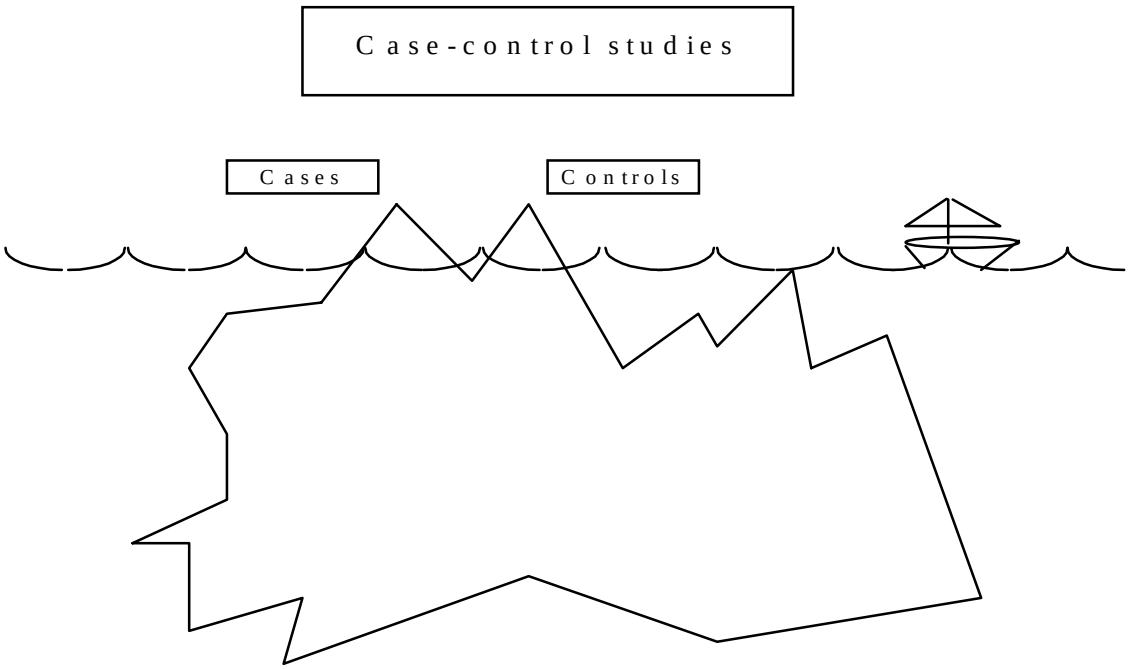
Subjects are selected on the basis of the outcome variable.

Key advantages

- Statistically efficient for rare conditions
- Logistically efficient for prolonged induction or latency diseases
- Can examine many exposures in one study
- Ethical - cannot affect onset of disease

Basic procedure

1. Identify cases, determine their characteristics - cases estimate the prevalence of the exposure in people who get the disease.
2. Select controls (noncases), determine their characteristics - controls estimate the prevalence of the exposure in people who have not developed the disease.
3. Compare the characteristics of cases with characteristics of noncases.
4. Draw inferences about the underlying processes that led to differences in characteristics of cases and controls. Odds ratio (OR = odds of exposure in cases/odds of exposure in controls) estimates the incidence density ratio (IDR = rate of disease in exposed persons/rate of disease in unexposed persons). For rare disease, IDR closely approximates cumulative incidence ratio (CIR, RR) of the disease for that exposure.



Example

If we want to test the hypothesis that exogenous estrogen is an etiologic factor in cancer of the uterine endometrium, we assemble a (case) group of women who have developed endometrial cancer (preferably newly-detected cases) and a (control) group of women whom we believe accurately reflect the population from which the cases have come. The case group will be used to estimate usage of estrogen by women who developed endometrial cancer; the control group will be used to estimate usage of estrogen by women in the source population (the "study base") which gave rise to the case group.

	Estrogen			
Endometrial cancer	Yes	No	Total	
Case	a	b	m_1	(a + b)
Control	c	d	m_2	(c + d)
Total	n_1 (a + c)	n_0 (b + d)	n	

$$OR_e \text{ (Exposure odds ratio)} = \frac{(a / m_1) / (b / m_1)}{(c / m_0) / (d / m_0)} = \frac{ad}{bc}$$

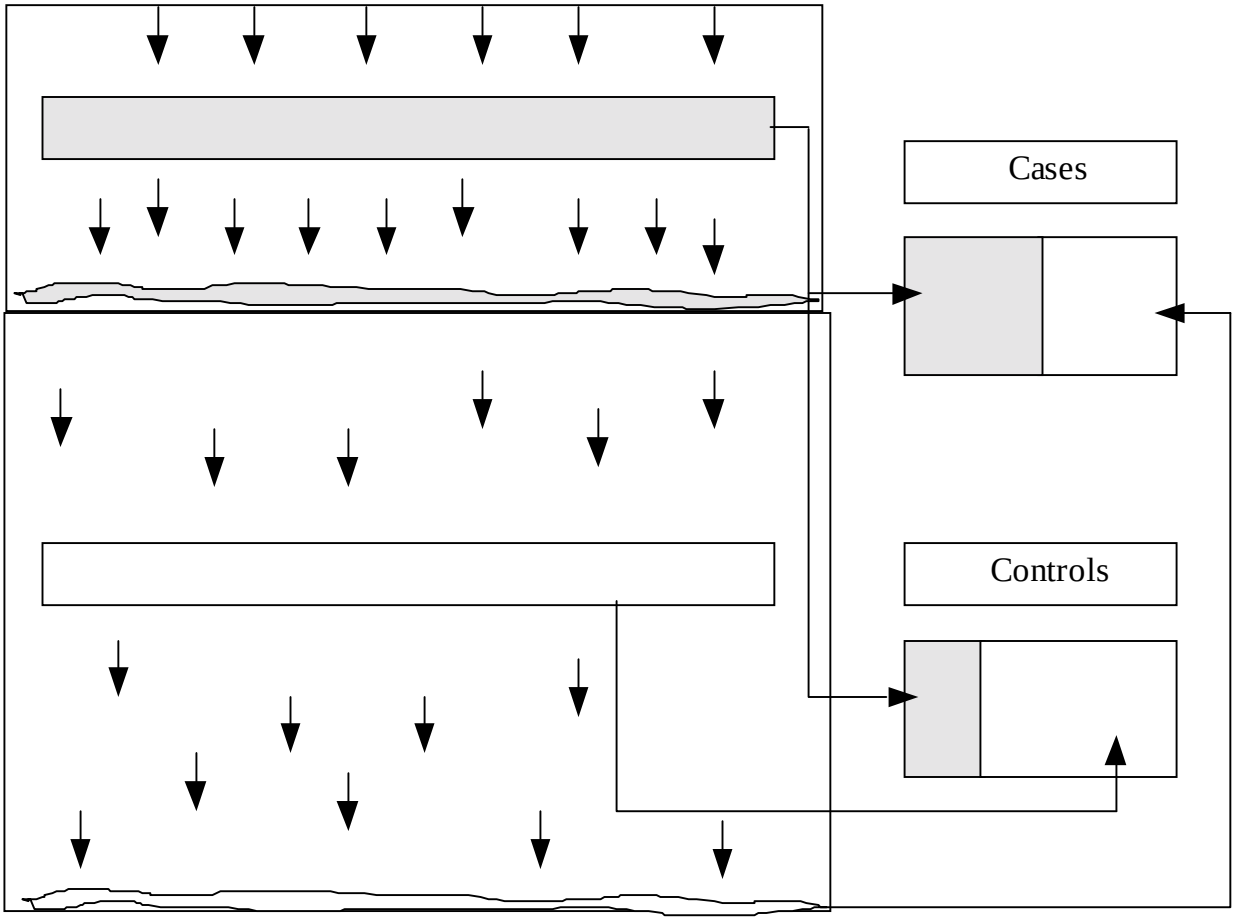
estimates IDR and CIR (rare disease)

If we wish to obtain an estimate of the incidence density ratio (or the relative risk) for endometrial cancer with respect to estrogen use, we can use the proportion or prevalence of estrogen use in the endometrial cancer cases to compute the odds of estrogen use among women who develop endometrial cancer [$p_{estrogen}|case/(1-p_{estrogen}|case)$] and the proportion or prevalence of estrogen use in the controls to compute the odds of estrogen use in the population [$p_{estrogen}|noncase/(1-p_{estrogen}|noncase)$]. The odds ratio for exposure is then the ratio of these two odds, and gives us the estimate of the relative risk (since endometrial cancer is a rare disease) and, if we have selected our cases and controls appropriately, of the incidence density ratio.

Rationale for the odds ratio

1. The cases provide an estimate of the prevalence of the exposure in people who get the disease.
2. The number of exposed cases (and therefore the proportion or prevalence of exposure among cases) reflects the rate of disease in exposed people in the population. The number of unexposed cases reflects the rate of disease in the unexposed population.
3. The odds of exposure in the cases (proportion exposed/proportion unexposed) therefore reflect the ratio of disease rates (or risks) in the population.
4. The controls provide an estimate of the prevalence of the exposure characteristic in people who have not developed the disease.
5. The odds of exposure in the controls (proportion exposed/proportion unexposed) reflect the odds of exposure in the population.
6. So the odds ratio (OR) [odds of exposure in cases/odds of exposure in controls] indicates the relative risk [incidence of disease in exposed persons/incidence of disease in unexposed persons].

The above rationale is presented to convey a "feel" for why the odds ratio from a case-control study conveys information about the strength of association between a disease and exposure.


Cases
Controls

Validity

The validity of a case-control study requires that:

- Cases in the study adequately represent the relevant cases (the population of cases about whom inferences are to be made) with respect to the variables of interest (notably, prevalence of exposure). This depends upon whether the cases available do in fact reflect the rates of disease in exposed and unexposed individuals undistorted by differential manifestation, detection, or short-term survival (e.g., selective survival, access to care, detection bias);
- Controls accurately reflect the exposure proportions in the study base (the source population for the cases). For example, hospitalized controls may overrepresent exposures associated with hospitalization for other conditions.

Both of these requirements, especially the latter, can be difficult to ensure. Therefore, case-control studies are regarded as highly susceptible to bias from problems with the:

Identification of cases

- Reliance on medical care system
- Often miss subclinical cases (detection bias?)
- Can miss rapidly fatal cases (selectively?)

Selection of controls

- Selection of controls can determine the study results
- Which controls are appropriate is often not obvious
- Trade-off between sampling and data collection
- Hospitalized controls, community controls, dead controls
- Controls may be reluctant to cooperate

Measurement of exposure for cases and controls

- Reliance on recall or records (differential?)
- Effect of disease on exposure assessment
- Effect of disease on exposure (confounding by indication)

There is also the thorny problem of establishing temporality, i.e., did the exposure precede the disease?

Interpretability of the odds ratio

Why does the OR from the cases and controls we have assembled estimate anything in the population? Consider what the cells in the table below represent. Assume that the cases were selected as newly occurring cases of endometrial cancer over a period of time in a defined

population and that the controls were selected at the same time as the cases from among women in that population (this is called "density sampling of controls").

Endometrial cancer	Estrogen: Yes	Estrogen: No	Total	
Case	a	b	m_1	(a + b)
Control	c	d	m_0	(c + d)
Total	n_1 (a + c)	n_0 (b + d)	n	

If this situation, the cases would be all (or some fraction f_1 of) cases of endometrial cancer in the population. If the incidence rate of endometrial cancer is ID and the amount of population-time is N women-years, then:

$$m_1 = (f_1)(ID)(N)$$

[f_1 is included only for purposes of generality — if all cases are included, then f_1=1 and can be ignored.]

Cases among women taking estrogen (cell "a") would be:

$$a = (f_1)(ID_1)(N_1)$$

where ID_1 and N_1 are the incidence rate and population-time, respectively, for women taking estrogen.

Similarly, cases among women not taking estrogen (cell "b") would be:

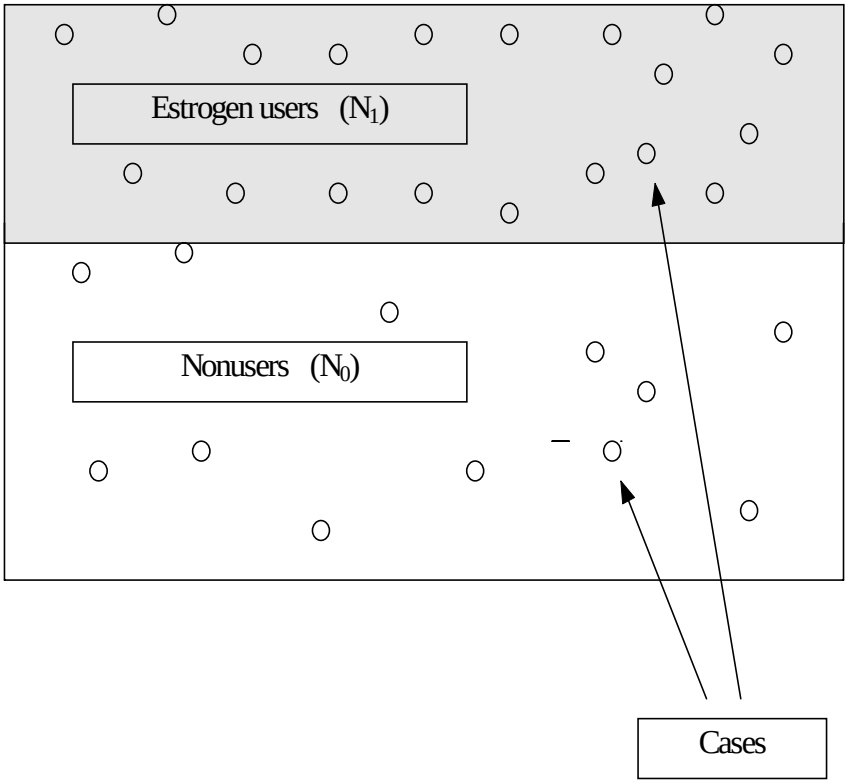
$$b = (f_1)(ID_0)(N_0)$$

with ID_0 and N_0 applying to women not taking estrogen.

Note: Whether N_1 and N_0 represent women-years of estrogen use or women-years in estrogen users (i.e., are person-years for a women after she stops taking estrogen counted as exposed or unexposed) would depend upon whether the estrogen effect endures after the drug is discontinued.

We now see where the cases have come from. What about the controls? The control group is typically, though not necessarily, chosen in some fixed ratio to the number of cases, such as two controls per case.

Since the strategic advantage of the case-control method is that we do not need to enroll the entire population from which the cases arise, the number m_0 of controls will be some small fraction f_0 of the noncases in the population. If we have 200 cases and have decided to select 400 controls, then f_0 would be 400 divided by the population size or the amount of population-time. f_0 is included to demonstrate the link between the case-control study and the population from which the cases arise, also referred to as the study base. In actual practice we establish the number of cases and controls required to meet certain sample size (statistical power) objectives; the sampling fraction f_0 is what results from the number of controls we seek. Since we do not – must not – choose our controls separately from each exposure group, the number of exposed (c) and unexposed (d) controls will be determined by the amount of population-time in each exposure category:



New cases among exposed	=	$ID_1 N_1$
New cases among unexposed	=	$ID_0 N_0$
Exposure odds in cases	=	
Exposure odds in noncases	≈	N_1 / N_0

$c = f_0N_1$ = number of exposed noncases

$d = f_0N_0$ = number of unexposed noncases

If the N's represent population-time, rather than simply population size, f_0 reflects sampling over time as well as over people.

The key point about f_0 is that for the control group to provide a valid estimate of the relative sizes of exposed and unexposed population-time, f_0 must be the same for both exposed controls (c) and unexposed controls (d).

The above discussion can be summarized in a revised 2 x 2 table:

	Exposed	Unexposed
Cases	$f_1ID_1N_1$	$f_1ID_0N_0$
Controls	f_0N_1	f_0N_0

With this background, we are ready to see how the OR can estimate the IDR:

$$OR = \frac{ad}{bc} = \frac{(f_1ID_1N_1)(f_0N_0)}{(f_1ID_0N_0)(f_0N_1)} = \frac{ID_1}{ID_0} = IDR$$

Numerical example

Assume a stable, dynamic population of 4 million women age 40 years or older, in which 1,000 incident cases of endometrial cancer occur each year (i.e., 1,000 cases/4 million women-years).

Increase rate to 2,500 cases/ 100,000 wy?

Assume:

- N_1 = 1 million women-years (1,000,000 wy or 1×10^6 wy) of estrogen use
- N_0 = 3 million women-years (3×10^6 wy) of unexposed person-time
- ID_1 (incidence density in exposed) = 40×10^{-5}/year (40/100,000 wy)
- ID_0 (incidence density in unexposed population) = 20×10^{-5}/year, so that the IDR is 2.0

In the 1×10^6 exposed women-years, there would be 400 cases.

In the 3×10^6 unexposed women-years, there would be 600 cases.

Of the 1,000 cases, 400 are exposed and 600 are unexposed. The prevalence of exposure among cases is 400/(400+600) = 40%; the exposure odds in cases would be .40/.60 = 0.67.

The expected prevalence of exposure in an unbiased sample of noncases would be, since the disease is so rare, $N_1/(N_1+N_0) = (1 \times 10^6) / (1 \times 10^6 + 3 \times 10^6) = 0.25$; the exposure odds among noncases would be 0.25/0.75 = 0.33.

The exposure odds ratio (OR) would therefore be:

$$OR = (.40/.60)/(.25/.75) = 0.67/.0.33 = 2.0$$

A case-control study that recruited (randomly) 200 cases and 400 controls (f_1 = 200/1,000 = 0.2; f_0 = 400/4,000,000 = 1/10,000 or 10^{-4}) would be expected to have the following results.

Expected Results from Hypothetical, Poplation-based Case-Control Study of Endometrial Cancer and Estrogen

	Exposed	Unexposed	Total
Cases	80	120	200
Controls	100	300	400
Total	180	420	600

$$OR_e = \frac{80 \times 300}{120 \times 100} = 2.0$$

It is apparent that given the rarity of the disease it makes no practical difference here whether the prevalence of exposure in the source population from which the cases emanate (i.e., the study base) is estimated from the total population or only from those without the disease.

Identifying the study base

Disease and other types of events occur in populations. Case-control studies provide a window into the process of disease occurrence in a population, without the necessity of studying the entire population. Thus, case-control studies are best understood by considering what is happening in the population (the study base) and by analyzing the relationship between it and the case-control study.

But how do we identify the study base? The study base or source population consists of those people who would have been available to be counted as cases had they developed the disease or

experienced the event under study. Thus, the source population must be at risk for the disease and for being selected as cases if they were to develop it. Moreover, the relevant exposure measure for both cases and for the source population is the time during which the disease was initiated or promoted in the cases. Identifying the relevant period in time can be an issue for a disease with a lengthy induction and/or latent period, such as most cancers, if the disease is common and/or the population or its exposure distribution is undergoing substantial change.)

The first step in identifying the study base is generally based on geography or membership. Thus, for cancer cases from a state with a tumor registry, the study base is the state (or a portion of the state if only cases from a certain portion of the state are being studied). For cases detected in a managed health care organization, the study base is its membership. If identification of cases is being made though hospitals, then the study base is the population from whom people would go to that hospital if they developed the disease. This last situation can be complicated by factors such as the extent to which some people go to hospitals not covered in the study and whether the disease is one which does not always lead to hospitalization.

An important next step is to identify that subset of the population that is truly at risk for the disease (and its detection). For endometrial cancer, obviously the study base does not include men. Does the study base include hysterectomized women? Certainly not, since women without a uterus obviously cannot develop endometrial cancer – though if the hysterectomy was recent, a woman could be part of the study base for cases detected prior to that time. (Also, there may be the potential for selective depletion of endometrial cancer susceptibles, but we will not consider that possibility here.)

Selecting a control group representative of the study base

At least as problematic as identifying the study base is coming up with a way to obtain a control group that will faithfully represent it. One obvious choice, which is now much more common than in earlier decades, is to carry out a random sample survey of the study base as it exists at the time of the study.

This approach is most likely to be valid if:

- an accurate sampling frame exists or is constructed
- a representative sample is drawn and adequately executed
- response rates are high and data are of adequate quality (high rate of accuracy)

Controls recruited from hospitals and other noncommunity-wide sources are nevertheless of interest because the cost and logistical challenges are often not as great, greater cooperation may be obtained, and data quality may be better than that from the general population. However, when controls are obtained from sources other than a random sample survey, validity depends upon whether these controls have the same exposure distribution as the study base. For example, selecting controls from friends of the cases ("friend controls") can lead to bias because people tend to choose friends because of shared interests, perspectives, affiliations, and so on which are often associated with exposures. Thus, the proportion of many exposures in friend controls will be more

similar to that in the case group than in the study base as a whole. The use of friend controls is an example of "over-matching".

What about if some subsets of the study base are at much higher risk than others, due to, for example, genetic factors or simultaneous exposures? If the difference in risk is great, then both the case group and study base should be demarcated on that risk factor, and separate (stratified) analyses carried out.

Variants in the basic case-control design

There are several ways in which the case-control study design can be implemented.

- Incident versus prevalent cases: Case-control studies can use only new cases (**incident cases**) of the disease, thereby avoiding some of the sources of bias inherent in the use of **prevalent cases** (e.g., influence of survival/duration of the condition), or they can use prevalent cases.

- **Defined population** or nesting: Case-control studies can be carried out in a geographically-defined population, e.g., a state where a cancer register provides notification of all incident cases from a known denominator population, or in a cohort that has been followed (e.g., an occupational group). Having a defined population offers further advantages (such as availability of an identified universe for selection of controls, knowledge of the denominator from which migration has occurred, measurement of key variables prior to the disease). A case-control study within an identified cohort is sometimes termed a "nested case-control" study. (Rothman and Greenland regard nearly all case-control studies as nested in their source population.)

- Eligible controls: Although classically, the controls in a case-control study were noncases, in some designs people who later develop the disease can still serve as controls.

Types of control groups - case-control, case-cohort

The controls in a case-control study can be selected from among (a) persons who have not developed the disease by the end of the period of case ascertainment (prevalence controls), (b) persons who have not developed the disease at the time each case occurs - such controls are usually matched in time to the cases (density sampling), or (c) persons at risk to become a case at the outset of case ascertainment.

These controls may be selected before or after case ascertainment. Rodrigues and Kirkwood (1990) call the three types of controls, respectively, "exclusive", "concurrent", and "inclusive". The traditional approach is method (a), "exclusive" controls. With this method, only people who remain free of the disease to the end of case ascertainment are accepted as controls. The odds ratio in this situation estimates the incidence (i.e., risk) odds ratio in the cohort from which the cases arose. For a rare disease, this incidence odds ratio estimates the CIR.

In the second sampling scheme (density or concurrent sampling [method (b)), a participant can be selected as a control at a given point even if that participant later develops the disease. With this

approach, the odds ratio computation estimates the relative rate (IDR) on the assumption that the IDR does not change during the follow-up period (assuming matching of controls to cases by time) (see Greenland and Thomas, 1982 and Rodrigues and Kirkwood, 1990). This study design has been referred to as a "density case-control study" (Hogue et al., 1983 referred to this design as a "case-exposure study"; however, Rodrigues and Kirkwood (1990) use that term for the third design [method (c)]). If a participant selected as a control later develops the disease, then that participant is also counted as a case; his/her data are used both as a case and as a control (his/her data appear in both categories).

The third design [method (c)] has been called "case-base" and "case-cohort" (also "case-exposure" – see Rodrigues and Kirkwood for citations). When such a case-control study is carried out within a fixed cohort, the odds ratio estimates the risk ratio with no rare disease assumption.

Rodrigues and Kirkwood show that the three ratio measures of association – CIR, IDR, and OR – can each be expressed so that its numerator is the odds of exposure in cases. Thus, all that differs are the denominators, and the three different approaches to sampling controls provide estimates for the respective denominators.

a. OR_e = Exposure odds ratio $= \dfrac{\text{Odds of exposure given disease}}{\text{Odds of exposure given non-diseased}} = \dfrac{a / b}{\mathbf{c / d}}$

(c / d is the odds of exposure in non-cases [never-cases at end of ascertainment period])

b. $IDR = \dfrac{ID_1}{ID_0} = \dfrac{a / py_1}{b / py_0} = \dfrac{a / b}{\mathbf{py_1 / py_0}}$

(py_1 / py_0 is ratio of exposed to unexposed person-years, from density sampling)

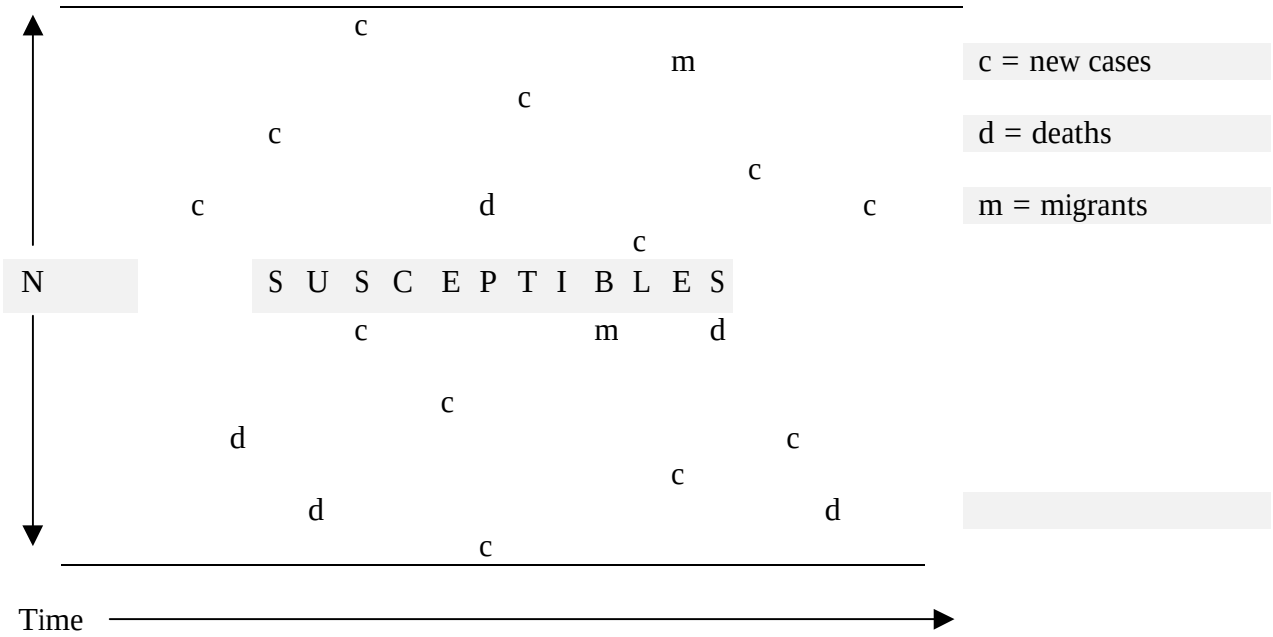
c. $CIR = \dfrac{CI_1}{CI_0} = \dfrac{a / n_1}{b / n_0} = \dfrac{a / b}{\mathbf{n_1 / n_0}}$

(n_1 / n_0 is the odds of exposure in the source population for the cases at the start of the follow-up)

where "a" = exposed cases, "b" = unexposed cases, and "n" and "py" represent persons and person-years for exposed (subscript 1) and unexposed (subscript 0).

A modern perspective

In general, design issues in a case-control study are best understood by considering how the issues would be dealt with in a randomized clinical trial (Feinstein, 1985) or a cohort study (Rothman and Greenland). In fact, students of epidemiology (including those of us on the other side of the podium) might have an easier time if the terms cohort study and case-control study had never been introduced, but rather the various approaches of studying disease presence and occurrence in a population classified in regard to the "windows" they provide into the development of the disease in the population.



The above diagram depicts a population of size N followed over time interval t. Suppose N_0 are susceptible (to a specific outcome) and that a surveillance system exists to detect cases (c's) of various diseases or events. For the moment, let us focus on a particular disease, and assume that M cases develop during the follow-up period shown. We will also focus on a particular exposure, to which N_1 of the population are exposed, leaving N_0 unexposed. We will designate the total population-time in the exposed group as N_{1t} and that in the unexposed group N_{0t}. The population distribution of disease and exposure are summarized in the following table.

	Exposure		
	Yes	No	Total
Cases	A	B	M_1
People	N_1	N_0	N
Incidence proportion	A/N_1	B/N_0	M_1/N
Incidence proportion difference	$(A/N_1) - (B/N_0)$		
Incidence proportion ratio	$(A/N_1) \,/\, (B/N_0)$		
Person-time	N_{1t}	N_{0t}	N_t
Incidence rate	$A/(N_{1t})$	$B/(N_{0t})$	$M_1/(N_t)$
Incidence rate difference	$(A/N_{1t}) - (B/N_{0t})$		
Incidence rate ratio	$(A/N_{1t}) \,/\, (B/N_{0t})$		

We can estimate any of the measures in this table with data from appropriately selected random samples and good historical information. For example, if we choose a random sample (n) from the original susceptible population (N), the ratio of exposed persons in the sample (n_1) to unexposed persons in the sample (n_0) estimates (N_1/N_0) the odds of exposure in the original population. If we then choose a random sample (of size m_1) of the M_1 cases (or obtain data from all M_1 cases), then ratio of cases in the sample who were exposed at the beginning of the period (a) to unexposed cases in the sample (b) estimates the odds of exposure in cases. By including only cases who were present in the population at the start of the period, we can then estimate the incidence proportion ratio $[(A/N_1)/(B/N_0)]$ as the ratio of the estimated odds of exposure in cases (a/b) divided by the estimated odds of exposure in the susceptible population at the start of the follow-up period (n_1/n_0). This estimate will be accurate if we have representative samples, accurate assessment of baseline exposure, and no loss to follow-up from outmigration or deaths. If in addition we know N, the size of the original susceptible population, then we can also estimate N_1 and N_0 as, respectively, $(n/N)n_1$ and $(n/N)n_0$, thereby allowing us to estimate incidence proportions and the incidence proportion difference. With this design we can estimate incidence density proportion ratios for any diseases for which a surveillance system (possibly our own) is available and any exposures for which we can obtain baseline data. Note that no rare disease assumption is involved in the above estimates.

If duration of follow-up time is important, we need to estimate the ratio of exposed and unexposed susceptible follow-up time. We can do this by sampling the susceptible population over time, instead of at baseline, in such a way that the probability of selecting a person is proportional to the amount of time he/she is susceptible ("density sampling"). One method for doing this is "risk-set" sampling, in which a susceptible person is sampled at the same date that each case occurs. The ratio

of exposed to unexposed persons sampled in this way estimates N_{1t}/N_{0t}, which we can use to estimate the incidence rate ratio.

Finally, if we choose to sample susceptibles at the end of the follow-up period (Rothman and Greenberg call this the "cumulative" design), then we can estimate the incidence odds ratio, which if the disease is rare will approximate the incidence rate ratio and the incidence proportion ratio. See Rothman and Greenland, chapter 7.

Bibliography

Hennekens and Buring. Rothman - *Modern Epidemiology*, Chapters 3-4, 6. Rothman and Greenland, Chapters 5-7. Lilienfeld and Lilienfeld - *Foundations of epidemiology*, Chapters 8-10; Mausner & Kramer - *Epidemiology: an introductory text*, Chapters 3 and 7-8. MacMahon & Pugh - *Epidemiology principles and methods*, Chapters 5 and 11-12. Fletcher, Fletcher, and Wagner - *Clinical epidemiology*, Chapters 5-7 (through page 118). Schlesselman. Case-control studies. Chapter 1. Kleinbaum, Kupper, Morgenstern - *Epidemiologic research*. Chapters 4-5.

Gray-Donald, Katherine; Michael S. Kramer. Causality inference in observational vs. experimental studies. *Am J Epidemiol* 1988; 127:885-892. See also correspondence from *Am J Epidemiol* 1989; 130:206-209.

"Multilogue" in the *Journal of Chronic Diseases / Journal of Clinical Epidemiology*:

> Kramer, Michael S.; Jean-Francois Boivin. Toward an "unconfounded" classification of epidemiologic research design. *J Chron Dis* 1987; 40(7): 683-688.
>
> *J Clin Epidemiol* 1988; 41(8):
>
> Miettinen, Olli S. Striving to deconfound the fundamentals of epidemiologic study design, 709-713.
>
> Greenland, Sander, Hal Morgenstern. Classification schemes for epidemiologic research designs, 715-716.
>
> Kramer, Michael S.; Jean-Francois Boivin. The importance of directionality in epidemiologic research design, 717-718.
>
> *J Clin Epidemiol* 1989; 42(6):
>
> Feinstein, Alvan R. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. 481-490.

Miettinen, Olli S. The clinical trial as paradigm for epidemiologic research, 491-496.

Feinstein, Alvan R. Unlearned lessons from clinical trials: a duality of outlooks, 497-498.

Miettinen, Olli S. Unlearned lessons from clinical trials: a duality of outlooks, 499-502.

Szklo, Moyses. Design and conduct of epidemiologic studies. *Preventive Medicine* 1987; 16:142-149 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

Ecologic studies

Cohen, Bernard L. Invited commentary: in defense of ecologic studies for testing a linear-no threshold theory. *Am J Epidemiol* 1994;139:765-768.

Greenland, Sander; James Robins. Invited commentary: ecologic studies – biases, misconceptions, and counterexamples. *Am J Epidemiol* 1994;139:747-60.

Greenland, Sander; James Robins. Accepting the limits of ecologic studies. *Am J Epidemiol* 1994;139:769-771.

Greenland, Sander. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987;125:761-768.

Koopman, James S.; Longini, Ira M. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *AJPH* 1994;84:836-842]

Piantadosi, Steven. Invited commentary: ecologic biases. *Am J Epidemiol* 1994;139:761-764.

Poole, Charles. Ecologic analysis as outlook and method. Editorial. *AJPH* 1994;84(5):715-716)

Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *AJPH* May 1994;84:819-824.

Susser, Mervyn. The logic in ecological: I. The logic of analysis. *AJPH* 1994;84:825-829. II. The logic of design. *AJPH* May 1994;84:830-835.

Cohort studies:

Liddell FDK. The development of cohort studies in epidemiology: a review. *J Clin Epidemiol* 1988; 41(12):1217-1237.

Samet, Jonathan M., Alvaro Muñoz. Cohort studies. *Epidemiologic Reviews* 1998; 20(1).

Case-control studies:

Armenian, Hartoune K. Applications of the case-control method. *Epidemiologic Reviews* 1994;16(1):1-164.

Cornfield, Jerome. A method of estimating comparative rates from clinical data. *JNCI* 1951;11:1269-1275. (A classic)

Feinstein, Alvan R. Experimental requirements and scientific principles in case-control studies. Editorial. *J Chron Dis* 1985; 38:127-134.

Gordis, Leon. Should dead cases be matched to dead controls? *Am J Epidemiol* 1982; 115:1-5.

Greenberg RS, Ibrahim MA. The case-control study. In: *Textbook of Public Health.* Holland W, Detels R, and Knox EG (ed). NY, Oxford, 1983.

Greenland, Sander. Control-initiated case-control studies. *Int J Epidemiol* 1985; 14:130-134.

Greenland, Sander; Duncan C. Thomas. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982; 116:547-53.

Greenland, Sander; Duncan C. Thomas, Hal Morgenstern. The rare-disease assumption revisited: a critique of "Estimators of relative risk for case-control studies". *Am J Epidemiol* 1986; 124:869-876.

Hogue, Carol J. R., David W. Gaylor, Kenneth F. Shulz. Estimators of relative risk for case-control studies. *Am J Epidemiol* 1983; 118:396-408.

Hogue, Carol J. R.; David W. Gaylor, Kenneth F. Schultz. The case-exposure study: a further explication and response to a critique. *Am J Epidemiol* 1986; 124:877-883.

Horwitz, RI, Feinstein AR. Methodologic standards and contradictory results in case-control research. *Am J Med* 66:556-564, 1979.

Horwitz, Ralph I. The experimental paradigm and observational studies of cause-effect relationships in clinical medicine. *J Chron Dis* 1987; 40:91-99.

Horwitz, Ralph I., and Alvan R. Feinstein. The application of therapeutic-trial principles to improve the design of epidemiologic research. *J Chron Dis* 1981; 34:575-583.

Kramer, Michael S.; Jean-Francois Boivin. Toward an "unconfounded" classification of epidemiologic research design. *J Chron Dis* 1987; 40:683-688.

Mantel, N. and Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. *J National Cancer Institute* 22:719-748, 1959. Read 719-732 only [Read past page 732 at your own risk]). (Another classic)

McLaughlin, Joseph K.; William J. Blot, Eric S. Mehl, Jack S. Mandel. Problems in the use of dead controls in case-control studies. I. General results. *Am J Epidemiol* 1985; 121:131-139. II. Effect of excluding certain causes of death. *Am J Epidemiol* 1985; 122:485-94.

Miettinen, Olli. Estimability and estimation in case-referent studies. *Am J Epidemiol* 1976; 103: 226-235.

Newman, Stephen C. Odds ratio estimation in a steady-state population. *J Clin Epidemiol* 1988; 41:59-65 (advanced).

Pearce, Neil; Harvey Checkoway. Case-control studies using other diseases as controls: problems of excluding exposure-related diseases. *Am J Epidemiol* 1988; 127:851-6 (advanced).

Poole, Charles. Exposure opportunity in case-control studies. *Am J Epidemiol* 1986; 123:352-358.

Poole, Charles. Critical appraisal of the exposure-potential restriction rule. *Am J Epidemiol* 1987; 125:179-183.

Rodrigues L, Kirkwood BR. Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. *Intl J Epidemiol* 1990;19:206-213.

Schlesselman, James J.; Bruce V. Stadel. Exposure opportunity in epidemiologic studies. *Am J Epidemiol* 1987; 125:174-178.

Stavraky, Kathleen M.; Aileen Clarke. Hospital or population controls? An unanswered question. *J Chron Dis* 1983; 36:301-308, and responses on pages 309-316.

Wacholder, Sholom; Joseph K. McLaughlin, Debra T. Silverman, Jack S. Mandel. Selection of controls in case-control studies. I. Principles. *Am J Epidemiol* 1992;135(9):1019-1028

Wacholder, Sholom; Debra T. Silverman, Joseph K. McLaughlin, Jack S. Mandel. Selection of controls in case-control studies. II. Types of controls. *Am J Epidemiol* 1992;135(9):1029-1041

Wacholder, Sholom; Debra T. Silverman, Joseph K. McLaughlin, Jack S. Mandel. Selection of controls in case-control studies. III. Design options. *Am J Epidemiol* 1992;135(9):1042-1050

Wingo, Phyllis A.; Howard W. Ory, Peter M. Layde, Nancy C. Lee, and the Cancer and Steroid Hormone Study Group. The evaluation of the data collection process for a multicenter, population-based, case-control design. *Am J Epidemiol* 1988; 128:206-217.

Plus several articles and comments in *J Chron Dis* 1985; 38(7) and responses in 1986; 39(7):565-571.

Intervention trials:

Chalmers, Thomas C.; Paul Celano, Henry S. Sacks, Harry Smith, Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983; 309:1358-61.

Farquhar, John W.; Stephen P. Fortmann, June A. Flora et al. Effects of communitywide education on cardiovascular disease risk factors: The Stanford Five-City Project. *JAMA* 1990;264:359-365.

Feinstein, A.R. *Clinical Epidemiology: The Architecture of Clinical Research*. Chapter 29, pp. 683-718.

Feinstein, Alvan R. Current problems and future challenges in randomized clinical trials. *Circulation* 1984; 70:767-774.

Friedman L.M., Furberg, C.D., DeMets, D.L. *Fundamentals of clinical trials*. Boston: John Wright PSG Inc., 1982.

Koepsell TD et al. Symposium on community trials: Invited commentary, 594-599 plus articles by Allan Donner, Stephen Fortmann, Sylvan Green. *Am J Epidemiol* 1995;142(6)

Lavori, Philip W.; Thomas A. Louis, John C. Bailar III, Marcia Polansky. Designs for experiments–parallel comparisons of treatments. *N Engl J Med* 1983; 309:1291-8.

Lilienfeld, Abraham M. *Ceteris Paribus*: The evolution of the clinical trial. *Bulletin of the History of Medicine* 56:1-18, 1982.

Luepker R, et al. Community education for cardiovascular disease prevention: Risk factor change in the Minnesota Heart Health Program. *Am J Public Health* 1994;84:1383-1393.

MacMahon and Pugh - *Epidemiology: principles and methods*, Chapter 13; Mausner and Kramer - *Epidemiology: an introductory text*, Chapter 8; Lilienfeld and Lilienfeld - *Fundamentals of epidemiology*, Chapters 10-11.

Moses, Lincoln E. The series of consecutive cases as a device for assessing outcomes of intervention. *N Engl J Med* 1984; 311:705-710

Peto, R., Pike, M.C., Armitage, P. et al. Design and analysis of randomized clinial trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 34:585-612, 1976.

Peto, R., Pike, M.C., Armitage, P., et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 35:1-39, 1977.

Sacks, H., Chalmers, T.C., Smith, H. Randomized versus historical controls for clinical trials. *Amer J Med* 72:233-239, 1982.

Schwartz, D., Lellouch, J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis* 20:637-648, 1967.

Simon, Richard and Robert E. Wittes. Methodologic guidelines for reports of clinical trials. Editorial. *Cancer Treatment Reports* 1985; 69:1-3.

Weinstein, Milton C. Allocation of subjects in medical experiments. *N Engl J Med* 291:1278-1285, 1974.

Analytic study designs - Assignment

Part I

The following questions are based on the attached paper "Cancer and Tobacco Smoking" (Morton L. Levin, Hyman Goldstein, and Paul R. Gerhardt, *JAMA*, May 27, 1950, pages 336-338).

1. What study design has been employed (circle one)

   A. case-control
   B. Cohort (prospective)
   C. retrospective cohort (or historical cohort)
   D. ecologic
   E. May be either a or c

   In one or two sentences justify your choice.

2. The comparison group was composed of "those with symptoms referable to the same site but which proved not to be due to cancer." Briefly discuss the appropriateness of this comparison group and its likely effect on the association observed in the study. (3-6 sentences).

3. Histories of tobacco usage were obtained routinely from all admitted patients before the final diagnosis had been established. "This procedure is considered especially important from the standpoint of excluding bias." The authors are referring to bias from:

   A. selective survival
   B. cohort effect
   C. antecedant-consequent confusion
   D. ecologic fallacy
   E. selective recall

   Explain.

4. The authors state that "There were more than twice as many cases of lung cancer among cigarette smokers as among any other group". Briefly discuss the meaningfulness of this statement.

5. Which interpretation of the data in Table 4 is correct? [Watch out--this is tricky.]

   A. Cumulative incidence of lung cancer in cigarette smokers* is approximately 2 1/2 times the cumulative incidence in nonsmokers (20.7/8.6).

   B. Incidence density rate of lung cancer in cigarette smokers* is approximately 2 1/2 times that in nonsmokers (20.7/8.6).

   C. Prevalence of lung cancer among all cigarette smokers* admitted to Roswell Park Memorial Institute is approximately 2 1/2 times that among all nonsmokers admitted to Roswell Park Memorial Institute (20.7/8.6).

   D. Prevalence of cigarette smoking* among lung cancer cases is approximately 2 1/2 times the prevalence of nonsmoking among lung cancer cases (20.7/8.6).

   E. None of the above statements is correct.

   In one or two sentences justify your answer.

_____

*Note: 25 years' duration or longer.

Note: The two articles for this assignment were included the packet of copyrighted material sold at the Health Affairs Bookstore. The page numbers jump to reflect the insertion of those pages.

Part II

The next five questions are based on the attached paper "The Mortality of doctors in relation to their smoking habits", (Richard Doll, Austin Bradford Hill, *British Medical Journal*, June 26, 1954, pages 1451-1455).

The questions are designed to direct your thoughts in reviewing the study. Your answers can be brief! [It is suggested that you read through the entire paper before answering the questions.]

1. Doll and Hill state (page 1451, column 1) that "Further retrospective studies of that same kind would seem to us unlikely to advance our knowledge materially or to throw any new light upon the names of the association."

   a. What specific design does "that same kind" refer to?

   b. Why do the authors regard it necessary that there be studies employing a different design (a "prospective" one). What are some advantages of a prospective, rather than a retrospective, approach?

2. What study design have Doll and Hill in fact employed (check one):

   A. case control with incident cases
   B. cohort (prospective)
   C. cohort (historical
   D. cross-sectional
   E. ecologic
   F. case control nested within a cohort

   In one or two sentences justify your choice.

3.

   a. The study group was recruited from members of the medical profession in the United Kingdom in 1951. Briefly, discuss the appropriateness of using this target population. What problems would one anticipate in recruiting and following these subjects? How might these problems be minimized?

   b. In reading the Methods Section, what drawbacks of a prospective investigation of this nature are vividly apparent?

4.

   a. Think about the approach to measuring smoking status in the Levin et al. study in Part I and the Doll and Hill study. Are you more confident in the validity of either approach? What is the "real" exposure factor or factors?

   b. Lung cancer is a disease presumed to have a long latency period. How do Doll and Hill justify their use of current smoking habits as the exposure variable?

5. In contrast to many earlier investigations, Doll and Hill studied deaths rather than diagnosed cases. Cite some advantages and disadvantages of using deaths rather than incident cases.

6. What is the reason given for standardizing rates by age (p.ÿ1452)? Do you agree?

7. What is the reason for concern about the criteria upon which the diagnosis of lung cancer was based (page 1452, 2nd column)?

8. Read the first paragraph under "Method of Smoking" (page 1453) carefully. We will take up this concept (nondifferential misclassification bias) later in the course. [No response required.]

9. Are the results from the "prospective" study basically consistent with those of the retrospective studies? Cite several factors that might account for the lower rates in the present study. [Thought question: how were the authors able to estimate death rates from their retrospective study?]

10. Read the paragraph headed "The Diagnoses" (page 1454) carefully. We will consider this concept ("Detection bias") later in the course. [No response required]

11. Doll and Hill end their conclusion: "It seems clear that smoking cannot be a major factor in [the] production [of deaths attributable to coronary thrombosis], but the steady increase in mortality with the amount of tobacco smoking recorded suggests that there is a subgroup of these cases in which tobacco has a significant adjuvant effect," (p.1455). Was that "data dredging" or prescience? [No need to write, just think.]

12. How would you, speaking on behalf of Doll and HIll, respond to the argument that people who have never smoked, or who have stopped smoking, are often more health conscious, eat better diets, get more exercise, use more preventive health services, and handle stress better than people who smoke, and that these reasons, rather than smoking, may be responsible for the results observed in the study?

Analytic study designs - Assignment solutions

Part I

1. A case control study design was employed. Subjects were selected on the basis of their having or not having the disease of interest (i.e., selected cancers). A strength of this particular case control study, however, is that exposure information had been obtained before the diagnosis was known.

2. The case-control design has a number of subtleties, and a full understanding of the central issue of control selection came only after several decades of case-control studies. The principle for selection of the control group is that it should permit estimation of the exposure distribution in the study base, i.e., the population from which the cases arose. However, previously it was thought that, on analogy with an experimental comparison, the cases and controls should be "comparable" or "as similar as possible except for their disease status".

   Belief in the importance of comparability between cases and controls has probably been the rationale for control selection procedures in many published studies and perhaps also for the widespread use of matched controls. There is a role for matching in selection of controls, primarily where a known strong risk factor is prominent in the cases but much less frequent in the study base. In this situation, most cases will have the known risk factor (e.g., smoking in a study of lung cancer) but most controls in an unmatched control group will not. That difference will make it difficult or at least inefficient to compare the exposure distributions of cases and controls while taking account of the known risk factor. Selecting the control group in a way that forces it to have the same distribution of the established risk factor as will be found in the case group avoids this problem. However, if the matching criterion is not a strong risk factor, then there is the danger that a matched control group will have an exposure distribution that is more like that of the case group than is the exposure distribution in the study base, a situation referred to as "over-matching".

   There are other considerations that enter into the selection procedure for controls, however, including feasibility, cost, anticipated response rate, and availability of exposure information. For example, people selected from the general population may not be able to provide their exposure status (e.g., if measuring it involves a costly or invasive procedure, such as coronary angiography, or access to records they will not have, such as workplace exposure to various chemicals). Also, if the condition under study is not rare but requires a costly or invasive procedure to detect, a control group randomly selected from the general population could include within it a meaningful proportion of people with the condition. So the actual selection procedure often represents a compromise between the objective of constructing a window into the study base and constructing a window that is clear enough to see through, even if one knows that the view through it may not quite be the study base. Thus, even though they are not fully representative of the population from which cases arose, hospital controls may be more motivated and available than community controls and may provide opportunities for detection of exposure more similar to those for cases.

In this study the use of hospitalized controls produces an overestimate of the prevalence of smoking in the population from which cases arose. The reason is that many of the controls are symptomatic for diseases associated with smoking. Thus, this control group has a greater proportion of smokers than would a general population sample and the contrast with the lung cancer case group is therefore diminished, resulting in a negative bias in the measure of association, similar to Berkson's or referral bias. (Note: this is *not* what was meant by "Briefly"!)

3. E. Selective Recall - Cases generally have an increased motivation or tendency to recollect exposure to a putative causal agent. To avoid bias, it is helpful if controls have a similar stimulus to remember and report.

4. This statement is imprecise in the present context. In a case control study, such as this, one directly estimates and compares the proportion of exposed (smokers) among the cases to the proportion of exposed among the noncases, not the proportion of diseased among the exposed and unexposed as implied in this statement. The author's statement would be more appropriate to a cohort design or to most cross-sectional studies.

5. "E" - none of the statements is correct: Statement "C" comes the closest to describing the data presented in Table 4, though its usefulness is doubtful in a case control study. Interpretation of findings would be more appropriate in terms of the proportion exposed among the cases and controls. [Statement "C" is not actually correct because Table 4 does not deal with all smokers and nonsmokers admitted to Roswell Park, but only with those with certain cancers or symptoms referable to those same sites (see first paragraph of second column.]

Part II

1. A. case control

   - Exposure variable can be measured more reliably, with no danger of its measurement being influenced by the disease outcome;
   - Selection bias in the choice of a control group is avoided, while selection bias from loss to follow-up can often be controlled or at least its likelihood guaged;
   - The temporal sequence (i.e., that exposure precedes disease) is clearly demonstrated;
   - Risk and rates can be estimated directly.

2. B. Cohort -- The investigators measured initial exposure status (by survey) before death (and ideally before disease onset) and subsequent mortality experience by surveillance of death records.

3. a.

   - Better informed than general population
   - Motivated to respond to medical research

- Possible motivation to change behavior
- Busy schedule may lead to poor response rate from causes other than lung cancer
- Better medical care may decrease mortality from causes other than lung cancer.

Techniques to minimize problems of survey:

- Keep questionnaire brief
- Do not reveal research hypothesis
- Follow-up interviews with non-responders

b.

- Large sample required and lengthy follow-up, with attendant expense, effort, and delay in obtaining results
- Difficult to assess prior exposure history
- Poor response rate from initial survey
- Loss to follow-up.

4. a. Since both involve questionnaire responses, the differences are in: the timing of the survey and the quantification of exposure. Doll and Hill might be favored for less selective recall, whereas Levin et al. might be selected for a more thorough inquiry with more precise quantification of exposure. The "real" exposure factor is presumably some constituent or combination of constituents of tobacco smoke. The "true" agents could vary with type of cigarette, smoking practices, cigarette brand contents, and so on. (Cigarettes smoke contains thousands of substances, including many that are known carcinogens.)

   b. The assumption is that current smoking practice, which is more reliably measured, is a good indicator of previous smoking practice. A prior case-control study showed current smoking history was almost as great a risk factor for lung cancer as total smoking history.

5. Advantages:

- Routine notification is available;
- Easy access to records;
- Necropsy confirmation available in many cases.

Disadvantages:

- Differential survival may influence the findings, though this should be less of a problem with a rapidly fatal disease like lung cancer;
- Cause of death may be inaccurately recorded, particularly where multiple conditions are present;

- Follow-up must be longer for any given number of cases.

6. The distribution of smoking habits varies considerably with age, as does cancer incidence. Therefore, age adjustment or age-specific comparisons are needed.

7. To demonstrate that disease diagnosis was not based not only on clinical judgment but documented by examination of tissue specimens. Particularly with a rare disease, like lung cancer, false positives can readily distort the observed association, leading to information (misclassification) bias.

9. Yes: Prevalent cases may have failed to respond initially, and therefore been deleted from the cohort; short period of follow-up of disease with long induction; doctors may receive better medical care, thereby reducing mortality.

12. Ultimately, it is not possible to exclude all possible alternate explanations except by conducting an experiment in which smoking (or stopping smoking), can be randomly allocated in sufficiently large groups with sufficient adherence to the experimental regimen. For ethical and practical reasons, such an experiment cannot be conducted with smoking onset as the experimental variable. Smoking cessation programs have not been sufficiently effective to serve as an alternate intervention. If they were, it is doubtful that they could be used in a lung cancer trial, since the evidence for the harmful effects of smoking is so abundant that ethical considerations would probably preclude such a trial.

    Though "absolute" proof requires human experiments, the weight of observational and animal evidence is convincing. Any particular factor proposed as an alternate explanation (e.g., exercise, diet) can be examined simultaneously with smoking, to see which factor is associated with lung cancer when the other factor is controlled. There are no serious contenders. Traditional criteria for causal inference are well satisfied:

    a. strong association (RR about 8 for smokers overall);
    b. dose-response effect exists;
    c. replication in many studies;
    d. cohort studies demonstrate that exposure precedes disease;
    e. biologic explanation (animal and tissue culture models, autopsy studies);
    f. experimental confirmation available in animal models;
    g. analogy to other carcinogen-cancer associations.

    Smoking and lung cancer is one of the strongest of epidemiologically-established relationships. Controversy remains, but that should not paralyze policy planning (see M.A. Ibrahim, The cigarette smoking/ lung cancer hypothesis. Editorial. *Am J Public Health* 1976; 66:131-132). Nevertheless, the debate has continued even into recent years (see references in Bibliography).

9. Causal inference*

The desire to act on the results of epidemiologic studies frequently encounters vexing difficulties in obtaining definitive guides for action. Weighing epidemiologic evidence in forming judgments about causation.

> "The world is richer in associations than meanings, and it is the part of wisdom to differentiate the two." — John Barth, novelist.
>
> "Who knows, asked Robert Browning, but the world may end tonight? True, but on available evidence most of us make ready to commute on the 8.30 next day." — A. B. Hill

Historical perspective

Our understanding of causation is so much a part of our daily lives that it is easy to forget that the nature of causation is a central topic in the philosophy of science and that in particular, concepts of disease causation have changed dramatically over time. In our research and clinical practice, we largely act with confidence that 21st century science has liberated us from the misconceptions of the past, and that the truths of today will lead us surely to the truths of tomorrow. However, it is a useful corrective to observe that we are not the first generation to have thought this way.

In the 1950s, the middle of what will soon become the last century, medical and other scientists had achieved such progress that, according to Dubos (page 163), most clinicians, public health officers, epidemiologists, and microbiologists could proclaim that the conquest of infectious diseases had been achieved. Deans and faculties began the practice of appointing, as chairs of medical microbiology, biochemists and geneticists who were not interested in the mechanisms of infectious processes. As infectious disease epidemiology continues to surge in popularity, we can only shake our heads in disbelief at the shortsightedness of medical and public health institutions in dismantling their capabilities to study and control infectious diseases, whose epidemics have repeatedly decimated populations and even changed the course of history.

On the other hand, perhaps the connection is not so direct. According to Dubos, the 19th-century fall in death rates from infectious disease and malnutrition actually began in mid-century, several decades before the medical discoveries of the scientific era could be turned into actual policies. Medical science and the germ theory received an inordinate share of credit because the decline was not widely recognized until the end of the century. Moreover, he charges

> The present generation [presumably the pre-World War II birth cohorts going back to 1910] goes still further and now believes that the control of infectious and nutritional disease dates from the widespread use of antibacterial drugs and from the availability of vitamins and processed foods. So short and parochial are our memories!" (page 365)

* (An earlier version of these notes was prepared by Sandra Martin, Ph.D.)

While acknowledging the very important roles played by local boards of health and other municipal bodies, Dubos attributes most of the improvement in health to improvements in prosperity and transportation that enabled many people to affort "at least one square meal a day":

> No medical discovery made during recent decades can compare in pratcial importance with the introduction of social and economic decency in the life of the average man. The greatest advances in the health of the people were probably the indirect results of better housing and working conditions, the general availability of soap, of linen for underclothing, of glass for windows, and the humanitarian concerns for higher living standards. (page 365)

Before proceeding to our investigation of causal inference, it will be helpful to take a brief look at the history of public health and disease in the 17th-19th centuries. [The following account comes primarily from Wilson Smillie (*Public health: its promise for the future*), Mervyn Susser (*Causal thinking in the health sciences*), and Lisa Berkman and Lester Breslow (*Health and ways of living*).]

In the early seventeenth century, medical science was "just emerging from the morass of the Middle Ages" (Smillie, 1955:18). The most deadly disease of the American colonies in that century was smallpox. The disease was even more devastating to the indigenous populations of the New World and is believed to have killed over half the Indian population in Mexico following the Spanish conquest (Smillie, 1955:21). In Europe, smallpox was an endemic disease of childhood; but in the more isolated situation in the colonies, recurrent epidemics devastated settlements. According to Smillie (p22), a 1633 smallpox epidemic in the Massachusetts Bay colony spread to the Indians and killed "whole plantations" of them. A 1689-1690 epidemic in New England killed 1,000 people in one year (by comparison, Boston had a total population of about 7,000) at that time.

During the eighteenth century, the practice of smallpox inoculation (published by the Greek Timonius in 1714) successfully aborted epidemics in the American colonies, although the practice was at first resisted. Smallpox inoculation was banned outright in New York City in 1747, required the governor's permission in Carolina in 1764, and required the consent of the selectmen in towns in Massachusetts (Smillie, 1955, p28).

Nevertheless, smallpox inoculation soon proved its worth. At the beginning of the Revolutionary War, in 1776, smallpox arrived in Boston. A heroic campaign inoculated 9,152 nonimmune people in three days. Although the inoculations produced 8,114 smallpox cases resulting in 165 deaths (1.8%), the 232 natural infections in susceptible persons who had not been inoculated resulted in 33 deaths (14.2%) (data from Shattuck, Lemuel, reported in Smillie, 1955:29). Two decades later, Edward Jenner, an obscure country practitioner in England, demonstrated immunity to smallpox in ten persons who had previously developed cowpox. Although his paper to the Royal Society was refused, he published his classic monograph in 1798 to become known as the father of vaccination. (Take-home message: don't let a manuscript rejection discourage you!)

The second great communicable disease in eighteenth century North America, yellow fever, also took a fearsome toll on communities in the New World. The first American article on yellow fever (by Dr. John Lining of Charleston) described the disease as both contagious and imported (Smillie,

1955:35). Quarantine of sick persons and of ships suspected of having yellow fever on board was sometimes instituted to prevent or abort epidemics.

Towards the end of the eighteenth century, though, the miasma theory of disease arose – the theory that all disease was due to bad air – contaminations (miasma) emanating from a great variety of sources (Smillie, 1955:3). So strong was the power of this new theory, that Dr. Benjamin Rush, the greatest American physician of that time, was sure that the great 1793 yellow fever epidemic in Philadelphia was neither contagious nor had come from the West Indies, but was rather due to a pile of spoiled coffee thrown on a wharf (Smillie, 1955:9).

By the early nineteenth century, medicine and the public health movement were dominated by the miasma theory (Susser, *Causal Thinking in the Health Sciences*). The line of investigation was to prove the ill-effects of miasma; the line of prevention was to eliminate the sources of miasma in slums and poor sanitation. Although the concept of miasma, overthrown later in the century, is ridiculed today, the sanitation measures that the miasma theory called for were often dramatically effective in reducing death rates. During the nineteenth century, as Susser writes, Jacob Henle formulated the conditions that needed to be met to prove the germ theory, and some 20 years later, Louis Pasteur demonstrated the existence of microorganisms. Now, the causes of disease could actually be seen – microbiology had progressed from a science of inference to a science of direct observation. Microorganisms then became the object of the search for causes. The containment of the spread of microbes became the object of prevention. Asepsis, antisepsis, and disinfection - measures taken on the basis of germ theory - also proved effective. Moreover, the new paradigm proved superior to the miasma theory through its greater specificity and in its ability to explain and predict certain phenomena outside the miasma theory, such as immunization and chemotherapy.

The discovery of microorganisms and the ascendence of the germ theory of disease brought with them the view that illness consisted of many discrete clinical entities, each caused by a different agent, and each with certain morbid manifestations yielding distinct syndromes (Berkman and Breslow, 1983:5). This concept prevails even today, as illustrated in the dictionary definitions shown in the chapter on the Phenomenon of Disease. The search for specific agents has led to great breakthroughs in medicine and public health, such as the effective control of many infectious diseases in the developed world and the worldwide eradication of smallpox. Even where the germ theory did not apply, as in the case of vitamin deficiency diseases, the concept of specificity of cause has also proved effective for etiology and control.

There were those who resisted the one-disease-one-cause model. But the tide was against them. As Dubos (1965, quoted in Berkman and Breslow, p. 6) observed:

> These vague arguments were no match for the precise experimentation by which Pasteur, Koch, and their followers defended the doctrine of specific causation of disease. Experimental science triumphed over the clinical art, and within a decade the theory of specific etiology of disease was all but universally accepted, soon becoming, as we have seen, the dominant force in medicine.

At the same time, this great spurt in medical research diminished awareness of the rarity of one-to-one relationships and of the complex relationships between causes and effect that exist in the real

world. Even as late as the 1950's, for example, it was very difficult to conceptualize that smoking can cause so many diseases (noted by the late Ernst Wynder in a 1997 seminar at UNC at Chapel Hill), and the fact that so many diseases were associated with cigarette smoking was put forward as an argument against interpeting the associations as causal. (According to Sir Richard Doll, in 1992 the Sloan-Kettering Institute's new director "told Wynder that his conclusiion that a causal relationship existed between smoking and lung cancer was irresponsible and that all future publications by his goup would have to be cleared through the director's office", Ernst Wynder, 1923-1999, *AJPH* 1999;89:1798-9: 1799).

Lord Bertrand Russell has written, "every advance in a science takes us further away from the crude uniformities which are first observed into a greater differentiation of antecedent and consequent and into a continually wider circle of antecedents recognized as relevant (*Mysticism and Logic*, London: Longmans, Green, 1918, p. 188, quoted in E.H. Carr, *What is history*, NY: Knopf, 1963, p. 118). A number of developments undermined the supremacy of the one-cause-one-disease model.

One was the growing predominance of microbial diseases of endogenous origin, diseases caused by organisms that are carried by many people in the population (Dubos, 176). Contempory examples are bacterial infections secondary to acute viral illnesses, opportunistic infections in persons with AIDS, and urinary tract infections with *E. coli*. A second was the recognition that many pathogens, including the tubercle bacillus, can be carried for long periods of time, only to cause disease when the host's immunity becomes weakened. A third was the shift of attention from infectious diseases to heart disease and cancer, where various factors are related to risk but none absolutely necessary; thus the term "multifactorial" disease. (Although CHD is the classic multifactorial disease, there have been recent suggestions that infectious processes may be an important dimension.) Finally, as epidemiology has expanded to study behavioral and environmental maladies (e.g., automobile injuries, alcoholism, homicide, and unprotected intercourse), a unicausal model does not even have meaning.

Nevertheless, in practice much of epidemiology focuses on single risk factors. Ideally we could make use of an overall model combining multiple etiologic agents into a comprehensive system. But often epidemiologic research has its greatest role in stages of investigation before a comprehensive causal picture is possible. Indeed, epidemiologic studies are one of the primary avenues towards beginning to define the factors that might make up such a picture. So the usual approach is to take one or two suspected factors at a time and then see if, taking into account what has already been discovered about the disease, the suspected factors increase the explanatory or predictive power of the investigation. This one factor-at-a-time approach is the essence of risk factor epidemiology, of the concepts confounding and effect modification to be presented later, and of epidemiologic approaches to causal inference.

The concept of causality

In *Modern Epidemiology*, Rothman and Greenland illustrate the process of understanding a cause with a description of a toddler learning that moving a light switch causes the light to turn on. But what we take as a cause depends upon the level at which we seek understanding or the constituency we represent. Thus:

The **mother** who replaced the burned-out light bulb may see her action as the cause for the light's turning on, not that she denies the effect of the light switch but has her focus elsewhere.

The **electrician** who has just replaced a defective circuit breaker may cite that as the cause of the light's turning on, not that he denies the importance of the switch and the bulb, but his focus is elsewhere still.

The **lineman** who repaired the transformer that was disabled by lightning may regard his repair as the cause of the light's turning on.

The **social service agency** that arranged to pay the electricity bill may regard that payment as the cause of the light's turning on, since with the electricity cut off, neither the switch nor the circuit breaker matters.

The **power company**, the **political authority** awarding the franchise, the investment bankers who raised the financing, the **Federal Reserve** that eased interest rates, the **politician** who cut taxes, and the **health care providers** who contributed to the toddler's safe birth and healthy development might all cite their actions as the real cause of the light's turning on.

The National Rifle Association's slogan "Guns don't kill people; people kill people" is not a public health stance, but it does illustrate the complexities of apportioning causation.

Mervyn Susser proposes that for epidemiologists a causal relation has the following attributes: association, time order, and direction. A cause is something that is associated with its effect, is present before or at least at the same time as its effect, and acts on its effect. In principle, a cause can be **necessary** – without it the effect will not occur – and/or **sufficient** – with it the effect will result regardless of the presence or absence of other factors. In practice, however, it is nearly always possible to conceive of other factors whose absence or presence could avert an effect since, as with the light switch example above, assumptions are always present. A fall from a five story building would appear to be a sufficient cause of death. But it could be argued that death would not have resulted had there been a safety net below!

Rothman has elaborated a component causes model that attempts to accomodate the multiplicity of factors that contribute to the occurrence of an outcome. In his model, a sufficient cause is represented by a complete circle (a "causal pie"), the segments of which represent component causes. When all of the component causes are present, then the sufficient cause is complete and the outcome occurs. There may be more than one sufficient cause (i.e., circle) of the outcome, so that the outcome can occur through multiple pathways. A component cause that is a part of every sufficient cause is a necessary cause. The induction period for an event is defined in relation to each particular component cause, as the time required for the remaining component causes to come into existence. Thus, the last component cause has an induction period of zero. This model is useful for illustrating a number of epidemiologic concepts, particularly in relation to "synergism" and "effect modification", and we will return to it in a later chapter.

Causal Inference

Direct observation vs. inference:

Much scientific knowledge is gained through direct observation. The introduction of new technology for observation along optical, aural, and chemical dimensions of perception, through such tools as microscopes, x-rays, ultrasound, magnetic resonance scans, and biochemical assays has greatly expanded our opportunities for direct observation and contributed to major advances in scientific knowledge. For example, a recent Nobel Prize was awarded for measuring ion channels in cells, a process that previously had to be inferred. With direct observation, it is possible to "see" causation, especially if one can manipulate the process. Thus, it has been said that the advances in molecular biological techniques have been converting the science of genetics from one of inference to one of direct observation.

In general, however, challenges to understanding transcend that which can be observed directly, so that inference is an essential aspect of scientific activity. It is typically not possible to observe all aspects of a phenomenon of interest, and this situation is very much the case for relationships under epidemiologic investigation. Moreover, even observation involves inference.

Consider the difficulties that arise from latency and induction. The rapidity with which scurvy improved after Lind began his treatments was a great aid in recognizing the effect of lemons. The two-week induction period of measles and its infectiousness before the appearance of symptoms must at one time have been a barrier to understanding its transmission. At the time of Goldberger's investigations, pellagra typically developed about four months after the onset of a niacin-deficient diet. The longer induction period must have made it that much more difficult to associate cause with effect. For example, an interval of four months confounded the seasonality, so that cases were higher in spring and summer (when food was becoming more available) than in winter (when the disease was really developing). At times, opponents of the acceptance of the causal relationship between tobacco and lung cancer have pointed to low rates of lung cancer in populations with high smoking rates (for example, American women in the 1950's) as contradictory evidence, neglecting to take into account the long interval between the onset of cigarette smoking and the development of lung cancer.

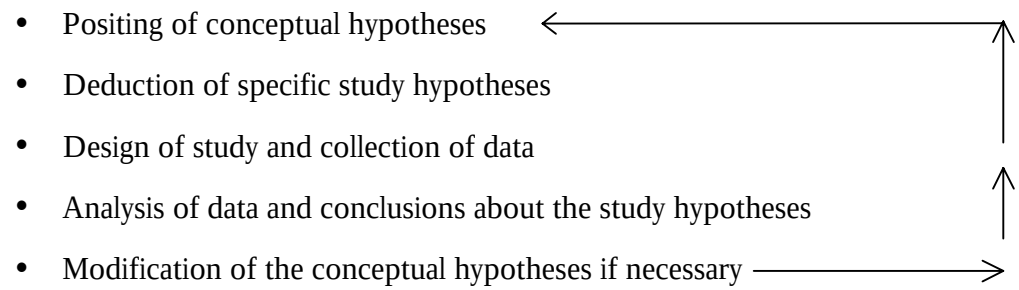
Similarly, rare diseases require observation of many subjects, greatly restricting the level of detail that can be visualized or examined. Severe constraints on measurement are also imposed by the need to rely largely on noninvasive measurement methods.

Idealized view of the scientific process

For reasons such as these, a primary recourse in epidemiology is to inference, through:

- positing of conceptual models (conceptual hypotheses);
- deduction of specific, operational hypotheses; and
- testing of operational hypotheses.

As presented in Kleinbaum, Kupper, and Morgenstern, the cycle of scientific progress proceeds as follows:



This admittedly idealized portrait appropriately emphasizes the importance of conceptual models. As the distinguished historian Edward Hallett Carr has written (*What is history*, NY: Knopf, 1968, p. 136) "The world of the historian, like the world of the scientist, is not a photographic copy of the real world, but rather a working model which enables him more or less effectively to understand it and to master it. The historian distils from the experience of the past, or from so much of the experience of the past as is accessible to him, that part which he recognizes as amenable to rational explanation and interpretation, and from it draws conclusions which may serve as a guide to action. A recent popular writer, speaking of the achievements of science, refers graphically to the processes of the human mind which, 'rummaging in the ragbag of observed 'facts,' selects, pieces, and patterns the relevant observed facts together, rejecting the irrelevant, until it has sewn together a logical and rational quilt of "knowledge"' (Leslie Paul: *The annihilation of man*. London: Faber & Faber, 1944, p. 147)."

Carr continues, in a passage that applies much more broadly than to historical reasoning alone, "History therefore is a process of selection in terms of historical significance. To borrow Talcott Parson's phrase once more, history is 'a selective system' not only of cognitive but of causal orientations to reality. Just as from the infinite ocean of facts the historian selects those which are significant for his purpose, so from the multiplicity of sequences of cause and effect he extracts those, and only those, which are historically significant; and the standard of historical significance is his ability to fit them into his pattern of rational explanation and interpretation. Other sequences of cause and effect have to be rejected as accidental, not because the relation between cause and effect is different, but because the sequence itself is irrelevant. The historian can do nothing with it; it is not amenable to rational interpretation, and has no meaning either for the past or the present." (E.H. Carr, op.cit., p. 138). Thus in a hypothetical situation Carr (p. 137) presents in which Jones, driving from a party where he has drunk too much, in a car whose brakes are defective, at an intersection with poor visibility runs down and kills Robinson, who was crossing the road to buy cigarettes, we would entertain alcohol, defective brakes, and poor visibility as causes (and potential targets for preventive action), but not cigarette smoking even though it is true that had Robinson not been a cigarette smoker he would not have been killed that evening.

Conceptual hypotheses arise from inductive reasoning, based on available observations and theory, analogies to known processes, and so forth. For example, the effects of passive smoking on lung cancer and of oral contraceptives on breast cancer were first posited based on knowledge of the effects of active smoking on lung cancer and of oral contraceptives on estrogen-sensitive tissues. Existing knowledge may be compatible with more than one alternative model. For example, existing

data on the effects of radiation on cancer risk are compatible with a linear relationship, in which there is no threshold below which risk is absent, or with a curvilinear model in which a threshold of risk exists.

From these conceptual hypotheses, deductive reasoning can generate specific predictions or study hypotheses which ought to be true if the conceptual model is correct. If these predictions or study hypotheses are incompatible with valid data from empirical studies, then the conceptual model that gave rise to the predictions is called into question. This situation forces a re-appraisal or modification of the conceptual hypotheses and lays the basis for advancing understanding.

Karl Popper – power of falsification:

This aspect of the process of scientific investigation has been emphasized by the philosopher Karl Popper. In Popper's conceptualization, falsification of a hypothesis appears to be more informative than corroboration of a hypothesis. There could be innumerable data sets that are consistent with a false hypothesis. A single counter example, however, forces a modification. Therefore, in Popper's view, studies should attempt to refute, rather than to confirm, hypotheses being entertained. A hypothesis that has survived numerous attempts to refute it gains in strength more than one that has merely been corroborated repeatedly.

Although Popper's model is appealing, how well does it describe how science actually proceeds? One problem with this orderly process of induction-deduction-testing requires a large body of knowledge from which to conceptualize and deduce. Particularly in the early stages of research in an area, there is typically a need for descriptive investigations to generate a body of data that can give some direction to thinking about the issues and provides some basis for inductive reasoning. More serious is the fact that in epidemiologic research, a negative result (finding of no association) often cannot refute the original hypothesis because of the many sources of bias that work towards masking underlying associations.

A further point at which the orderly progression outlined above is inadequate is the situation in which the existing conceptual models have been found wanting yet no new ones have been advanced to break through the stalemate. In physics, for example, Einstein's theory of relativity – a revolutionary reconceptualization of physical phenomena – broke through an impasse that had been reached toward the latter part of the 19th century, and opened the way for dramatic advances in knowledge. Goldberger's investigations of pellagra provide a less dramatic but important illustration of the role of a reconceptualization in studying a specific disease. So it is important to bear in mind that advances in knowledge can come from careful observation, precise description, and creative thinking – though in many cases this thinking proceeds through the implicit positing of hypothesis and testing them against available knowledge. Indeed, even the process of direct observation involves paradigms that guide our observation and interpretation.

According to D.C. Stove ("Karl Popper & the Jazz Age"), Popper's philosophy of science can be understood only in reference to the social circumstances of its origins (Vienna in the years after the First World War). In Stove's view, Popper's philosophy is based on reversal of traditional notions of science and philosophy. Traditionally, propositions in science are verifiable. For Popper, they are distinguished by being falsifiable. The method of science has been regarded as essentially inductive.

Popper maintains that it is fundamentally deductive. To many, the essence of science is caution; Popper says that audacity is the essence of science. Science was supposed to be distinguished from guesswork and everyday opinion by the fact that its conclusions are certain or at least have a vast preponderance of probability in their favor; Popper would say that scientific conclusions are never more than guesswork, hypotheses, conjectures, and that no theory ever becomes more probable. For historical reasons, according to Stove, Popper's philosophy of science received broad acceptance by the public and the scientific community. Particularly in epidemiology, where it is impossible to control many sources of extraneous influence, the possibility that a true relationship will be obscured makes it hard to refute an epidemiologic hypothesis and therefore limits the applicability of Popper's model. (See the Bibliography for other points of view.)

"Common sense":

An alternative model of scientific progress is that of "common sense", a phenomenon of increasing interest to researchers in artificial intelligence. Consider the following situation (Judea Pearl, Cognitive Systems Laboratory, UCLA, as described in M. Mitchell Waldrop, "Causality, Structure, and Common Sense", *Science* 11 September 1987; 237:1297-1299):

> You go outside in the morning and notice the grass is wet.
>
> The obvious inference is that it rained during the night.

However, suppose you now learn that someone left the lawn sprinkler on during the night. Suddenly your confidence in the rain goes down considerably – upon receiving a new fact, you withdraw your original conclusion.

According to presentations at the American Association for Artificial Intelligence in July 1987 (recounted in Waldrop's article), this kind of logical flip-flop ("nonmonotonic reasoning" in the artificial intelligence community) is the epitome of common sense. It is also a blatant violation of the conventional theory of logic (based on axioms, theorems, proof of theorems). But it is typical of the kind of judgment under uncertainty that characterizes both human experts and computer-based expert systems. In common sense, causes compete, evidence cooperates. The more clues we have to support a given hypothesis, the more confident we are that the hypothesis is true.

Statistical inference and causal inference

Statistical inference is not the same as causal inference, though there is a parallelism in the inferential process itself, and statistical inference is generally employed in evaluating the data for use in causal inference. In statistical inference, data from a sample of observations are used to make inferences about the population from which they are assumed to derive. A statistical model, expressed in a null hypothesis (H_0), is "tested" against data. Based on the data, the statistical model is either accepted or rejected as an adequate explanation of the data. Rejection is a stronger statement and is usually based on a more stringent criterion (a 5% significance level means that results as strong as those observed would occur by chance only 5% of the time, whereas a typical 80% level of statistical power means that a real relationship will not appear to be "significant" 20% of the time).

But excluding an explanation based on chance does not establish causality, since there are many other possible noncausal reasons for an association to exist. The association could conceivably reflect some peculiarities of the study group, problems with the measurement of disease or exposures, or the effects of some other factor that might affect both the disease AND the putative cause. In fact, the putative risk factor may have occurred AFTER (even as a result of) the disease. In causal inference, one examines the structure and results of many investigations in an attempt to assess and, if possible, eliminate all possible noncausal reasons for observed associations.

Influence of knowledge and paradigms

Since causal inference is a process of reasoning, it is conditioned by what is believed to be true and by prevailing concepts of disease. These concepts are based on knowledge of the time, as well as on ignorance and erroneous beliefs.

Consider the case of microbial agents. The Henle-Koch Postulates (1884) for implicating a bacteria as the cause of a disease held:

1. The parasite (the original term) must be present in all who have the disease;
2. The parasite can never occur in healthy persons;
3. The parasite can be isolated, cultured and capable of passing the disease to others

have been a useful model for diseases such as anthrax, tuberculosis, and tetanus. But these postulates are not adequate for many other diseases, especially viral diseases, because of (Rivers, 1937; Evans 1978):

1. Disease production may require co-factors.
2. Viruses cannot be cultured like bacteria because viruses need living cells in which to grow.
3. Pathogenic viruses can be present without clinical disease (subclinical infections, carrier states).

When pathogens are not so toxic or virulent that their presence always brings disease, then we need to consider multiple factors and a "web" of causation.

Criteria for causal inference in epidemiology

Criteria for causal inference became an issue of importance and controversy with the establishment of the first Advisory Committee to the Surgeon General on the Health Consequences of Smoking. In its 1964 report, the Committee presented a list of "epidemiologic criteria for causality" which Sir Austin Bradford Hill subsequently elaborated in his classic 1965 Presidential Address to the newly formed Section of Occupational Medicine of the Royal Society (Hill AB. The environment and disease: association or causation? *Proc Royal Soc Medicine* 1965;58:295-300). Hill's criteria are widely recognized as a basis for inferring causality.

The basic underlying questions are:

1. Is the association real or artefactual?
2. Is the association secondary to a "real" cause?

The Bradford Hill criteria

1. Strength of the association – The stronger an association, the less it could merely reflect the influence of some other etiologic factor(s). This criterion includes consideration of the statistical precision (minimal influence of chance) and methodologic rigor of the existing studies with respect to bias (selection, information, and confounding).

2. Consistency – replication of the findings by different investigators, at different times, in different places, with different methods and the ability to convincingly explain different results.

3. Specificity of the association – There is an inherent relationship between specificity and strength in the sense that the more accurately defined the disease and exposure, the stronger the observed relationship should be. But the fact that one agent contributes to multiple diseases is not evidence against its role in any one disease.

4. Temporality – the ability to establish that the putative cause in fact preceded in time the presumed effect.

5. Biological gradient – incremental change in disease rates in conjunction with corresponding changes in exposure. The verification of a dose-response relationship consistent with the hypothesized conceptual model.

6. Plausibility – we are much readier to accept the case for a relationship that is consistent with our general knowledge and beliefs. Obviously this tendency has pitfalls, but commonsense often serves us.

7. Coherence – how well do all the observations fit with the hypothesized model to form a coherent picture?

8. Experiment – the demonstration that under controlled conditions changing the exposure causes a change in the outcome is of great value, some would say indispensable, for inferring causality.

9. Analogy – we are readier to accept arguments that resemble others we accept.

Strength of the association

- Pronounced excess of disease associated with the exposure.
- The magnitude of the ratio of incidence in the exposed to incidence in the unexposed.
- How strong is "strong"? A rule-of-thumb:

Relative risk	"Meaning"
1.1-1.3	Weak
1.4-1.7	Modest
1.8-3.0	Moderate
3-8	Strong
8-16	Very strong
16-40	Dramatic
40+	Overwhelming

Strong associations are less likely to be the result of other etiologic factors than are weak associations.

> Egs., Smoking and lung cancer; smoking and CHD.

Consistency

The association has been "repeatedly observed by different persons, in different places, circumstances and times." Consistency helps to guard against associations arising out of error or artifact. But consistently observed results are not necessarily free of bias, especially across a small number of studies, and results in different populations may differ if a causal relationships is influenced by the presence or absence of modifying variables.

Specificity

The relationship between exposure and disease is specific in various ways – a specific disease is linked with a specific exposure, specific types of exposure are more effective, etc. There is an intimate relationship between specificity and strength in the sense that the more accurately defined the disease and exposure, the stronger the observed relative risk should be.

> e.g., Schildkraut and Thompson (*Am J Epidemiol* 1988; 128:456) reasoned that the familial aggregation they observed for ovarian cancer was unlikely to be due to family information bias because of the specificity of the relationship in that case-control differences in family history (a) involved malignant but not borderline disease and (b) were greater for ovarian than for other cancers.

But the fact that one agent contributes to multiple diseases is not evidence against its role in any one disease. For example, cigarette smoke causes many diseases.

Temporality

First exposure, then disease.

It is sometimes difficult to document sequence, especially if there is a long lag between the exposure and the disease, subclinical disease, exposure (e.g., a treatment) brought on by an early manifestation of the disease).

Biological gradient

The verification of a dose-response relationship consistent with the hypothesized conceptual model.

Need to consider threshold and saturation effects, characteristics of the exposure.

[See Noel Weiss, Inferring causal relationships, *Am J Epidemiol* 1981; 113:487.]

Plausibility

Does the association make sense biologically.

> E.g.s, estrogen and endometrial cancer, estrogens and breast cancer, oral contraceptives and breast cancer

Coherence

Does a causal interpretation fit with known facts of the natural history and biology of the disease, including knowledge about the distributions of the exposure and disease (by person, place, time) and the results of laboratory experiments. Do all the "pieces fit into place"?

(For an exquisite example of the evaluation of coherence, see Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernest L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Nat Cancer Inst* 1959;22:173–203.)

Experimental evidence

Certain types of study designs may provide more convincing evidence than other types of study designs. Intervention studies can provide the strongest support, especially when the exposure can be randomly assigned. Since it is unethical and/or impractical to assign many of the exposures that epidemiologists study. One possible alternative is to remove the exposure and see if the disease decreases, unless the causal process is regarded as irreversible.

> E.g.s, pellagra, scurvy, HDFP, LRC-CPPT, MRFIT.

Analogy

Have there been similar situations in the past? (e.g., rubella, thalidomide during pregnancy)

Except for temporality, no criterion is absolute, since causal associations can be weak, relatively nonspecific, inconsistently observed, and in conflict with prevailing biological understanding. But each criterion that is met strengthens our assurance in reaching a judgment of causality. [See also Hill's comments on tests of statistical significance.]

Several of his criteria (for example, coherence, biological gradient, specificity, and perhaps strength) may be reformulated in terms of a more general issue of consistency of observed data with a hypothesized etiological (usually biological) model. For example, a biological gradient need not be monotonic, as in the case of high doses of radiation which may lead to cell-killing and therefore a lower probability of tumor development. Similarly, specificity applies in certain situations but not others, depending upon the pathophysiologic processes hypothesized.

Search for Cause versus Decision-making

Causal inference is of fundamental importance for advancing scientific knowledge. The Popperian stance is that in an ultimate sense, every theory is tentative. Any theory can potentially be overthrown by incompatible data that cannot themselves be called into question. So in the view of many, scientific knowledge advances through concerted attempts to refute existing theories.

In considering issues in causal inference in epidemiology, though, it is useful to draw a distinction between inference aimed at establishing etiology and inference aimed at reaching a decision to act or not to act. The Popperian stance has less applicability in causal inference in support of decision-making, because of the importance of timely action. Even though individual and collective decisions are often made based on considerations other than scientific knowledge, and even without any valid causal data, causal inference is fundamental for decision-making. Moreover, judgments of causality – ultimately by governmental authorities and the public at large – are a critical basis for the resolution of controversial issues, e.g., restrictions on products such as tobacco, saccharin, coffee, oral contraceptives, handguns; pollution controls, etc. Those moved to action can cite Hill's words:

> All scientific work is incomplete - whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not

confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time.

A.B. Hill, The environment and causation, p. 300

Parallel concepts in epidemiologic inference and the legal process

One can draw an interesting analogy between the process of decision-making in epidemiology and in the legal process. In both processes, a decision about facts must be reached on the evidence available. In the absence of revealed truth (e.g., mathematical proof), both approaches emphasize integrity of the process of gathering and presenting information, adequate representation of contending views, rules of evidence, standards of certainty for various potential consequences. Both areas emphasize procedural (methodological) safeguards, since the facts in a given situation are generally established only as the findings of an adequate investigatory process. Similarly, it is important for both epidemiology and law not only that justice (i.e., proper procedures/methodology) be done but also that it be seen to be done. In law, pattern jury instructions provide a basis for the jury to use in weighing evidence. Similarly, epidemiology has criteria for causal inference.

The legal rules of evidence offer several parallels to the epidemiologic approach to weighing evidence and inferring causality. In both systems, reliability of the information (data) is a prime rationale. Some examples are:

- The Hearsay Rule: evidence is not admissible if based on hearsay rather than direct observation.

  Example: If the doctor testifies that the patient said he was driving on the wrong side of the road, that testimony is hearsay evidence and therefore not admissible. The doctor did not see the patient driving on the wrong side of the road.

  There are exceptions: official government sources, business records obtained in the regular course of business (without an eye to a lawsuit), other records routinely made are admissible in evidence.

- Dead man's statute: testimony about conversation with person who is now deceased is not admissible (because he/she cannot respond).

In both law and epidemiology, there is a relationship between the seriousness of the action and the degree of evidence required for that action. Some examples concerning searches, seizures and judgments:

- To issue a search warrant, the magistrate must find that there is a reasonable suspicion that the object of the search will be found.
- To issue an arrest warrant, the magistrate must find that there is probable cause that the person committed the crime.
- For a police officer to arrest an individual without a warrant, he must have reasonable cause to believe that a crime may be imminent or just committed.

- To issue an indictment, the grand jury must find that there is a prima facie case that the individual did commit the crime.
- For a decision against the defendant in a civil suit, the judge or jury must find a "preponderance of the evidence".
- To convict the defendant in a criminal trial, the jury must find that the evidence establishes his/her guilt "beyond a reasonable doubt".
- For a verdict of guilt based entirely on circumstantial evidence, the jury must be satisfied that every reasonable hypothesis has been excluded except guilt. (If there is some real evidence, the requirement is not so strict.)

(In Scotland, there is a verdict of "not proved", which certainly has parallels in epidemiologic "judgments".)

In both law and epidemiology, the facts in any individual case always factor importantly into the decision, and the decision is generally influenced by considerations of:

- How imperative is it to act?
- How imminent is a possible harm?
- How serious is the potential harm?

It is generally better to err on the side of safety (though in law that's kept implicit, never given as explicit reason).

Bibliography

Buck, Carol. Popper's philosophy for epidemiologists. *Intl J Epidemiol* 1975; 4:159-168.

Evans AS. Causation and disease: a chronological journey. The Thomas Parran Lecture. *Am J Epidemiol* 1978;108:249-258. See also his book of the same title, NY, Plenum, 1993.

Faust, David; Jay Ziskin. The expert witness in psychology and psychiatry. *Science* 1988; 241:31-35.

Ginzburg, Harold M. Use and misuse of epidemiologic data in the courtroom: defining the limits of inferential and particularistic evidence in mass tort litigation. *American Journal of Law & Medicine* 1986; 12 (3-4):423-439.

Hill, Austin Bradford. The environment and disease: association or causation? *Proceedings Royal Society Medicine* 1965;58:295-300. (a classic)

Jacobsen, M. Against Popperized epidemiology. *Intl J Epidemiol* 1976; 5:9-11.

Lave, Lester B.; E.P. Seskin. Epidemiology, causality and public policy. *American Scientist* 1979; 67:178-180.

Maclure, Malcolm. Popperian refutation in epidemiology. *Am J Epidemiol* 1985; 121:343-350.

Pearce, Neil; Douglas Crawford-Brown. Critical discussion in epidemiology: problems with the Popperian approach. *J Clin Epidemiol* 1989 42(3):201-208. Response by Carol Buck: 185-188.

Research Methodology: Strenghening Causal Interpretations of Nonexperimental Data (DHHS Publication No. (PHS)90-3454) - single copies available from the Agency for Health Care Policy and Research. Contains the proceedings of an AHCPR-sponsored (HS05306) conference.

Rothman and Greenland, *Modern Epidemiology*; Rothman, *Modern Epidemiology*. Ch 2.

Rothman, Kenneth, Stephen F. Lanes (eds). *Causal inference*. Chestnut Hill, MA, Epidemiology Resources Inc. W 61 C374 1988. (Includes four papers from the 1985 Society for Epidemiologic Research symposium, with solicited written criticism and rejoinders).

Schlesselman, James J. "Proof" of cause and effect in epidemiologic studies. *Preventive Medicine* 1987; 16:195-210 (from a Workshop on Guidelines to the Epidemiology of Weak Associations)

Schlesinger, George. *The intelligibility of nature*. New Jersey, Humanities Press, 1985.

Stove, David. *Popper and after*. Oxford, 1983.

Susser, Mervyn. Judgment and causal inference. Criteria in epidemiologic studies. *Am J Epidemiol* 105:1-15, 1973.

Susser, Mervyn. What is a cause and how do we know one? *Am J Epidemiol* 1991; 133:635-648.

Susser, Mervyn. The logic of Sir Karl Popper and the practice of epidemiology. *Am J Epidemiol* 1986; 124:711-718.

Weed, Douglas L. On the logic of causal inference. *Am J Epidemiol* 1986; 123:965-979 (and correspondence in 1987; 126:155-158.

Weiss, Noel S. Inferring causal relationships. *Am J Epidemiol* 1981; 113:487-

10. Sources of error

A systematic framework for identifying potential sources and impact of distortion in observational studies, with approaches to maintaining validity

We have already considered many sources of error in epidemiologic studies: selective survival, selective recall, incorrect classification of subjects with regard to their disease and/or exposure status. Because of the limited opportunity for experimental controls, error, particularly "bias", is an overriding concern of epidemiologists (and of our critics!) as well as the principal basis for doubting or disputing the results of epidemiologic investigations.

Accuracy is a general term denoting the absence of error of all kinds. In one modern conceptual framework (Rothman and Greenland), the overall goal of an epidemiologic study is accuracy in measurement of a parameter, such as the IDR relating an exposure to an outcome. Sources of error in measurement are classified as either random or systematic (Rothman, p. 78).

Rothman defines **random error** as "that part of our experience that we cannot predict" (p. 78). From a statistical perspective, random error can also be conceptualized as sampling variability. Even when a formal sampling procedure is not involved, as in, for example, a single measurement of blood pressure on one individual, the single measurement can be regarded as an observation from the set of all possible values for that individual or as an observation of the true value plus an observation from a random process representing instrument and situational factors. The inverse of random error is **precision**, which is therefore a desirable attribute of measurement and estimation.

Systematic error, or **bias**, is a difference between an observed value and the true value due to all causes other than sampling variability (Mausner and Bahn, 1st ed., p. 139). Systematic error can arise from innumerable sources, including factors involved in the choice or recruitment of a study population and factors involved in the definition and measurement of study variables. The inverse of bias is validity, also a desirable attribute.

These various terms – "systematic error", "bias", "validity" – are used by various disciplines and in various contexts, with similar but not identical meanings. In statistics, "bias" refers to the difference between the average value of an estimator, computed over multiple random samples, and the true value of the parameter which it seeks to estimate. In psychometrics, "validity" most often refers to the degree to which a measurement instrument measures the construct that it is supposed to measure. The distinction between random and systematic error is found in many disciplines, but as we shall see these two types of error are not wholly separate. We will return to the issue of terminology below, in the section on "Concepts and terminology".)

Precision

The presence of random variation must always be kept in mind in designing studies and in interpreting data. Generally speaking, small numbers lead to imprecise estimates. Therefore, small differences based on small numbers must be regarded with caution since these differences are as likely the product of random variation as of something interpretable.

Estimates of ratio measures (e.g., the relative risk) based on sparse data are very susceptible to instability. For example, a relative risk of 5.0 based on the occurrence of 4 cases among the nonexposed becomes twice as large (10.0) if two unexposed cases are missed through the vagaries of sampling, measurement, missing data, or other reasons. If three are missed, the relative risk will be 20.

Example to illustrate the concept of precision

Consider the following data, from Table 4 of Hulka et al., "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen, *Am J Epidemiol* 112:376-387, 1980:

Effect of duration of estrogen use on relative risks [age-adjusted] using three control groups among white women, North Carolina, 1970-76

Duration of use	No. of Cases	D&C Controls		Gynecol. Controls		Community Controls	
		No.	RR	No.	RR	No.	RR
None used	125	136		118		172	
Less than 6 months	8	13	0.7	12	0.7	20	0.8
6 months - 3.5 yrs.	9	14	0.7	9	0.9	21	0.7
3.5 yrs. - 6.5 yrs.	9	16	0.8	1		7	1.7
6.5 yrs. - 9.5 yrs.	9	11	1.2	2	3.8	5	2.5
More than 9.5 yrs	19	10	2.0	2	5.1	4	5.5

Note: "D&C Controls" = dilatation and curetage patients as controls
"Gyn Controls" = other gynecology clinic patients as controls

First, let's recall from our previous topic that since these data come from a case-control study, the "relative risks" in the table are odds ratios. Since the disease is a rare one, however, odds ratios, risk ratios, and incidence density ratios will all be about the same. Also from that lesson we should be able to reformulate the above data as a series of 2 x 2 tables if for some reason we wished to. Such reformulation would make it easier to see how to calculate crude relative risk estimates (OR's) from

the data in the table. (Note that the OR's in the table are age-adjusted, through a mathematical modeling procedure called multiple logistic regression, so our crude OR's will differ in some cases.)

Notice that the relative risk estimates for the GYN and community controls in the longer duration categories are based on very few controls. For example, if two of the community controls classified as duration of use "3.5 years - 6.5 years" were instead duration "6.5 years - 9.5 years", then the relative risk estimates would be reversed and the consistent dose-response picture would no longer appear. [For the moment we are ignoring the age adjustment, though for these two particular duration groups in the community controls the adjusted OR's are the same as if they were calculated from the data in the table.] Similarly, the RR's over 5 for the longest duration subjects are based on 2 and 4 controls in the GYN and community groups, respectively. On the other hand, the fact that similar results were found in two control groups strengthens the assessment that a dose-response relationship truly exists, rather than being a chance finding.

Quantifying the degree of precision or imprecision – confidence intervals

Statistical techniques such as standard errors and confidence inteverals are used to quantify the degree of precision or imprecision of estimates; there are also rules of thumb (e.g., see Alvan Feinstein, *J Chron Dis* 1987; 40:189-192). A **confidence interval** provides a range of values that is expected to include the true value of the parameter being estimated. The narrower the confidence interval, the more precise the estimate.

For example, suppose we are estimating the relative risk for endometrial cancer in women who have used replacement estrogens for 3.5 years to 6.5 years (compared to women who have not taken estrogens) and that the "true" (but unknown) relative risk is 1.5. Suppose also that the estimate we obtain from the data for community controls in the above table are unbiased, though they do reflect random error. Hulka et al. computed the age-adjusted value as 1.7 (the crude value is very similar: 1.77). The 1.7 is a point estimate and provides our best single estimate of the (unknown) true relative risk of 1.5.

We do not expect that our estimate is exactly correct, however, so we also compute an interval estimate, or confidence interval as an indicator of how much data were available for the estimate. Suppose that the 95% confidence interval is (0.6,4.7). The interpretation would be that 1.7 is the best single estimate of the (unknown) true relative risk and that there is "95% confidence that the true relative risk is somewhere between 0.6 and 4.7". "Confidence" does not mean the same thing as "probability". In this case "95% confidence" means that we obtained the confidence limits 0.6 and 4.7 through a procedure that yields an interval that will contain the true value in 95% of instances in which we use it and will not contain the true value in the remaining 5% of instances.. Loosely speaking, a 95% confidence interval of 0.6-4.7 means that the observed value of 1.7 is "compatible", by conventional usage, with true relative risks anywhere between 0.6 and 4.7 inclusive.

Another way of describing the meaning of "compatible" is the following. The limits 0.6 and 4.7 are obtained from the point estimate of 1.7 and the estimated standard error of that estimate. The estimated standard error is a function of the size of the numbers (i.e., the amount of data) on which

the point estimate is based and thus a measure of its imprecision. A 95% confidence interval (0.6,4.7) means that if our study had yielded a point estimate anywhere in that interval, the 95% confidence interval around that point estimate would contain the value 1.7. In that sense the observed value of 1.7 is compatible with true relative risks anywhere between 0.6 and 4.7.]

In fact, the original table in Hulka et al. did include confidence intervals for the odds ratios. Attention to confidence intervals or to sparseness of data is an important aspect of interpreting results.

Reducing random variation (increasing precision)

Confidence intervals and other procedures for assessing the potential for random variation in a study do not increase precision, but merely quantify it. The major strategies for reducing the role of random error are:

1. Increase sample size – a larger sample, other things being equal, will yield more precise estimates of population parameters;
2. Improve sampling procedures – a more refined sampling strategy, e.g., stratified random sampling combined with the appropriate analytic techniques can often reduce sampling variability compared to simple random sampling;
3. Reduce measurement variability by using strict measurement protocols, better instrumentation, or averages of multiple measurements.
4. Use more statistically efficient analytic methods – statistical procedures vary in their efficiency, i.e., in the degree of precision obtainable from a given sample size;

Bias

Bias is by definition not affected by sample size. Rather, bias depends on enrollment and retention of study participants and on measurement. [A technical definition of "bias" in its epidemiologic usage (based on Kleinbaum, Kupper, and Morgenstern) is the extent to which an estimate differs from the true value of the parameter being estimated, even after sample size is increased to the point where random variation is negligible. This definition is based on consistency; in statistics, an estimator is **consistent** if its value comes progressively closer to the value of the parameter it estimates as the sample size increases.]

Concepts and terminology

In the area of bias and validity, as in so many other areas that cross disciplines, terminology can be a significant source of confusion. Such dangers are particularly apparent when the terms are also used in a nontechnical sense in ordinary discourse. An additional source of confusion for terminology concerning validity is overlap among the concepts. For example, measurements are an ingredient of studies, but studies can also be regarded as measurement procedures applied to populations or

associations. So the same terms may be applied to individual measurements and to entire studies, though the meaning changes with the context.

Internal validity

Epidemiologists distinguish between internal validity and external validity. **Internal validity** refers to absence of systematic error that causes the study findings (parameter estimates) to differ from the true values as defined in the study objectives. Systematic error can result from inaccurate measurements of study variables, nonuniform recruitment or retention of study participants, or comparisons of groups that differ in unknown but important characteristics. Thus, internal validity concerns bias in estimates for the target population specified in the study objectives.

External validity

External validity refers to the extent to which a study's findings apply to populations other than the one that was being investigate. Generalizability to populations beyond the target population for which the study was designed and/or beyond the circumstances implicit in the study is a matter of scientific inference, rather than a technical or statistical question (see Rothman and Greenland). Therefore external validity is probably better considered in relation to causal inference and interpretation of study results. (Rothman and Greenland regard "external validity" as a misnomer, preferring to draw the distinction between validity and generalizability.)

Validity pertains to a specific measure

Since different types of errors affect specific findings in different ways, validity must generally be discussed in regard to a specific measure or measures. A study aimed at testing an etiologic hypothesis typically seeks to estimate strength of association measured as the ratio or difference of incidences in different groups. Lack of internal validity in this context means inaccuracy (bias) in these estimates. In fact, Kleinbaum, Kupper, and Morgenstern (*Epidemiologic Research*, ch. 10) define (internal) validity and bias in terms of systematic distortion in the "measure of effect". A study can yield a valid (unbiased) measure of effect despite systematic errors in the data if the errors happen to offset one another in respect to the measure of effect. However, a study with no systematic error can yield a biased estimate of a measure of effect (for example, due to random variability in an important measurement – see appendix). Much of the methodologic writing about bias concerns distortion in effect measures.

Not all studies have as their objective the estimation of a measure of effect, and even studies that do also report estimates of other parameters (e.g., incidence rates, prevalences, means). Thus, even if the measure of effect is accurately estimated, the possibility and extent of bias in other measures must be considered.

Measurement validity

In Rothman and Greenland's perspective, measurement is the purpose of all studies, so the concept of validity of measurement is the same as that of validity. However, validity of the **measurements** carried out in conducting a study raises issues of its own and is addressed in another category of

methodological literature. Validity of measurement (I have to confess that this is my own term to differentiate this type of validity) concerns the avoidance of systematic error in measuring or detecting a factor (e.g. blood pressure, smoking rate, alcoholism, HIV infection). The sociologic and psychologic literature deals extensively with measurement validity, particularly in relation to data collected via questionnaires and interviews. Cognitive psychology studies the thinking processes by which study participants decode questionnaire items and retrieve the information from memory (e.g., Warnecke *et al*., 1997a; Warnecke *et al*., 1997b). Psychometrics studies statistical aspects of psychological measurement instruments (Nunnally, 1994). These disciplines are especially pertinent for epidemiologists interested in sophisticated measurement of self-report measures.

Direction of bias – "which way is up"

Concepts and terminology can also complicate descriptions of the direction in which a bias may distort a measure of effect. The sources of confusion are: (1) an association can be positive (RR > 1.0) or inverse (RR < 1.0, also referred to as "negative"), (2) a source of bias can make a measure of effect increase in magnitude, decrease in magnitude, move towards 1.0 from either above or below, and move away from 1.0 in either direction, and (3) it is easy to lose sight of whether the measurement of association being referred to is that observed in the study or the "true" one that exists in the target population. [Try plotting some relative risks on a line as you read the next two paragraphs.]

Describing the direction of bias – example:

Suppose that "aggressive" people are more likely to survive an acute myocardial infarction (MI) than are nonaggressive people. A case-control study of MI that recruits its cases from among (live) hospitalized MI patients will therefore overrepresent aggressive MI cases, since proportionately more of them will live long enough to enroll in the study. If this is the only source of systematic error, then we expect the observed relative risk (RR) to be greater than the true relative risk for incidence of acute MI (since the true relative risk would include the victims who died before they could be enrolled in the study). The direction of bias is in the positive direction (toward higher values of the RR), regardless of whether the true RR is greater than 1.0 (i.e., aggressive people also more likely to have an MI) or less than 1.0 (aggressive people are less likely to have an MI).

In contrast, uniform random error in the measurement of aggressiveness independent of other variables typically moves the observed RR "toward the null" (closer to 1.0 than the true RR). Bias toward the null can produce a lower observed RR (if the true RR is greater than 1.0) or a higher observed RR (if the true RR is less than 1.0), but not an RR that is farther from the null than the true RR. On the other hand, the bias from greater survival of aggressive MI cases in the above hypothetical case-control study will be closer to 1.0 only if the true RR is less than 1.0 and farther from 1.0 only if the true RR is greater than 1.0.

For these reasons we need four terms to characterize the potential effects of sources of bias:

"**Positive bias**" – The observed measure of effect is a larger number than the true measure of effect is (if it could be known);

"**Negative bias**" – The observed measure of effect is a smaller number than the true measure of effect is (if it could be known);

"**Towards the null**" – The observed measure of effect is closer to 1.0 than the true measure of effect is (if it could be known);

"**Away from the null**" – The observed measure of effect is farther from 1.0 than the true measure of effect is (if it could be known);

Another way of describing the direction of bias is to say that the observed measure of effect overestimates (underestimates) the true measure. With this phraseology, however, more information must be available, since "overestimates" could be taken as meaning higher in numerical value or greater in strength (farther from the null).

In the interests of precise communication, we will try to adhere to the above usage, which does not appear to be standard in the profession. However, terminology is only one source of confusion. Consider the longstanding proposition that nondifferential misclassification (covered below) of a dichotomous exposure or disease variable, in the absence of confunding (see next chapter) always produces bias that is "toward the null". This proposition holds as long as nondifferential (independent) misclassification is no worse than what would result from classification each observation by tossing a coin. However, extreme nondifferential misclassification (in the limiting case, misclassification of every participant), however, can bias the measure of effect beyond and then away from the null value.

Types of bias

Students of epidemiology often wish for a catalog of types of bias in order to be able to spot them in published studies. David Sackett (Bias in analytic research. *J Chron Dis* 32:51-63, 1979) once attempted to develop one. Nine sample entries he describes are:

1. Prevalence-incidence (Neyman) bias

   This is Sackett's term for, among other things, selective survival. Also included are the phenomena of reversion to normal of signs of previous clinical events (e.g., "silent" MI's may leave no clear electrocardiographic evidence some time later) and/or risk factor change after a pathophysiologic process has been initiated (e.g., a Type A may change his behavior after an MI), so that studies based on prevalence will produce a distorted picture of what has happened in terms of incidence.

2. Admission rate (Berkson) bias

   Where cases and/or controls are recruited from among hospital patients, the characteristics of both of these groups will be influenced by hospital admission rates.

3. Unmasking (detection signal) bias

   Since by necessity, a disease must be detected in order to be counted, factors that influence

disease detection may be mistakenly thought to influence disease occurrence. This possibility is particularly likely where the disease detection process takes place outside of the study (e.g., in a case-control study), where the disease has an occult, or asymptomatic, phase, and where the exposure leads to symptoms that induce the individual to seek medical attention.

4. Non-respondent bias

   Non-respondents to a survey often differ in important ways from respondents. Similarly, volunteers often differ from non-volunteers, late-respondents from early respondents, and study dropouts from those who complete the study.

5. Membership bias

   Membership in a group may imply a degree of health which differs systematically from others in the general population. For example, the observation that vigorous physical activity protects against CHD was initially thought likely to be a result of fitter people (with lower innate CHD risk) being more likely to engage in vigorous activity. Another example would be if people who participate in a health promotion program subsequently make more beneficial lifestyle changes than nonparticipants due not to the program itself but to the participants' motivation and readiness to change.

6. Diagnostic suspicion bias

   The diagnostic process includes a great deal of room for judgment. If knowledge of the exposure or related factors influences the intensity and outcome of the diagnostic process, then exposed cases have a greater (or lesser) chance of becoming diagnosed, and therefore, counted.

7. Exposure suspicion bias

   Knowledge of disease status may influence the intensity and outcome of a search for exposure to the putative cause.

8. Recall bias

   Recall of cases and controls may differ both in amount and in accuracy (selective recall). Cases may be questioned more intensively than controls.

9. Family information bias

   Within a family, the flow of information about exposures and illnesses is stimulated by, and directed to, a family member who develops the disease. Thus a person who develops rheumatoid arthritis may well be more likely than his or her unaffected siblings to know that a parent has a history of arthritis.

The appendix to Sackett's article gives his entire catalog of biases.

Classifying sources of bias

Inspite of David Sackett's initiative, a complete catalog of biases does not yet exist. Instead, following Olli Miettinen's work in the 1970's, epidemiologists generally refer to three major classes of bias:

1. **Selection bias** – distortion that results from the processes by which subjects are selected into the study population:
2. **Information bias** (also called **misclassification bias**) – distortion that results from inaccuracies in the measurement of subject characteristics, and incorrect classification therefrom:
3. **Confounding bias** – distortion in the interpretation of findings due to failure to take into account the effects of disease risk factors other than the exposure of interest.

Confounding bias is somewhat different from the other two forms in that the actual data collected by the study may themselves be correct; the problem arises from a misattribution of observed effects (or their absence), i.e., an apparent effect is attributed to the exposure of interest, whereas in fact it ought to have been attributed to some other factor. We will discuss confounding in the following chapter.

Of course, as in so many other areas of epidemiology, the divisions among the classes are only relative, not absolute!

Selection bias

Ignoring the questions of random error in sampling (i.e., assuming that all samples are large enough so that random variation due to sampling is negligible), we can see that if the process by which subjects are recruited favors or overlooks certain types of subjects, then the study population we obtain will not be representative of the population for which we are attempting to obtain estimates. For example, if we are studying characteristics of persons with diabetes and obtain all of our subjects from among hospital patients, the characteristics of this study population will yield a distorted or biased estimate of the characteristics of diabetics in general.

In case-control studies, situations that can produce selection bias include:

- the exposure has some influence on the process of case ascertainment ("detection bias"): the exposure prevalence in cases will be biased;
- selective survival or selective migration – the exposure prevalence in prevalent cases may be biased compared to that in incident cases;
- the exposure has some influence on the process by which controls are selected (e.g., use of chronic bronchitis patients as controls for a study of lung cancer and smoking): the exposure prevalence in controls will differ from that in the base population.

In cohort studies, the primary source of selection bias is generally differential attrition or loss to follow-up. Example (hypothetical):

Complete cohort:

	Type A	Type B	
CHD	40	20	
CHD	160	180	
Total	200	200	RR=2.0

Observed cohort:*

	Type A	Type B	
CHD	32	18	
CHD	144	162	
Total	176	180	RR=1.82

*based on a 10% loss rate among subjects, except that Type A subjects who developed CHD are assumed to have been lost at a 20% rate. If all subjects, including the CHD/Type A group had experienced a 10% loss rate, the incidence in each behavior type group, and therefore the risk ratio, would be undistorted.

Conceptual framework

[After Kleinbaum, Kupper and Morgenstern, *Epidemiologic Research* and *Am J Epidemiol* article on selection bias (see bibliography)].

External population: the population of ultimate interest, but which we are not attempting to study directly – e.g., we may wish to study the relationship between hypertension and stroke in general, but study only subjects in North Carolina, recognizing that generalizing to other areas will require consideration of differences between North Carolina and those other areas. We will not concern ourselves with generalizability in this chapter.

Target population: the population for which we intend to make estimates.

Actual population: the population to which our estimates actually apply. This population may not be obvious or even knowable.

Study population: the group of participants for whom we have collected data. In Kleinbaum, Kupper, and Morgenstern's framework, the study population is regarded as an unbiased sample of the actual population, differing from it only from through unsystematic sampling variability error.

The study population is a subset of the actual population. Bias is the discrepancy between the actual and target populations. Generalizability deals with inference from the target population to an external population (see previous page).

In thinking about selection bias and its potential effect on study results, we find it useful to consider the probabilities according to which people in the target population could gain access to the actual population. These probabilities are called (population) selection probabilities.

For simplicity, consider a dichotomous disease and dichotomous exposure classification, and let the fourfold table in the target population and actual population be as follows:

	E	$\bar{E}$
D	A	B
$\bar{D}$	C	D

Target

	E	$\bar{E}$
D	A^o	B^o
$\bar{D}$	C^o	D^o

Actual

We can then define four selection probabilities:

alpha (α) = (A^o/A) the probability that a person in cell A (in the target population) will be selected into the actual population from which the study population is a random sample

beta (β) = (B^o/B) the probability that a person in cell B (in the target population) will be selected into the actual population

gamma (γ) = (C^o/C) the probability that a person in cell C (in the target population) will be selected into the actual population

delta (δ) = (D^o/D) the probability that a person in cell D (in the target population) will be selected into the actual population

Example: assume that selective survival exists, such that cigarette smokers who suffer an MI are more likely to die before reaching the hospital. Then a case-control study of MI and smoking, using hospitalized MI patients as cases will have alpha lower than beta (exposed cases are less available to study than are nonexposed cases). This bias will produce a distortion in the odds ratio that will understate a true association between smoking and MI (i.e., negative bias).

The assignment for this lecture has an exercise that asks you to apply this conceptual framework to a detection bias issue involving endometrial cancer and estrogen. The basic issue is that use of estrogen might lead to uterine bleeding, which would result in a woman seeking medical attention and receiving a dilation and curettage (D&C). If an occult (asymptomatic) endometrial cancer were

present, then the D&C would detect it. According to the detection bias scenario, women with occult endometrial cancer are therefore more likely to come to medical attention if they are estrogen users, creating a detection bias situation.

This scenario was vigorously disputed, since it depends upon the existence of a sizable reservoir of asymptomatic endometrial cancer, and is now widely discounted. Nevertheless, the endometrial cancer and estrogen issue provides abundant illustrations for concepts related to selection bias and information bias. We will take up this case study presently. (Note that though bias in case-control studies has attracted the most theoretical interest, all study designs are vulnerable.)

Recourse — Minimize loss to follow-up, obtain representative study populations, anticipate sources of bias and avoid them. Sometimes the factors associated with selection bias can be measured, in which case the analysis of the data can attempt to take these factors into account. Logic in the interpretation of the data may be able to marshall evidence for or against selection bias as having been responsible for an observed association. But if you can avoid it, that's the best!

Estrogen and endometrial cancer case example

During the 1970s, case-control studies reported a strong (OR about 10) association between endometrial cancer and use of postmenopausal estrogens. The association was biologically plausible, since the endometrium of the uterus is an estrogen-responsive tissue. Also, endometrial cancer rates were rising in geographical areas where use of postmenopausal estrogens was growing most rapidly.

Criticism of case-control studies had also been rising, however. For one, case-control studies reporting an association between breast cancer and the anti-hypertensive medication reserpine had received wide attention, but the association was later discounted. Also, critics of the case-control design (notably Alvan Feinstein, who labelled the design the "trohoc" study ["cohort" spelled backwards]) had become prominent. The *Journal of Chronic Disease* (now called the *Journal of Clinical Epidemiology*) hosted a conference of leading epidemiologists to discuss the validity of the design (proceedings published in Michel A. Ibrahim and Walter O. Spitzer. *The case-control study: consensus and controversy*. Pergamon, New York, 1979).

At about this time, Barbara Hulka, Carol J.R. Hogue, and the late Bernard G. Greenberg (then all at the UNC School of Public Health) published a comprehensive review of methodological issues involved in the estrogen-endometrial cancer association (Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen. *Amer J Epidemiol* 1978; 107:267-276). The case-control design is particularly susceptible to selection bias, because since the disease has already occurred, the validity of the study is critically dependent upon the selection of cases and controls. The Hulka et al. review made the following points (more material from this review is presented in the appendix to this chapter):

1. Ascertainment of cases

   Cases provide an estimate of estrogen exposure in women who develop endometrial cancer.

This estimate of the prevalence of exposure among cases can be expressed in probability terms as Pr(E|D) – the probability of exposure conditional on having the disease.

Cases in the study should therefore be representative of all similarly described (i.e., age, geography, subdiagnosis) persons who develop the disease with respect to exposure status. For endometrial cancer, two issues are -

a. Heterogeneity of cases (stage, grade, histological type) may reflect different underlying etiology or relationship to exposure.
b. Sources of cases and diagnostic process may have implications for exposure status (e.g., cases from rural hospitals may have had less access to postmenopausal estrogens).

2. Selection of controls

Controls provide an estimate of the prevalence of exposure in the source population from which the cases arose (now referred to as the "study base"). This prevalence can be expressed in probability terms as Pr(E) - the probability that a person selected at random from the study base is exposed to exogenous estrogens. Controls must therefore be representative of the study base with respect to exposure status, so that the prevalence of estrogen use in controls (in probability terms, Pr(E|not D) accurately estimates exposure in the study base. In addition, controls should be able to provide exposure and other data with accuracy equivalent to that obtainable from cases (this point concerns information bias and will be discussed later in this chapter).

Therefore, controls should be similar to cases in terms of:

a. Data sources, so that the opportunity to find out about prior estrogen use is equivalent to that for cases;
b. Other determinants of the disease that cannot be controlled explicitly

But controls should not be too similar to cases on nondeterminants of the disease.

Overmatching and the selection of controls

This last qualification was directed at the issue of detection bias raised by Feinstein (see above). Ralph Horwitz's and Feinstein's recommendation for reducing detection bias was to select controls from among women who had had the same diagnostic procedure as had the cases (dilation and curettage), thereby ensuring that controls did not have occult disease and making them more similar to the cases. Hulka et al.'s response was that such a selection procedure for controls constitutes overmatching.

The concept of overmatching and Horwitz and Feinstein's proposed "alternative controls" (*NEJM*, 1978) focus on the relationship of selection bias and the selection of the control group in a case-control study, which is why the estrogen – endometrial cancer topic is such an excellent one for understanding control selection.

Controls in an experiment

In a true experiment, in which one group is given a treatment and another serves as a control group, the optimum situation is generally for the treatment and control groups to be as close to identical as possible at the time of the treatment and to be subjected to as similar as possible an environment apart from the treatment. If randomization of a large number of participants is not feasible, the control group is matched to the experimental group to achieve as much similarity as possible in anything that might affect development of the outcome.

Earlier generations of epidemiologists were often taught that, by analogy, the control group in a case-control study should be similar to the case group in all characteristics other than disease (and exposure status, which the study seeks to estimate). In that way, exposure differences could more readily be attributed to the effects of the exposure on disease risk, the only other point of difference. Toward that objective, controls have often been matched to cases to increase similarity of the groups.

Analogies between experimental and case-control study designs

However, the analogy between the control group in a case-control study and the control group in an experiment, is faulty. In an experiment, exposure is introduced in one of two hopefully equivalent groups, and outcomes subsequently develop. The control group is chosen to have equivalent risk for the outcome in the absence of the exposure. In a case-control study, exposures exist in a population, and outcomes develop. The equivalence that is required for a valid comparison is that between exposed and unexposed persons. The case group – the members of the population who have developed the outcome – are not located in a corresponding position vis-a-vis the disease process as are the exposed group in a true experiment. The former is a group of people who develop the outcome; the latter are a group at risk for the outcome.

The correct experimental analog to the case group in a case-control study is the group of participants who develop the outcome during the experiment. In both designs, the cases arise from a population of both exposed (or "experimental") and unexposed (or "control") persons. Similarly, the correct analog for the control group in a case-control study is a random sample of all participants in the experiment at some point following the onset of exposure. The set of all participants in the experiment is the "study base" for the experiment. If a case-control study is conducted using the cases which arose in that experiment, then the control group should serve to estimate the proportion of exposure in that study base.

Matching and selection bias

Forcing the control group to be similar to the case group, either through matching or through using a source for recruitment of controls similar to that for recruitment of cases, will ordinarily make the control group less like the study base and may therefore introduce selection bias. Whether or not selection bias will be introduced depends upon the analysis methods used and whether or not the matching factors are related to prevalence of exposure. If the characteristics are unrelated to exposure then selection bias will not occur for that exposure, since both the matched and

unmatched control groups will presumably yield the same estimate of exposure prevalence. If the characteristics are risk factors for the disease, then although matching may introduce selection bias, this bias can be eliminated by controlling for the matching factors in the analysis (think of each matching factor as identifying subsets in both the cases and study base, so that the overall study can be regarded as a set of separate, parallel case-control studies, each itself valid).

Overmatching

However, if the characteristics are related to the exposure and are not risk factors for the disease, then forcing the controls to be more like the cases will distort both the exposure prevalence in controls (making it more like that in the cases and less like that in the study base) and odds ratio relating exposure and disease. This scenario is termed **overmatching**. If the matching factors are controlled in the analysis (which is not generally appropriate for factors other than risk factors for the outcome), then the estimated OR will be correct but less precise (i.e., have a wider confidence interval).

A hypothetical case-control study:

Suppose you are conducting an incident case-control study of endometrial cancer and exogenous estrogen. You arrange to be notified of any endometrial cancers diagnosed in 50-70-year-old female, permanent, full-time (or retired and on pension) state employees and retirees in a multi-state area. Assume that all receive medical care benefits; 100,000 are enrolled in fee-for-service plans, and 50,000 are enrolled in managed care (and no one changes!). This population is the study base.

During the five years of follow-up, 200 cases of endometrial cancer develop, for an overall cumulative incidence of endometrial cancer of 133 per 100,000 (0.00133). Of the 200 cases, 175 were exposed to estrogen, and 25 were not (these numbers were derived assuming a cumulative incidence of 200 per 100,000 (0.002) in women with estrogen exposure and 40 per 100,000 (0.0004) in women without exposure, but of course if you knew these incidences, you would not be conducting the study).

Suppose that a much larger percentage (75%) of women in fee-for-service plans are taking exogenous estrogen than are women in managed care (25%). However, you do not know that either, because the prescription records in the various organizations you are dealing with are not computerized (which is why you have resorted to a case-control study rather than following all 150,000 women as a cohort).

For your controls, you first choose a simple random (and by good fortune, precisely representative) sample of 600 women from the 150,000 master file of state employees and retirees. Your data then look as follows:

Southern endometrial cancer and estrogen study (SECES)

	Estrogen	No estrogen	Total
Endometrial cancer	175	25	200
Controls	350	250	600
Total	525	275	800

OR = 5.0
95% confidence interval: (3.19, 7.84)*

*(see chapter on Data Analysis and Interpretation)

(So far so good, since the CIR, by assumption, was 0.002/0.0004 = 5.0.)

However, you are concerned, since you anticipate that estrogen prescribing is very different in the two different types of health care plans. Your suspicion is further supported by the fact that 160/200=80% of the cases are in fee for service, compared to only two-thirds of the random sample controls (400/600) (and 100,000/150,000 in the study base). So even though you have no basis for believing that a woman's health care plan affects her risk for detecting endometrial cancer, you decide to make your control group more like the case group in regard to health plan membership (i.e., you overmatch).

Since 80% of the cases are in fee-for-service and 20% are in managed care, you use stratified random sampling to achieve that distribution in the controls. For 600 controls, that means 480 (80% of 600) from fee-for-service and 120 (20% of 600) from managed care. Since (unbeknownst to you), 75% of the women in fee-for-service take estrogen, as do 25% of the women in managed care, your control group will contain 390 women taking estrogen – 360 exposed women (75% × 480) from fee-for-service and 30 exposed women (25% × 120) in managed care. Thus, your data will now be:

Southern endometrial cancer and estrogen study (SECES)
MATCHED control group

	Estrogen	No Estrogen	Total
Endometrial cancer	175	25	200
Controls	390	210	600
Total	565	235	800

OR = 3.8
95% confidence interval: (2.40, 5.92)

The odds ratio for this table is 3.8, so your matched control group has indeed produced selection bias. Luckily your friend comes by and reminds you that when you use a matched control group, you need to control for the matching factor in your analysis. So you act as if you had conducted two separate studies, one among the women in fee-for-service and the other among the women in managed care (this is called a "stratified analysis" and will be discussed in the chapter on Multicausality – Analysis Approaches). Your two tables (and a combined total for cases and controls) are:

Southern endometrial cancer and estrogen study (SECES)
MATCHED control group, STRATIFIED analysis

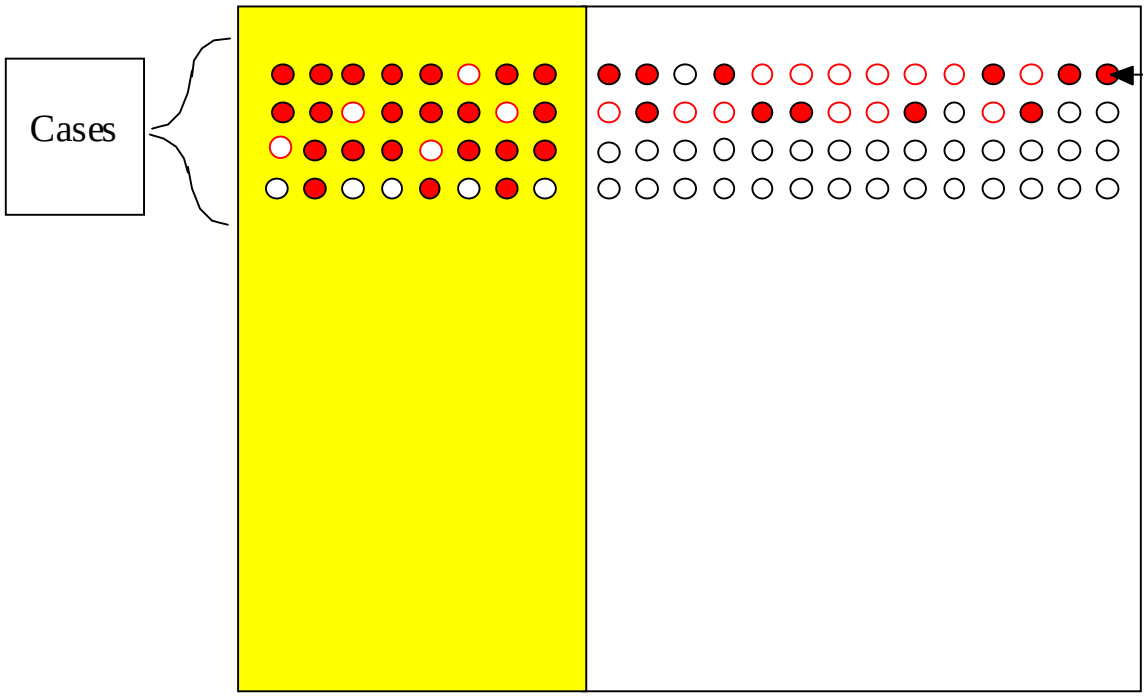
	Fee for service			Managed care			Both
	Estrogen	No Estrogen	Total	Estrogen	No estrogen	Total	Grand total
Cancer cases	150	10	160	25	15	40	200
Controls	360	120	480	30	90	120	600
Total	510	130	640	55	105	160	800
OR	5.00			5.00			
95% CI:	(2.55, 9.30)			(2.33, 10.71)			

Stratified analysis* (over both tables): OR=5.0 95% CI: (3.02, 8.73)

* See chapter on Multivariable analysis.

Each of the two case-control studies now has OR = 5.0. The control group within each type of health care plan was a simple random sample. The selection bias in the matched control group held only for the group as a whole compared to the study base as a whole. However, the interval estimate of the OR for the stratified analysis (the last table) is wider than the confidence interval for the OR in the unmatched analysis, indicating a less precise estimate.

` The "detection bias" hypothesis
for endometrial cancer and exogenous estrog



Selection bias in cohort studies

Selection bias is generally regarded as a greater danger in case-control than in cohort studies. The reason is that in cohort studies the investigator generally knows how many and which participants were lost to follow-up, so that s/he can assess the potential extent of bias. The investigator can also often examine baseline characteristics of participants who are lost to follow-up for indications that attrition is uniformly distributed and therefore less likely to result in selection bias.

Population cohort attrition

There is, however, a type of attrition that affects both cohort and case-control studies but which is unseen and difficult to categorize. The problem relates to the representativeness of people eligible for study. For simplicity we explain the situation in relation to cohort studies, but since a case-control study is simply an efficient method for studying the same phenomena as a cohort study of the study base, the problem is effectively the same in case-control studies.

A cohort consists of people who are alive at a point or period in calendar time or in relation to some event they undergo (e.g., graduating from college, joining a workforce, undergoing a surgical procedure) and then followed forward in time. The investigators' attention is for the most part directed at what happens after the cohort has been formed, but it is conceivable that mortality and migration occurring before that point have influenced who is available to enroll and thereby influenced what will be observed. If these early selection factors are related to an exposure under study they may diminish an observed effect.

Some examples:

If a cohort of HIV-infected persons is recruited by enrolling HIV seropositive persons identified through a serosurvey, those who have progressed to AIDS more quickly will be underrepresented as will persons involved in risk behaviors (e.g., injection drug use) that are associated with high mortality. Progression to AIDS in such a cohort will appear different than what would be observed if people were recruited at the time of initial HIV infection.

A study of the effect of hypertension in a cohort of elderly participants cannot enroll persons whose hypertension caused their death prior to the entrance age of the cohort. If those who died earlier had characteristics that made them more vulnerable to end-organ damage from hypertension, then the cohort study may observe less morbidity and mortality associated with hypertension than would be observed if the study had enrolled younger participants.

Even a cohort study in a population of newborns can enroll only infants from conceptions that result in a live birth. If environmental tobacco smoke (ETS) increases the rate of early fetal losses, possibly undetected, there may be differences between the fetuses who die and those who survive to birth. If fetuses who survive are more resistant to harm

from ETS, then a cohort study of harmful effects of ETS on infants may observe a weaker effect because the most susceptible of the exposed cases were never enrolled in the cohort.

Assuming that the target populations are defined as persons age 70 years and older (in the hypertension study) or newborns (in the ETS study), then internal validity as defined above would not appear to be affected. But study findings could nevertheless be misleading. If cholesterol lowering medication lengthens disease-free life in hypertensives, then more hypertensives taking this medication will survive to age 70 to enter the cohort. If these hypertensives have a higher rate of developing end-organ damage, then the observed rate of morbidity and mortality associated with hypertension will be higher, people taking cholesterol-lowering medication may now be observed to have higher morbidity and mortality, and the stronger effect of hypertension will be found to be associated with cholesterol-lowering medication. Similarly, a factor that reduces fetal loss in ETS-exposed pregnancies will increase the proportion of ETS-susceptible infants enrolled in the cohort study and will be associated with higher infant morbidity/mortality.

This problem would appear to be closer to lack of external validity (generalizability across time or setting), but it bears a strong resemblence to selective survival as encountered in a cross-sectional or case-control study (e.g., the example of a case-control study of aggressiveness and MI, used above). Thus losses prior to the inception of a cohort need careful consideration so that the investigator is not misled by selective factors operating at an earlier stage.

Selection bias due to missing data

One other potential cause of selection bias in studies of all kinds is missing data for a variable required in the analysis. Bias due to missing data is usually a topic considered under the heading of analysis, but its effect is akin to selection bias and its prevention requires avoidance of systematic differences in rates of missing data.

The problem can be particularly severe in analyses involving a large number of variables. For example, regression procedures often exclude an entire observation if it is missing a value for any one of the variables in the regression. This practice (called "listwise deletion") can exclude large percentages of observations and induce selection bias, even when only 5% or 10% of missing values for any one variable. Imputation procedures can often avoid the exclusion of observations, and depending upon the processes that led to the missing data (the missing data "mechanism") they can lead to less or unbiased analyses. There are also analytic procedures that can reduce the bias from nonresponse (inability to enroll participants) and/or attrition (loss of participants following enrollment).

Information bias

Information bias refers to systematic distortion of estimates resulting from inaccuracy in measurement or classification of study variables (misclassification bias is a subcategory of information bias when the variable has only a small number of possible values). For example, a disease may be present but go unrecognized, a blood pressure may be misread or misrecorded, recall of previous exposure may be faulty, or in extreme cases, data may have simply been fabricated by uncooperative subjects or research personnel. Typical sources of information/misclassification bias are:

1. variation among observers and among instruments – or variation across times by the same observer or instrument;
2. variation in the underlying characteristic (e.g, blood pressure) – and that variation has not been adequately accomodated by study methods;
3. misunderstanding of questions by a subject being interviewed or completing a questionnaire – or inability or unwillingness to give the correct response; or selective recall;
4. incomplete or inaccurate record data.

Systematic overview:

Information bias can occur with respect to the disease, the exposure, or other relevant variables. Sometimes, information bias can be measured, as when two methods of measurement are available, one being deemed more accurate than the other. Sometimes, information bias can be assumed to exist but cannot be directly assessed.

For example, if there is a true causal relationship between estrogens and endometrial cancer, i.e., a biological process by which estrogen molecules initiate or promote cancerous cell growth in the endometrium, then this pathophysiologic process presumably relates to certain specific molecular species, operating over a certain time period, and resulting in certain forms of endometrial cancer. To the extent that endometrial cancer is a heterogeneous entity, and the estrogen-related form is one subtype, then the observed association between endometrial cancer and estrogen is being diluted, as it were, by combining in one case group cancers caused by estrogens and cancers resulting from other mechanisms. Masking of the relationship also occurs by combining in one exposure group women whose exposure caused their cancers, women whose exposure to estrogen occurred only before or after the relevant time period in terms of the natural history of endometrial cancer, and women who were exposed to a nonpathogenic form of estrogen, nonpathogenic dose, or nonpathogenic mode of administration (should there be such).

Another example is the study of the health effects of exposure to lead. The traditional index of absorption, blood lead level, reflects only recent exposure, because the half-life of lead in blood is only about 36 days (see Landrigan, 1994). So there may be relatively little relationship between a single blood lead measurement and body lead burden. Pioneering studies by Herbert Needleman

employing lead exposure measures from deciduous teeth enabled the demonstration of a relationship between low lead exposure and cognitive and behavioral impairment in children. Now, the advent of K x-ray florescence analysis of lead in bone, where the half life is on the order of 25 years, may provide an important new tool in epidemiologic studies of lead exposure (Kosnett MJ et al., 1994).

For these reasons, rigorous study design and execution employ:

1. verification of case diagnosis, employing such procedures as multiple independent review of tissue samples, x-rays, and other diagnostic data;
2. definition of homogeneous subgroups, with separate analysis of data from each;
3. multiple data sources concerning exposure (and other relevant variables), permitting each to corroborate the other;
4. precise characterization of actual exposure, with respect to type, time period, dosage, etc.

Unfortunately, reality constraints impose compromises. For example, data from 20 years ago may be the most relevant in terms of the causal model, but data from two years ago may be much more available and accurate. In using the more recent data, one either assumes that recent exposure is a good proxy measure for previous exposure or that the recent exposure is also related to the disease, though perhaps not as strongly as the previous exposure.

[For more on the above, see Hulka, Hogue, and Greenberg, "Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen", and Kenneth J. Rothman, Induction and latent periods, *Am J Epidemiol* 114:253-259, 1981 (the Rothman article addresses the question of timing of exposure).]

One consideration raised by the above is the importance of developing specific hypotheses in advance of the study. Such hypotheses, if they can be elaborated, strengthen both the design and interpretation of the study. The design is strengthened because the hypotheses guide the investigator in selecting the relevant variables and their features (time of occurrence, etc.) on which to obtain data. It may not be possible to obtain just the right information, but at least the hypotheses protect guide the search. Hypotheses also provide guidance about what relationships to analyze and how to construct analysis variables (e.g., what disease subcategories to relate to which forms of exposure). Specific hypotheses – grounded in existing knowledge and theory – can also increase the persuasiveness of the findings.

Basic terms and concepts

Reliability (of measurement or classification) concerns the repeatability of a measurement – across time, across measurement instruments, across observers. If a measure is reliable, it may still not be accurate. But if a measure is not reliable, then the data values for it contain a substantial random component. This random component reduces the information content of the variable, the strength of associations involving it, and its effectiveness in controlling confounding (to be discussed in a

following chapter). The concept of reliability is relevant when two or more measures of comparable authoritativeness are being compared.

Validity (of measurement or classification) is the extent to which a measurement measures what it is supposed to measure. Assessment of validity, therefore, implies the availability of a measurement method that can be regarded as authoritative (often called a "gold standard"). Since one measure has more authority, our interest shifts from simple agreement between measures to evaluation of the less authoritative one. For example, if mean blood pressure measured with a random-zero mercury sphygmomanometer over a series of readings in a person lying on his/her back is our standard for "true" blood pressure, then a casual pressure in a person sitting down will be systematically inaccurate, since it will tend to be higher. Although we will examine agreement between the supine and casual blood pressures, our interest is on the accuracy of the latter with respect to the "gold standard".

Relationship of reliability and validity

"Validity" is used as a general term for accuracy or correctness. The procedure of assessing the accuracy of a measurement instrument is often referred to as validation. In many situations, though, we do not know the correct result, so the best we can do is to compare measurements that are assumed to be equally accurate. In these situations, agreement between measurements is termed "reliability".

In this sense, reliability is a subcategory of validity. However, reliability (repeatability, consistency) can be present without validity (two faculty can agree completely yet both be wrong!). Also, a measurement procedure can be valid in the sense that it gives the correct value on average, though each measurement includes a large amount of random variation (e.g., 24-hour dietary recall for cholesterol intake). Sometimes it is said that "a measure that is unreliable cannot be valid". Whether this statement is true or not depends upon what aspect of validity is being considered. More commonly, random error (unreliability) and bias (lack of validity) are regarded as independent components of total error.

Psychometrics is the subdiscipline of psychology that addresses the evaluation of questionnaire items and scales. "Validation" as used in psychometrics encompasses both reliability (consistency) and validity. However, due to the scarcity of classifications and measures that can be regarded as authoritative, much of psychometric validation concerns assessment of reliability. Common situations where reliability is important to examine are comparisons of performance of two raters or interviewers of equal stature (inter-rater reliability), of results from repeated measurements of a characteristic that is believed to be stable (test-retest reliability), and of scores from equivalent items that make up a scale (inter-item reliability – often termed "internal consistency").

Assessment of reliability

Validation involves the measurement of agreement between two or more measurements or classifications. Agreement is not identical to "association", but is rather a special case – the case in

which the both measures increase in the same direction and have the same scale. (An obvious example of an association which does not imply agreement is an inverse association.)

Percent agreement

For categorical variables, a simple measure of reliability is the percentage of instances in which the two measurement instruments agree. Supposed that 100 electrocardiograms (ECG) are given to two expert readers to code independently as "abnormal" or "normal", and that the two readers agree on 90 (30 that they both call "abnormal" and 60 that they both call "normal"). But the 90 percent agreement is not as good as it seems, since it gives "credit" for agreement that we would expect to occur just by chance. What if the two readers, preferring to play golf, left the ECG's with their secretaries, with instructions for each secretary to code each ECG independently by rolling a die and coding the ECG as "abnormal" if the die came up 6. To the unfortunate investigator, when s/he checked the reliability of the coding, the two readers might appear to have 72% agreement (3 abnormals and 69 normals).

Categorical variables - Kappa

One measure of reliability that adjusts for agreement expected to occur by chance is **Kappa** (introduced by Cohen in 1960). For a categorical variable without inherent ordering (e.g., initial diagnosis of chest pain as ?angina, ?gastroesopheal reflux, or ?musculoskeletal), Kappa is computed as:

$$K = \frac{p_o - p_c}{1 - p_c}$$

p_o = observed proportion of agreement

p_c = proportion of agreement expected by chance

The proportion of agreement expected by chance is computed by using the marginal percentages, the same procedure that is used for computing a chi-square test for associatiion.

Suppose that a managed care organization (MCO) is investigating the reliability of physician diagnostic work-up for chest pain. Two physician's initial assessments and order for diagnostic testings are compared for 100 sequential patients presenting with uncomplicated occasional chest pain at their initial visit to the MCO.

Comparison of diagnoses by physicians A and B among 100 patients reporting chest pain

		Physician B			
Physician A		?Angina	?Reflux	?Musculo-skeletal	Total
	?Angina	12	1	1	14
	?Reflux	2	36	4	42
	?Musculo-skeletal	2	8	34	44
	Total	16	45	39	100

Since the physicians agree on the initial diagnosis for 12 + 36 + 34 patients, their percent agreement is 82/100 = 82%. However, based on the marginals we expect considerable agreement by chance alone. The expected proportion of agreement by chance is computed from the marginal distributions as follows:

Expected proportion of agreement =

	(Proportion ?Angina Physician A) × (Proportion ?Angina Physician B)	14/100 × 16/100
+	(Proportion ?Reflux Physician A) × (Proportion ?Reflux Physician B)	42/100 × 45/100
+	(Proportion ?Mus-Sk Physician A) × (Proportion ?Mus-Sk Physician B)	44/100 × 39/100
=	14/100 × 16/100 + 42/100 × 45/100 + 44/100 × 39/100	
=	0.0224 + 0.189 + 0.1716 = 0.383	

The value of Kappa for this table is therefore: $$K = \frac{0.82 - 0.383}{1 - 0.383}$$

For assessing agreement between ordinal variables with few categories, weighted versions of Kappa are used in order to assign varying weights to different degrees of disagreement. A discussion of Kappa may be found in Joseph Fleiss's text *Statistical Methods for Rates and Proportions*. The second edition suggests adjectives for characterizing values of Kappa.

Continuous variables

For continuous measures and ordinal variables with many categories, the data display is a scatterplot, rather than a crosstabulation. Perfect agreement means that all of the measurement pairs lie on a straight line with slope 1 and intercept 0 (i.e., the line goes through the origin). The most direct index of the level of agreement between the two measures are the regression coefficient and intercept for the straight line that best fits the measurement pairs. The closer the regression

coefficient (slope) is to 1.0 and the regression intercept is to zero, and the narrower their confidence intervals, the better the level of agreement.

A common index of agreement is the correlation coefficient. The product-moment (Pearson) correlation coefficient (r) assesses the extent to which pairs of observations from the two measurements lie on a straight line. The Spearman (rank) correlation coefficient, *rho*, used for ordinal variables, assesses the extent to which the pairs of observations have the same ordering for the two measurement instruments (the lowest for the first instrument is close to the bottom for the second instrument, the tenth lowest for the the first is close to being tenth lowest for the second, and so on).

However, correlation coefficients ignore location and scaling. Thus, if the readings from one thermometer are always exactly two degrees below the readings from a second thermometer, agreement is certainly less than imperfect, yet the correlation coefficient between their readings will be 1.0 (for a perfectly straight line of slope 1.0, but not through the origin). If the readings from the first thermometer are always twice those of the second, the correlation will also be 1.0 (for a straight line through the origin, but with a slope of 2.0). Therefore a correlation coefficient alone is an inadequate assessment of agreement. It must be accompanied by a comparison of the location (mean, median) and scale (standard deviation or other measure of dispersion) for the readings of the two measures.

Reliability of a scale [optional]

A measure of reliability that is widely used in psychometrics is Cronbach's coefficient alpha. Coefficient alpha's conceptual basis (see Nunnally, *Psychometric Theory*) can be stated as follows.

Suppose that you have a set of questionnaire items each of which attempts to measure the same, unobservable construct (a "latent variable"). The response value for any individual item will reflect the value of that latent variable but also some amount of error, which is assumed to be random, independent of everything else, and symmetrically distributed with mean zero. Under these assumptions, the average of the response values for the set of items will provide a more reliable measure of the latent variable than is available from any single item (just as the average value for a set of otherwise equivalent blood pressure measurements will yield a more accurate (precise) value than any single measurement). The random components in the item responses should counterbalance each other, so that the average is a more precise measure of the latent variable.

In such a scenario, coefficient alpha assesses how much of the scale scores reflect the values of the latent variable and how much reflects measurement error. The the higher the "shared item variance" (the more the individual items in the scale agree with each other) and the larger the number of items, the higher the value of alpha. More precisely stated, coefficient alpha is the proportion of the total variance in the scale scores that represents the variance of the values of the latent variable (the rest being the variance of the random errors for each

item). Alpha values of 0.80 are are considered adequate for computing correlations and fitting regression models, and a sample size of 400 observations is regarded as adequate to estimate alpha (see Nunally).

Obstacles to realizing this ideal scenario include the probability that items are not perfectly equivalent, that people's responses to some items in the scale affect their answers to other items (so errors in item responses are not independent), and that factors other than the latent variable contribute non-random variation in item responses (thereby introducing systematic error, i.e., bias). Note that coefficient alpha does not address bias, only random variability.

Assessment of validity – sensitivity and specificity

As noted above, assessment of validity is directed toward the evaluation of a rater or measurement instrument compared to an authoritative rater or instrument. For detection of a characteristic or a condition, epidemiologists generally employ the concepts of sensitivity and specificity that were introduced in a previous chapter. Using the words "case" ("noncase") to refer, respectively, to people who have (do not have) the condition or characteristic (e.g., a disease, an exposure, a gene) being measured, then sensitivity and specificity are, respectively, the probabilities for correctly classifying cases and noncases.

Sensitivity is the ability to detect a case.

Specificity is the ability to detect a noncase.

Example:

If a procedure correctly identifies 81 of 90 persons with a disease, condition, or characteristic, then the sensitivity of the procedure is:

$$Se = 81/90 = 0.9 = 90\%$$

If the same procedure correctly identifies 70 of 80 persons without the disease, condition, or characteristic, then the specificity of the procedure is:

$$Sp = 70/80 = 0.875 = 88\%$$

In probability notation,

$$Se = Pr(D' \mid D) \qquad Sp = PR(\bar{D}' \mid \bar{D})$$

where D = case, $\bar{D}$ = noncase, D' = "classified as a 'case'", and $\bar{D}'$ = "classified as a 'noncase'".

The inverse of sensitivity and specificity are "false negatives" and "false positives". Some authors prefer to avoid the latter terms, because of the potential confusion about whether "negative" and "positive" refer to the test (in accordance with the definition in John Last's *Dictionary of Epidemiology* or to the true condition. However, the terms remain in common use, and we will follow the Dictionary's usage, whereby a "false negative" is a negative test result in a person who has the characteristic (i.e., an erroneous negative test) and "false positive" is an erroneous positive test result.

Sensitivity and specificity as defined above suffer from the same limitation that we have noted for percent agreement, that their calculation fails to take account of agreement expected on the basis of chance. Even a random process will classify some cases and noncases correctly. Methods for dealing with this limitation have been published (Roger Marshall, "Misclassification of exposure in case-control studies", Epidemiology 1994;5:309-314), but are not yet in wide use.

Impact of misclassification

The impact of misclassification on estimates of rates, proportions, and measures of effect depend on the circumstances. Consider the following example for a rare disease. Assume a cohort of 1,000 participants, of whom 60 develop CHD during the a four-year interval.

If the sensitivity of our diagnostic methods is only 0.80 (or 80%), then we will detect only 48 of those cases (48/60, i.e., 80% of 60). There will be 12 false negatives.

If the specificity of our diagnostic methods is 0.90 (or 90%), then we will incorrectly classify 94 of the 940 subjects who did not develop CHD (90% of the 940 noncases will be correctly identified as such, leaving 94 (940 minus 846) noncases to be incorrectly classified as "cases"). These 94 subjects will be false positives.

Thus, we will observe (or think we observe) 142 "cases" of CHD – 48 who in fact have CHD and 94 who actually do not. Note that in this case the majority of "cases" do not have the disease! This example illustrates the dilemma of false positives when studying a rare disease. The false positives and their characteristics will "dilute" or distort the characteristics of any "case" group we might assemble. Hence the emphasis on avoiding false positives through case verification, using such methods as pathological confirmation.

Suppose that the participants in this cohort are "exposed", and another similar cohort consists of 1,000 participants who are not "exposed". Assuming that the diagnostic accuracy is not influenced by exposure status, we expect the results for the two cohorts to be as follows:

Hypothetical scenario showing effect of misclassification bias on measures of association

	True (Se=1.0, Sp=1.0)		Observed: Se=0.8, Sp=1.0		Observed: Se=1.0, Sp=0.9		Observed: Se=0.8, Sp=0.9	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	60	30	48	24	154	127	142	121
$\bar{D}$	940	740	952	976	846	873	858	879
	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
RR	2.0		2.0		1.21		1.17	
RD	0.03		0.024		0.027		0.021	

From this example, we can see that:

1. Even rather high levels of sensitivity and specificity do not avoid bias;
2. Different epidemiologic measures are affected in different ways;
3. The RR need not be affected by imperfect sensitivity if specificity is sufficiently high; the RD will be affected, though.
4. The RR will be affected by imperfect specificity for detecting a rare disease even if sensitivity is high; however, the RD may be affected only slightly.
5. Specificity is of utmost importance for studying a rare disease, since it is easy to have more false positive tests than real cases, identified or not;
6. Bias in the classification of a dichotomous disease typically masks a true association, if the misclassification is the same for exposed and unexposed groups.

[Try creating a spreadsheet to see how various levels of sensitivity and specificity change the RR and RD. Convenient formulae are in the appendix.]

Types of misclassification

The above example deals with misclassification of a disease, misclassification that is independent of exposure status. Nondifferential misclassification of a dichotomous exposure variable, i.e., misclassification that occurs independently of disease status – will bias ratio measures of effect toward the null value of 1.0. This will also be the case for nondifferential misclassification of a dichotomous disease variable or of both a dichotomous disease variable and a dichotomous exposure simultaneously.

Differential misclassification, however, where errors in measuring one variable vary according to the value of another variable, can lead to bias in any direction. Common scenarios for differential misclassification are selective recall of past exposures or selective detection of disease based on knowledge of the patient's exposure history. Also, when the misclassified variable has more than two levels, even nondifferential misclassification can produce bias in any direction (because this last point has been emphasized only in recent years and because traditionally the teaching of epidemiology has focused on dichotomous disease and exposure variables, it is not uncommon to hear the maxim "nondifferential misclassification bias is towards the null" without mention of the exceptions).

In addition, measurement error for other variables involved in the analysis produces bias in a direction that depends on the relationships of the variables. For example, if we are performing age adjustment and have bias in the measurement of age, then the age adjustment will not completely remove the effect of age. A situation of this type is referred to as information bias in the measurement of a covariable and is discussed in Rothman and Greenland.

Direction and extent of bias

The importance of being able to discern the direction of the bias and, if possible, to assess its magnitude, is to enable interpretation of the observed data. For example, if a positive association is observed between two factors and the direction of misclassification bias can be shown to be toward the null, then such bias could not be responsible for the finding of a positive association. Similarly, if misclassification bias can be shown to be in the positive direction, then the failure to find an association cannot be due to that bias. In addition, techniques exist to correct for errors in measurement in a number of analytic procedures. However, these procedures often require some outside estimate of sensitivity and specificity.

Where the categories of bias break down

Earlier it was mentioned that the boundaries between random error and systematic error as well as those among the three classes of bias sometimes become unclear. Here are some situations that are challenging to classify.

False negatives in detecting cases for a case-control study.

If cases are missed due to information bias, then such persons will not be counted as cases in a case-control study. If this lack of sensitivity is in some way related to exposure status (e.g., greater detection of endometrial cancer among women who take estrogen – the detection bias issue), then the case group will not be representative of the population of cases.

From the viewpoint of the case-control study this type of bias will be classified as selection bias, since it is manifest through differential selection probabilities for cases. But the error mechanism in this scenario was misclassification of cases. Moreover, if some women with asymptomatic endometrial cancer happen to be selected as controls, their presence in the control group is

presumably classified as information (misclassification) bias, since in this situation the subjects belong in the study, except that they should be in the case group.

Variability in a parameter being measured can produce both random error (measurement imprecision) and information bias in a measure of effect

Blood pressure, for example, varies from moment to moment, so that every measurement of blood pressure reflects a degree of random variability (random error). If blood pressures are measured on a single occasion, and then disease incidence or some other endpoint is recorded during the ensuring five years, the observed association between blood pressure and the outcome will understate any true association.

The reason for this is that subjects who were classified as "high" on their initial measurement will include some who were "high" just by chance. Subjects classified as "low" will include some who were "low" just by chance. The resulting error in exposure measurement will muddy the contrast between the group outcomes compared to what would obtain if our "high" group contained only those who were truly "high" and low group contained only those who were truly "low".

Assuming that chance variability is independent of study outcomes, then the result is nondifferential misclassification, and the observed association will be weaker than the "true" association. Thus, random error can produce systematic error, or bias.

Blood pressure (and other physiological parameters) also varies in a diurnal pattern, being lower in the morning and rising during the day. Failure to provide for diurnal variation can produce several kinds of error. For example, if blood pressures are measured on study subjects at random times during the day (i.e., each subject's blood pressure is measured once, but any given subject may be examined at any time of day), then the diurnal variation adds a component of random error to that from the moment to moment variation. Therefore, estimates of group means and their differences will be more imprecise than if measurements had been conducted at the same time of day.

If for some reason subjects in one category (e.g., blue collar workers) are examined in the morning and subjects in another category (e.g., homemakers) are examined in the afternoon, then there will be a systematic difference between the mean blood pressures for subjects in the different categories, a systematic difference arising from the systematic variation in blood pressure combined with the systematic difference in time of measurement. The resulting systematic error could lead to selection bias or information bias depending upon the nature of the study.

Regression to the mean

A well-known phenomenon that illustrates how random variability can lead to systematic error is regression to the mean (Davis CE: The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 104:493-498, 1976). When a continuous variable, such as blood pressure or serum cholesterol, has a degree of random variability associated with it (or with its measurement), then each measurement can be thought of as based on the "true value" for the

subject plus or minus a random noise factor. If the distribution of the random variable is symmetric with a mean of zero, then the average value of a series of different readings will be close to the "true value". If the random component is large, however, any given measurement can be substantially above or below the average.

In such a situation, a variable for which a given measurement falls at the high end or the low end of the distribution for that variable will tend to be closer to the center of the distribution for a later measurement. For example, in the Lipid Research Clinics (LRC) Prevalence Study, populations were screened for cholesterol and triglyceride levels, and those with elevated levels were asked to return for additional evaluation. If, say, 15% of the subjects screened were asked to return, it can be expected (and did happen) that many of those subjects did not have elevated levels upon re-measurement.

The reason for this "regression" is that the group of subjects in the top 15% of the lipid distribution at their screening visit consists of subjects whose lipid measurement was high due to a large positive random component as well as subjects whose lipid levels are truly high. On re-measurement, the random component will, on average, be smaller or negative, so that subjects without truly high lipid levels will fall below the cutpoint as well as some subjects with truly high levels but who on this measurement have a large negative random component.

If by an extreme value we mean one that is "unusually high", that implies that usually it should be lower. The opposite is true for unusually low values. Therefore, the average serum cholesterol in an unselected population will not tend to "regress towards the mean", since in a random process the increases and decreases will balance each other. But if we select a portion of the population based on their initial measurements' being high (and/or low), then that selected population will tend to "regress" towards the population mean.

In regression toward the mean, we have a situation in which random variability can produce systematic distortion, in the sense that the mean of the cholesterol levels (or blood pressures) of the "elevated" subjects overstates their "true mean" (assuming that "true" is defined as an average of several measurements). Whether this distortion produces selection bias or information bias will depend upon the actual process of the study.

Suppose that "high risk" subjects (elevated cholesterol, blood pressure, and other CVD risk factors) are enrolled in a "wellness" program and their risk levels are measured several months later, there will probably be some decline in these levels regardless of the program's effects, simply due to regression to the mean. This process is one reason for the importance of a randomly allocated control group, which would be expected to experience the same regression.

[According to John R. Nesselroade, Stephen M. Stigler, and Paul B. Baltes, regression to the mean is not a ubiquitous phenomenon, but depends upon the characteristics of the underlying model or process involved. A thorough, but largely statistical, treatment of the topic can be found in "Regression toward the mean and the study of change," Psychological Bulletin 88(3):622-637, 1980.]

Appendix 1

Formulas to see the effects of various levels of sensitivity and specificity on the RR and RD

If a, b, c, d are the TRUE values of the cells in a four-fold table, then the observed RR and observed RD in the presence of Sensitivity (Se) and Specificity (Sp) for measuring disease are given by:

$$\text{Observed RR} = \frac{[(Se)a + (1\text{-}Sp)c]/n_1}{[(Se)b + (1\text{-}Sp)d]/n_0}$$

$$\text{Observed RD} = \frac{(Se)a + (1\text{-}Sp)c}{n_1} - \frac{(Se)b + (1\text{-}Sp)d}{n_0}$$

$$= Se\left(\frac{a}{n_1} - \frac{b}{n_0}\right) + (1 - Sp)\left(\frac{c}{n_1} - \frac{d}{n_0}\right)$$

Appendix 2

More on the concern to avoid false positive diagnoses of disease in case-control studies of a rare disease (e.g., endometrial cancer and estrogen) – the importance of verification of case status: [This is a simplified version of the presentation in the Hulka, Hogue, and Greenberg article in the bibliography.]

The case-control strategy aims to estimate the probability of exposure in cases and in noncases, the ideal for the latter being the general population from which the cases arose. Misclassification of disease leads to contamination of these probability estimates. In particular, false positives "dilute" the cases:

{Diagram)

The observed probability of exposure in subjects classified as "cases" equals:

1. the probability of exposure in true cases
2. plus a distortion equal to the proportion of false positive "cases" multiplied by the difference in exposure probability between true cases and false positives.

Algebraically,

$Pr(E|D') =$ — the **observed** exposure prevalence in "cases"

$Pr(E|D)$ — the **true** exposure prevalence in cases

$+ Pr(\overline{D}|D')\,[Pr(E|\overline{D}) - Pr(E|D)]$ — the **bias**

where E = exposure, D = a **true** case, $\overline{D}$ = a **true** noncase,
and D' = **any** subject who is classified as a "case" (correctly or incorrectly)

So $Pr(\overline{D}|D')$ is the probability that someone who is **called** a "case" is in fact a **noncase**

and $Pr(E|\overline{D})$ is the probability that a **true noncase** has the exposure.

Correspondingly, the **observed** probability of exposure in subjects classified as "**noncases**" equals:

1. the probability of exposure in true **noncases**
2. plus a distortion equal to the proportion of false negatives among persons classified as "noncases" multiplied by the difference in exposure probability between true noncases and false negatives.

Algebraically,

$\Pr(E|\bar{D}') =$ — the **observed** exposure prevalence in "noncases"

$\Pr(\bar{E}|D)$ — the **true** exposure prevalence in noncases

$+ \Pr(\bar{D}|D')\,[\Pr(E|D) - \Pr(\bar{E}|D)]$ — the **bias**

where $\bar{D}'$ = any subject classified as a "noncase" (correctly or incorrectly)

Numerical example:

If:

the probability of exposure in true cases = 0.4,

the probability of exposure in true noncases = 0.2,

the probability of exposure in false positives = 0.2 (i.e., the false positives are really just like other noncases)

then in a sample of subjects classified as "cases" in which one-third are falsely so classified (i.e., false positives) we expect to observe a probability of exposure of:

$$\Pr(E|D') = 0.4 + (1/3)\,[0.2\text{-}0.4] = 0.4 - (1/3)[0.2] = 0.333$$

or equivalently,

$$\Pr(E|D') = (2/3)(0.4) + (1/3)(0.2) = 0.333$$

(i.e., the prevalence of exposure is a weighted average of the exposure prevalence in the correctly classified cases and the exposure prevalence in the false positives).

Since the true probability of exposure in cases is 0.4, the observed results are biased downward. Since the proportion of false negatives in the control group (diseased subjects classified as "noncases") will generally be small if the disease is rare, the estimate of the probability of exposure in noncases will generally not be biased.

The true OR is 2.67 [{.4/(1-.4)}/{.2/(1-.2)}]; the observed OR is 2.0 [{.333/(1-.333)}/{.2/(1-.2)}]. The discrepancy would be greater if the true exposure probabilities were more different.

Bibliography

Henneckens and Buring. *Epidemiology in Medicine*. Rothman and Greenland, 1998. Rothman. *Modern Epidemiology*. Chapters 7, 8. Kleinbaum, Kupper and Morgenstern. *Epidemiologic Research: Principles and Quantitative Methods*. Chapter 10, Introduction to Validity.

Armstrong BK, White E, Saracci Rodolfo. Principles of exposure measurement in epidemiology. NY, Oxford, 1992, 351 pp., $59.95. Key reference (reviewed in Am J Epidemiol, April 15, 1994).

Brenner, Hermann and David A. Savitz. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. Am J Epidemiol 1990; 132:181-192.

Cohen, Bruce B.; Robet Pokras, M. Sue Meads, William Mark Krushat. How will diagnosis-related groups affect epidemiologic research? Am J Epidemiol 1987; 126:1-9.

Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 1990; 132:746-8; and correspondence in 1991;134:441-2 and 135(12):1429-1431

Feinleib, Manning. Biases and weak associations. Preventive Medicine 1987; 16:150-164 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiologic research. New Engl J Med 1982; 307:1611-1617.

Feinstein, Alvan R.; Stephen D. Walter, Ralph I. Horwitz. An analysis of Berkson's Bias in case-control studies. J Chron Dis 1986; 39:495-504.

Flanders, W. Dana; Harland Austin. Possibility of selection bias in matched case-control studies using friend controls. *Am J Epidemiol* 1986; 124:150-153.

Flanders, W. Dana; Coleen A. Boyle, John R. Boring. Bias associated with differential hospitalization rates in incident case-control studies. *J Clin Epidemiol* 1989; 42:395-402 (deals with Berkson's bias for incident case control studies - not a major work)

Flegal, Katherine M., Cavell Brownie, Jere D. Haas. The effects of exposure misclassification on estimates of relative risk. Am J Epidemiol 1986; 123:736-51.

Flegal, Katherine M.; Penelope M. Keyl, and F. Javier Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991; 134(10):1233-44.

Gregorio, David L.; James R. Marshall, Maria Zielezny. Fluctuations in odds ratios due to variance differences in case-control studies. *Am J Epidemiol* 1985; 121:767-74.

Horwitz, Ralph I. Comparison of epidemiologic data from multiple sources. J Chron Dis 1986; 39:889-896.

Horwitz, Ralph I. and Alvan R. Feinstein. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med 1978;299:1089-1094.

Hulka, Barbara S., Carol J.R. Hogue, and Bernard G. Greenberg. Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen. Amer J Epidemiol 1978; 107:267-276.

Hulka, BS, Grimson RC, Greenberg BG, et al. "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen. Am J Epidemiol 1980;112:376-387.

Hutchison, George B. and Kenneth J. Rothman. Correcting a bias? Editorial. N Engl J Med 1978;299:1129-1130.

Kosnett JM, Becker CE, Osterich JD, Kelly TJ, Pasta DJ. Factors influencing bone lead concentration in a suburban community assessed by noninvasive K X-ray fluorescence. JAMA 1994;271:197-203

Landrigan, Philip J. Direct measurement of lead in bone: a promising biomarker. Editorial. JAMA 1994;271:239-240.

Maclure, Malcolm; Walter C. Willett. Misinterpretation and misuse of the Kappa statistic. *Am J Epidemiol* 1987; 126:161-169.

Marshall, James R. and Saxon Graham. Use of dual responses to increase validity of case-control studies. *J Chron Dis* 1984;37:107-114. (also commentary by Stephen D. Walter, authors' reply, and Walter's reply in that issue).

Neugebauer, Richard and Stephen Ng. Differential recall as a source of bias in epidemiologic research. *J Clin Epidemiol* 1990; 43(12):1337-41.

Nunnally, Jum C. and Ira H. Bernstein. *Psychometric theory*. New York: McGraw-Hill, 1994.

Roberts, Robin S., Walter O. Spitzer, Terry Delmore, David L. Sackett. An empirical demonstration of Berkson's bias. *J Chron Dis* 1978; 31:119-128.

Sackett, D.L.: Bias in analytic research. *J Chron Dis* 32:51-63, 1979. (and comment). In Ibrahim: *The Case-Control Study*.

Schatzkin, Arthur; Eric Slud. Competing risks bias arising from an omitted risk factor. *Am J Epidemiol* 1989; 129:850-6.

Sosenko, Jay M.; Laurence B. Gardner. Attribute frequency and misclassification bias. *J Chron Dis* 1987; 40:203-207.

Walker, Alexander M. Comparing imperfect measures of exposure. *Am J Epidemiol* 1985; 121:783-79

Warnecke RB, Johnson TP, Chavez N, Sudman S, O'Rourke DP, Lacey L, Horm J. Improving question wording in surveys of culturally diverse populations *Ann Epidemiol* 1997; 7:334-342.

Warnecke RB, Sudman S, Johnson TP, O'Rourke D, Davis AM, Jobe JB. Cognitive aspects of recalling and reporting health-related events: Pap smears, clinical breast exams, and mammograms. *Am J Epidemiol* 1997; 13(3):305-315.

White, Emily. The effect of misclassification of disease status in follow-up studies: implications for selecting disease classification criteria. Am J Epidemiol 1986; 124:816-825.

Sources of error - Assignment

1. The continuing controversy over the health effects of low frequency magnetic fields began with an epidemiologic study of electrical wiring configurations and childhood cancer (Wertheimer N. *Am J Epidemiol* 1979; 109:273-284).

   Cases consisted of persons dying of cancer in Colorado before age 19 in the years 1950-1973 who also had a Colorado birth certificate and whose birth or "death" address was in the greater Denver area and had been occupied from 1946-1973. Controls consisted of next (non-sibling) Denver-area birth certificates. Residential status for cases and controls is shown below.

   For all study participants, birth and death addresses were visited and a map of the wires and transformers of the electric power distribution was drawn. Homes were categorized as having "high-current configurations" (HCC), "low-current configurations" (LCC), or "very low current configurations" (VLCC), according to their proximity to high current distribution lines. Table 2 shows the distribution of expected current of residence addresses for cases and controls:

Table 1
Residential status of cases and controls

Residential status	Cases	Controls
Stable, same birth & death address	109	128
Moved, birth & death addresses available	145	128
Only one address available, either birth or death	88	88

Table 2
Case-control distribution for the amount of current expected from different wiring configurations, based on all known addresses for study participants

Wiring configuration	Expected current	Cases	Controls	% Cases
HCC	High-very high	182	103	64
LCC	Low	289	324	47
VLCC	Very low	20	45	31

   a. Calculate the odds ratios comparing HCC to LCC houses, and LCC to VLCC houses.

   b. What is the meaning of the percentages in the right-most column? Are they incidences? Do they indicate a dose-response relationship between electric current and cancer occurrence?

   c. Identify likely sources of selection and information bias in this study, with particular attention to those sources that are common in case-control studies. Can you suggest methods to minimize these sources of bias? (Use your imagination—you are not expected to consult the article, though if you wish to do so, try answering this question first).

   d. What was the purpose of selecting controls by taking the next birth certificate?

e. OPTIONAL: If the reported cumulative incidence (CI) of childhood cancer (by age 19) is 10 cases per 10,000 children in the general population, what would you estimate the cumulative incidence to be among children living next to HCC's assuming that one out of five children live next to HCC's? What proportion of childhood cancer might be attributable to HCC's?

2. This question is based on the attached extract from Rosenberg et al., "Oral contraceptive use in relation to nonfatal myocardial infarction". (*Am J Epidemiol* 1980; 111:59-66).

   a. What type of study design has been used?

      A. case-control with prevalent cases

      B. case-control with incident cases

      C. prospective cohort

      D. retrospective (historical) cohort

      E. ecologic

      F. case-control nested in a cohort

   b. Give 2 possible sources of selection bias that might interfere with using these data to obtain estimates of the relative risk of MI in women taking oral contraceptives. For each source, give an example of how it might cause the true relative risk to be overstated in the data.

   c. Briefly assess the likelihood of misclassification of the outcome measure, reported MI.

   d. Briefly (2-3 sentences) assess the likelihood of misclassification of the exposure measure, reported oral contraceptive use.

3. The issue of "detection bias" sparked a vigorous controversy in the investigation of the relationship between endometrial cancer and the use of exogenous estrogen preparations. The case for the importance of "detection bias" was presented by Horwitz and Feinstein (Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *N Engl J Med* 1978; 299:1089-1094). The following questions refer to that article and Horwitz and Feinstein's detection bias argument.

   a. Give a definition of the term "detection bias" as it is applied by Horwitz and Feinstein to studies of the endometrial cancer-exogenous estrogen relationship.

   b. Would "detection bias" tend to overstate or understate a truly positive association between estrogen use and endometrial cancer? Explain briefly.

   c. Is "detection bias" (in the above sense) a form of selection bias or information (misclassification) bias? Justify your position.

   Consider "detection bias" to be a form of selection bias, and let alpha, beta, gamma, and delta be probabilities by which individuals in the target population are included in a study, according to the following scheme:

Alpha (α) is the probability by which individuals with endometrial cancer and a history of estrogen use are included in the study;

Beta (β) is the probability by which individuals with endometrial cancer but without a history of estrogen use are included in the study;

Gamma (γ) is the probability by which individuals without endometrial cancer but with a history of estrogen use are included in the study;

Delta (δ) is the probability by which individuals without endometrial cancer and without a history of estrogen use are included in the study.

Selection probabilities of inclusion in the actual population, by disease and exposure characteristics of individuals in the target population

	E	$\overline{E}$
D	α	β
$\overline{D}$	γ	δ

d. Assuming that no other source of selection bias is present, what relationship among or between selection probabilities most closely characterizes the detection bias situation described by Horwitz and Feinstein? Justify your answer.

e. Characterize, in terms of the above selection probabilities, the impact of the approach adopted by Horwitz and Feinstein to correct for the effect of "detection bias" (i.e., their use of "alternative controls"). Justify your answer.

f. What is an alternate, and presumably theoretically preferable, approach to avoiding detection bias and how would it be characterized in terms of the above selection probabilities? Comment on the practicality of the preferable approach.

4. The major community studies of CVD, such as the Framingham study, began before the availability of exercise ECGs, echocardiography, and other sophisticated methods of detecting CHD. For example, in the Evans County Study (Cassel JC *et al.*, *Arch Intern Med* 1971 (December); 128 [entire issue]), CHD case detection in living subjects was accomplished using clinical judgment based on history, symptoms, physical examination, resting ECG, and chest X-ray. Opportunities for bias from misclassification include both an incorrect exclusion decision at enrollment (so that a subject later diagnosed as having CHD may have been a prevalent case at the outset and therefore not a new case) and misclassification of CHD status at follow-up.

   The following questions are based on data from the Evans County Cardiovascular Disease Study (Cornoni JC, Waller LE, Cassel JC, Tyroler HA, Hames CG. The incidence study—study design and methods. *Arch Intern Med* 1971 [December];128:896-900):

Results from the Evans County Cardiovascular Study

Persons examined in 1960-62	3102
CHD cases at initial examination	93
Died before 1967-1969 reexamination (includes 36 deaths among those with CHD in 1960-62 and 56 CHD deaths among those without CHD in 1960-62)	320
Moved, vital status could not be determined	40
Known alive, but not re-examined (migrated or refused)	212
Re-examined in 1967-69 (includes 57 subjects who had CHD in 1960-62)	2530
CHD cases detected at reexamination in 1967-69 among survivors initially free of CHD	87

a. Diagram the above data.

b. Calculate the observed (87-month) cumulative incidence of CHD in the Evans County incidence study (exclude subjects who migrated, refused re-examination, were lost to follow-up, or died of non-CHD causes).

c. Estimate the "true" incidence of CHD between 1960-62 and 1967-69 that would be expected if the sensitivity and specificity of the diagnostic procedure in the Evans County study were, respectively, 70% and 98%. Assume that the 93 CHD cases detected in 1960-62 had in fact constituted all those and only those with CHD at the time and that there was no misclassification of cause of death for subjects who did not survive until reexamination.

d. *Optional*: It is reasonable to suppose that the sensitivity and specificity in 1960-62 would have been worse than in 1967-69. What is the lowest that specificity could have been given the data (you may set sensitivity at any level you like).

Sources of error - Assignment solutions

1.

a. $$OR_{HCC/LCC} = \frac{ad}{bc} = \frac{(182)(324)}{(289)(103)} = 1.98$$

$$OR_{LCC/VLCC} = \frac{(289)(45)}{(20)(324)} = 2.01$$

b. The percentages indicate the proportion of individuals within each exposure category (i.e., wiring configuration) who were cases. These percentages do not represent incidences (though Rothman and Greenland call them "pseudo-incidence rates". To calculate incidence one must know the size of the population at risk. The controls here are at best a (very small) sample of the population at risk. The percentages do suggest a dose response relationship, for although the total number of controls was arbitrary and fixed, their distribution among the various exposure categories was not.

c. Potential Sources of Biases.

i) Misclassification of exposure. The measurement of exposure was extremely imprecise. Magnetic fields in the living space were not measured directly. The child's address at death did not necessarily represent where he had lived during the majority of his life (and thus the exposure he had received). Also, wiring configurations could have changed between the time of actual exposure and the time of measurement. Similarly, no note was taken of exposures received when not at home (for instance at school). Assuming that in-home exposure measurement was not feasible, misclassification of exposure could have been reduced if the study had been limited to participants who had the same birth and death address and to those participants who lived in multiple dwellings, all of which were evaluated and had the same "current expected" classification.

ii) Specification of the outcome variable was imprecise. "All cancers" is a very heterogenous group (with more than 1 etiology) rendering the results suspect. The study apparently relied on cause of death from death certificates, without obtaining supporting evidence from medical records and pathological reports.

iii) Only children who had died as a result of their cancer were included in the study. It may be that HCCs are not carcinogenic, but rather are somehow related to prognosis once the cancer has developed.

d. The purpose of selecting controls by taking the next birth certificate was to match on age and to select controls who would have had similar environmental exposures other than magnetic fields. Also, the use of a systematic procedure avoids unwanted variability that creates opportunities for introducing bias.

e. Let: I = overall incidence

I_1 = incidence in the exposed

P_1 = proportion of population exposed

I_0 = incidence in the nonexposed

R_0 = proportion of the population nonexposed

RR = relative risk

We know that:

$I = I_1P_1 + I_0P_0$

$I_e = (RR)(I_0)$, so that

$I = (RR)(I_0)(P_1) + I_0P_0$, and

$$I_0 = \frac{I}{(RR)(P_1) + P_0}$$

From the problem,

$I = .001$

$P_1 = .2$; $P_0 = .8$

RR = 2.11 (combining LCC & VLCC into 1 group)

$$I_0 = \frac{0.001}{2.11(.2)+.8} = 0.0008$$

$$\text{Population Attributable Risk \%} = \frac{I - I_0}{I} = \frac{0.001-0.0008}{0.001} = 20\%$$

(For a review of the 20 years of epidemiologic studies stimulated by this one, see Greenland S, Sheppard AR, Kaune WT, Poole C, Kelsh MA. A pooled analysis of magnetic fields, wire codes, and childhood leukemia. *Epidemiology* 2000;11:624-634.)

2.

a. Study design - a case control study with prevalent cases.

b. Selection bias:

i) Non-response--about 29% of the sample did not return completed questionnaires. If nonrespondents were disproportionately distributed so that the nonresponse rates among MI nonusers of OCs or non-MI users of OC were higher than the other non-response rates, then the odds ratio would overstate the true association.

ii) Selective survival--if OC users and nonusers had different fatality rates from MI, then the prevalent (surviving) MI subjects would not provide an accurate estimate of OC use among women who later develop MI. If the fatal MI rate is higher among nonusers of OC, then the odds ratio observed would overstate the true association.

c. Misclassification of the outcome measure is likely, since:

i) MI may go undetected ("silent MI")

ii) MI may not be diagnosed, despite symptoms (due to, for example, lack of sensitivity of diagnostic tests used)

iii) MI history may be "denied" or mistaken, though for nurses one expects greater accuracy of reporting. No check of hospital records appears to have been made. Probably the true prevalence will be understated.

Misclassification of the exposure measure is also likely, particularly concerning time periods and duration of OC use. Memory in this case is unverified by pharmacy or physician records. Furthermore, the differences among types of OC preparations have not been noted.

The above limitations do not by any means invalidate the study, nor should the investigators necessarily have attempted to collect additional data. But it is important to be aware of the limitations in interpreting the data and in reconciling results with other investigations.

3.

a. Detection bias, in the sense in which Horwitz and Feinstein have applied the term to studies of endometrial cancer and exogenous estrogen, refers to a distortion in the observed proportion of estrogen users among women diagnosed as having endometrial cancer. The distortion in the case would result from the allegedly greater likelihood of diagnostic testing for endometrial cancer in women who take estrogens. The series of events envisioned is: women who take estrogen tend to have vaginal bleeding, prompting them to see their doctor, who then performs a diagnostic procedure (dilitation and curretage ["D&C"]). If an asymptomatic cancer is present, it will come to medical attention. A similar cancer present in a woman not receiving estrogens would go undetected. So the additional diagnostic attention given to the women taking estrogens leads, according to Horwitz and Feinstein, to additional detection of asymptomatic cancers. The end result is that in any series of endometrial cancer cases, the proportion of estrogen users is artificially inflated.

b. "Detection bias", as described above, would tend to overstate a truly positive association between estrogen use and endometrial cancer, because by artificially inflating the proportion of estrogen users among endometrial cancer cases, the difference between cases and controls would become more marked.

c. "Detection bias". in the above sense, is a form of selection bias, since it deals with the selection or ascertainment of cases into the study population. It is true that there is a form of misclassification at work, in that women with asymptomatic endometrial cancer are going unrecognized as cases – and one or two of them might conceivably appear among the control group of a study population. Processes that influence who becomes part of the study population lie in the realm of selection bias. The misclassification of a possible control or two in a study population could cause information bias, but only to a trivial degree. So it

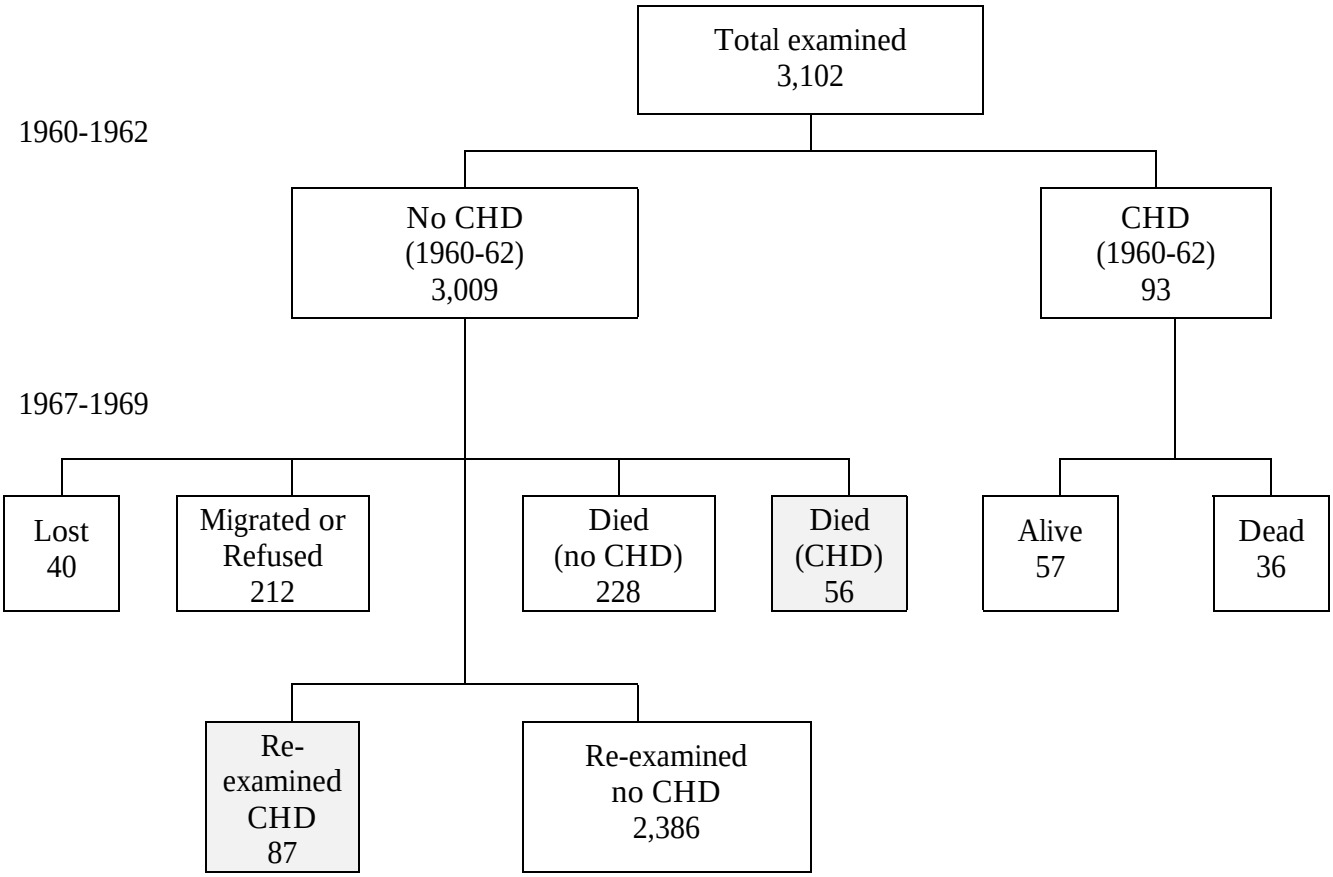
makes most sense to view detection bias as a form of selection bias resulting from over representation of estrogen users among the cases.

d. Detection bias, as described by Horwitz and Feinstein for this situation, is characterized by alpha greater than beta: the probability of coming to medical attention, therefore of being available for the case group of a study, is greater for women using estrogen than for women not using estrogen.

e. The approach adopted by Horwitz and Feinstein attempts to introduce a compensatory distortion in the control group, by recruiting controls from a population that is known to have higher estrogen usage. They therefore seek to increase gamma relative to delta, to increase the proportion of estrogen users among controls. Unfortunately, there is no way to know how great is the distortion of alpha relative to beta, nor to know how much distortion is being introduced to "compensate." Two biases don't necessarily make a right!

f. A presumably preferable alternative, theoretically, would be to increase beta so that it equals alpha, i.e., to introduce some measure to detect asymptomatic cancers in nonusers of estrogen (or in all women, without regard to estrogen use). With present technology, this would require subjecting asymptomatic women to D&C's, an impractical and ethically dubious approach given the low prevalence of endometrial cancer, the nature of a D&C, and the curability of symptomatic endometrial cancer.

a. Flow diagram:



b. Observed cumulative incidence (removing from the denominator subjects lost to follow-up or dying free from CHD):

$$CI = \frac{87 + 56}{87 + 56 + 2{,}386} = \frac{143}{2{,}529} = 0.0565 = 56.5 \text{ per } 1{,}000$$

c. Assume that (1) achieved sensitivity and specificity were, respectively, 70% and 98% for CHD detection among the 2,473 (2,386 + 87) persons free of CHD in 1960-62 who were re-examined in 1967 and (2) there was 100% for both sensitivity and specificity for CHD detection at death).

Since we are assuming 100% sensitivity and specificity for the 56 CHD deaths we will remove them from the following computations.

Let T = "true" nonfatal incident cases

Persons counted as nonfatal incident cases are "true" cases correctly classified PLUS "true" *non*cases *in*correctly classified (see table on next page):

Correctly classified true cases + Incorrectly classified true noncases = Total observed "cases"

$$\text{Sensitivity} \times \text{cases} + (1 - \text{Specificity}) \times \text{noncases} = 87 \text{ observed cases}$$

$$(Se) \times T + (1 - Sp) \times (2473 - T) = 87 \text{ observed cases}$$

$$0.7 \times T + (1 - 0.98) \times (2473 - T) = 87 \text{ observed cases}$$

$$0.7 \times T + 0.02 \times (2473 - T) = 87 \text{ observed cases}$$

$$0.68 \times T + 49.46 = 87 \text{ observed cases}$$

$$T = 55 \text{ "true" nonfatal incident cases}$$

To obtain <u>total</u> new incident cases we add the new cases to the CHD deaths (assumed classified correctly):

$$55 + \text{CHD deaths} = 55 + 56 = 111 \text{ new cases}$$

and substitute the new number into the computation in part b, above:

$$87\text{-month CI} = \frac{111}{2{,}529} = 0.022 = 22 \text{ per } 1{,}000$$

The above computations can be summarized in the following table, which compares the true diagnosis to the study diagnosis assuming 70% sensitivity and 98% specificity for participants who were re-examined in 1967. For example, the upper left-hand cell has the number of persons with CHD (T) who were counted as such:

		"True" diagnosis		
		CHD	$\overline{\text{CHD}}$	Total
Study diagnosis	CHD	$(Se) \times T$ $(0.70) \times T$	$(1 - Sp) \times (2{,}473 - T)$ $(0.02) \times (2{,}473 - T)$	87
	$\overline{\text{CHD}}$	$(1 - Se) \times T$ $(1 - 0.70) \times T$	$Sp \times (2{,}473 - T)$ $(0.98) \times (2{,}473 - T)$	2,386
	Total	T	$(2{,}473 - T)$	2,473*

Given: Se = 0.70, Sp = 0.98

* (Only survivors were re-examined.)

d. The lowest that specificity **in 1960-62** could have been given these data can be found by supposing that all prevalent cases were false positives. In that worst case scenario, the following relationships would hold:

$Se \times T + (1 - Sp) \times (3{,}102 - T) = 93$ observed cases in 1960-62

If there were no true cases, then $T = 0$, $(1 - Sp) \times (3102) = 93$, and $Sp = 0.97$

So the examination procedures in 1960-62 must have achieved at least 97% specificity.

11. Multicausality: Confounding

Accounting for the multicausal nature of disease – secondary associations and their control

Introduction

When "modern epidemiology" developed in the 1970s, Olli Miettinen organized sources of bias into three major categories: selection bias, information bias, and confounding bias. If our focus is the crude association between two factors, selection bias can lead us to observe an association that differs from that which exists in the population we believe we are studying (the target population). Similarly, information bias can cause the observed association to differ from what it actually is. Confounding differs from these other types of bias, however, because confounding does not alter the crude association. Instead, concern for confounding comes into play for the interpretation of the observed association.

We have already considered confounding, without referring to it by that term, in the chapter on age standardization. The comparison of crude mortality rates can be misleading, not because the rates are biased, but because they are greatly affected by the age distributions in the groups being compared. Thus, in order to be able to interpret the comparison of mortality rates we needed to examine age-specific and age-standardized rates in order avoid or equalize the influence of age. Had we attemped to interpret the crude rates, our interpretation would have been **confounded** by age differences in the populations being compared. We therefore **controlled for** the effects of age in order to remove the confounding. In this chapter we will delve into the mechanics of confounding and review the repertoire of strategies to avoid or control it.

Counterfactual reasoning

Epidemiologic research, whether descriptive or analytic, etiologic or evaluative, generally seeks to make causal interpretations. An association between two factors prompts the question what is responsible for it (or in the opposite case, what is responsible for our not seeing an association we expect). Causal reasoning about associations, even those not the focus of investigation, is part of the process of making sense out of data. So the ability to infer causal relationships from observed associations is a fundamental one.

In an "epidemiologists' ideal world", we could infer causality by comparing a health outcome for a person exposed to a factor of interest to what the outcome would have been in the absence of exposure. A comparison of what would occur with exposure to what would occur in the absence of exposure is called counterfactual, because one side of the comparison is contrary to fact (see Rothman and Greenland, p49, who attribute this concept to Hume's work in the 18th century). This counterfactual comparison provides a sound logical basis for inferring causality, because the effect of the exposure can be isolated from the influence of other factors.

In the factual world, however, we can never observe the identical situation twice, except perhaps for "instant replay", which does not allow us to alter exposure status. The plethora of factors that can influence an outcome vary from person to person, place to place, and time to time. Variation in these factors is responsible for the variability in the outcomes we observe, and so a key objective in both experimental and observational research is to minimize all sources of variability other than the one whose effects are being observed. Only when all other sources of variability are adequately controlled can differences between outcomes with and without the exposure be definitively attributed to the exposure.

Experimental sciences

Experimental sciences minimize unwanted variability by controlling relevant factors through experimental design. The opportunities for control that come from laboratory experimentation are one of the reasons for their power and success in obtaining repeatable findings. For example, laboratory experiments can use tissue cultures or laboratory animals of the same genetic strain and maintain identical temperature, lighting, handling, accommodation, food, and so forth. Since not all sources of variability can be controlled, experiments also employ control groups or conditions that reflect the influence of factors that the experimenter cannot control. Comparison of the experimental and control conditions enables the experimenter to control analytically the effects of these unwanted influences.

Because they can manipulate the object of study, experiments can achieve a high level of assurance of the equivalence of the experimental and control conditions in regard to all influences other than the exposure of interest. The experimenter can make a before-after comparison by measuring the outcome before and after applying an "exposure". Where it is important to control for changes that occur with time (aging), a concurrent control group can be employed. With randomized assignment of the exposure, the probability of any difference between experimental and control groups can be estimated and made as small as desired by randomizing a large number of participants. If the exposure does not have lingering effects, a cross-over design can be used in which the exposure is applied to a random half of the participants and later to the other half. The before-after comparison controls for differences between groups, and the comparison across groups controls for changes that occur over time. If measurements can be carried out without knowledge of exposure status, then observer effects can be reduced as well. With sufficient control, a close approximation to the ideal, counterfactual comparison can be achieved.

Comparison groups

In epidemiology, before-after and cross-over studies are uncommon, partly because the exposure often cannot be manipulated by the investigator; partly because of the long time scale of the processes under study; and partly because either the exposure, the process of observation, or both often have lasting effects. The more usual approximation to a counterfactual comparison uses a comparison group, often called a "control group" on analogy with the experimental model. The comparison group serves as a surrogate for the counterfactual "exposed group without the exposure". Thus, the adequacy of a comparison group depends upon its ability to yield an accurate

estimate of what the outcomes would have been in the exposed group in the absence of the exposure.

Randomized trials

The epidemiologic study design that comes closest to the experimental model is the large randomized, controlled trial. However, the degree of control attainable with humans is considerably less than with cell cultures. For example, consider the Physicians Health Study, in which Dr. Charles Hennekins and colleagues at Harvard University enrolled U.S. physicians (including several faculty in my Department) into a trial to test whether aspirin and/or beta carotene reduce risk of acute myocardial infarction and/or cancer. The study employed a factorial design in which the physicians were asked to take different pills on alternate days. One group of physicians alternated between aspirin and beta carotene; another group alternated between aspirin and a placebo designed to look like a beta carotene capsule; the third group alternated between an aspirin look-alike and beta carotene; and the fourth group alternated between the two placebos). In this way the researchers could examine the effects of each substance both by itself and with the other – two separate experiments conducted simultaneously.

With 20,000 participants, this study design ensured that the four groups were virtually identical in terms of baseline characteristics. But there was clearly less control over physicians during the follow-up period than would have been possible with, say, laboratory rats. For example, the physician-participants may have increased their exercise levels, changed their diets, taken up meditation, or made other changes that might affect their disease risk. Such changes can render a study uninformative.

The MRFIT debacle

Just such an unfortunate situation apparently developed in the Multiple Risk Factor Intervention Trial (MRFIT), a large-scale (12,000 participants, over $100 million) study sponsored by the National Heart, Lung, and Blood Institute (NHLBI) of the U.S. National Institutes of Health (NIH). As evidence mounted that blood cholesterol was an etiologic risk factor for multiple forms of cardiovascular disease, particularly coronary heart disease (CHD), the possibility for a trial to verify that changing cholesterol levels would reduce CVD was being intensively explored. However, in the late 1960's suitable drugs were not available; the only cholesterol-lowering intervention was dietary modification. A "diet-heart" trial would require over one million participants and last many years – not an appealing scenario.

The idea of a diet-heart trial persisted, however, eventually metamorphosizing into a study to verify that cardiovascular disease rates could be lowered by changing the three most common CVD risk factors: cigarette smoking, elevated serum cholesterol, and hypertension. Thus was born MRFIT.

The trial was launched in the early 1970's. Men (because they have higher CHD rates) whose risk factors placed them at high CHD risk (based on a model from the Framingham Study) were randomized to "Special Intervention" (SI) or Usual Care (UC). SI participants received intensive, state-of-the-art, theoretically-based interventions to improve diet and promote smoking cessation.

Hypertensive SI participants were treated with a systematic protocol to control their blood pressure. UC participants had copies of their regular examinations sent to their personal physicians, but received no treatment through MRFIT. In this pre-"wellness" (health promotion / disease prevention through individual behavior change) era, the trial's designers projected modest risk factor changes in SI participants and little if any change in UC participants. Even though UC participants' physicians were to receive examination results, in those years few practicing physicians became involved in dietary change, smoking cessation, or even blood pressure control for healthy patients.

The planned sample size of about 12,000 men, about 6,000 in SI and 6,000 in UC, was achieved, and follow-up was maintained for seven years. By the end of the follow-up period, risk factor levels in the SI group had reached the target levels, and 46% of SI smokers quit smoking. But to the surprise (and consternation) of the MRFIT investigators, cholesterol levels and blood pressures also declined among UC participants, and 29% of UC smokers quit. During the years of the trial, smoking, diet, and hypertension had risen on the agendas of both the medical profession and the public (presumably aided by another NHLBI initiative, the National High Blood Pressure Control Program). Mortality among the UC participants was not only considerably lower than the projection based on data from the Framingham study but was even (slightly) below that for SI participants. Needless to say, there were many uncomfortable epidemiologists when the results came out.

Nonrandomized studies

Most epidemiologic studies do not have the opportunity to compare groups formed by a random assignment procedure. Whether we study smoking, alcohol, seat belts, handgun ownership, eating, exercise, overweight, use of particular medications, exposure to toxic agents, serum cholesterol, blood pressure, air pollution, or whatever, there is no assurance that the comparison group (the unexposed participants) is just like the exposed participants except for the exposure under study. Indeed, the opposite is more likely, since all sorts of factors are related to family and physical environment, occupation (e.g., workplace exposures), lifestyles (e.g., nutrition, physical activity), social influences (e.g., social support, injustice), health care, health conditions (e.g., medications), genetic endowment, and other characteristics.

Confounding

Thus, whenever we compare groups with respect to factors of interest, we must always consider that group differences in other, "extraneous" factors could be responsible for what we observe (or do not observe) (extraneous factors = factors other than the relationships under study). **Confounding** (from the Latin *confundere*, to mix together) can be defined as a "situation in which a measure of the effect of an exposure on risk is distorted because of the association of exposure with other factor(s) that influence the outcome under study" (Last, *A dictionary of epidemiology*). Confounding is a problem of comparison, a problem that arises when extraneous but important factors are differently distributed across groups being compared. The centrality of the concept of confounding and its control in epidemiology derives from the limited opportunities for experimental control.

A hypothetical example (with apologies to the Western Collaborative Group Study)

To investigate how confounding can arise and how it can be dealt with, consider the following hypothetical data based on the Western Collaborative Group Study of coronary heart disease (CHD) risk in managers and white collar workers exhibiting the coronary prone behavior pattern. This pattern, most often referred to as the Type A behavior pattern, is described as hard-driving, time-urgent, and hyperaggressive. In contrast, Type B people are regarded as more relaxed and easy-going.

In this study, Meyer Friedman, Raymond Rosenman, and their colleagues recruited 3,154 white male managers, aged 39-59, employed at ten California companies. The men were given medical examinations for CHD and a standardized, structured interview to determine their behavior type. Behavior type was determined by reviewing videotapes of the interviews. The 2,648 participants judged to be free of CHD at baseline were followed-up with annual physical examinations to detect new CHD cases during the subsequent 8-1/2 years. The (actual) results of the study are shown in the following diagram and are tabulated in Table 1.

Figure 1
Western Collaborative Group Study 8-1/2 year follow-up

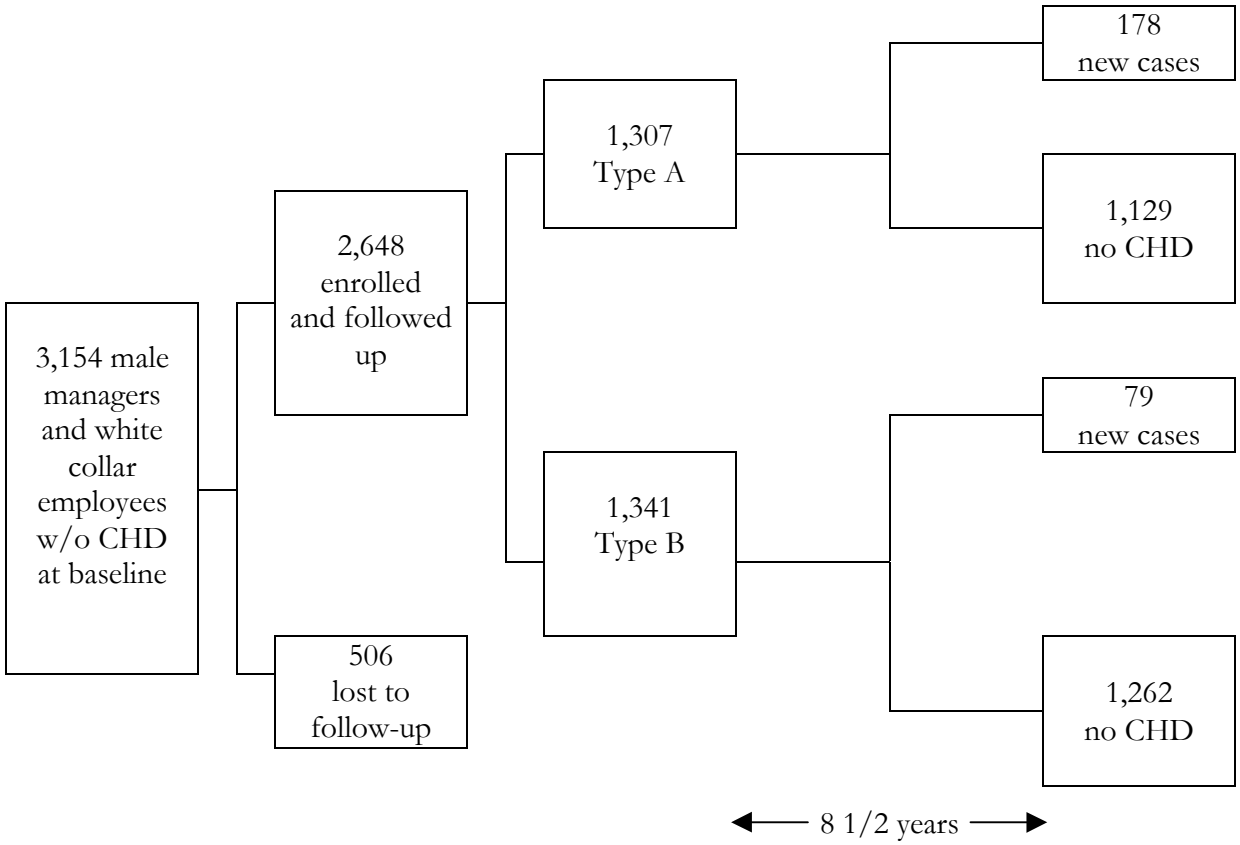


Table 1
Relationship of CHD to Behavior Pattern

	Behavior pattern		
	A	B	Total
CHD cases	178	79	257
No manifest CHD	1,129	1,262	2,391
Total	1,307	1,341	2,648

Since these data come from a cohort study, we would analyze them by estimating the incidence of CHD for the Type A and Type B groups. Even though the risk period for CHD extends beyond the period of observation, we will use cumulative incidence (CI) for simplicity. In these data, the CI is 178/1307 = 0.14 for the Type A group, and 79/1341 = 0.06 for the Type B group. The relative risk (risk ratio, cumulative incidence ratio) is therefore 0.14/0.06 = 2.3

Questions to ask:

There are many aspects of the design and conduct of this study that we would want to inquire about. For example:

> What were the criteria for classifying participants as Type A or Type B?
>
> How many participants were lost to follow-up?
>
> How was CHD defined and diagnosed?
>
> Were the physicians who determined whether a participant was a new case or not aware of the participant's behavior type?

But since our topic today is confounding, we are most interested in the question:

> Do the Type A and Type B groups differ in other factors that might have affected their observed CHD rates?

or, equivalently,

> Are there factors other than behavior pattern that may have been responsible for the observed rates?

(It might be interjected here that the same question would be relevant whether a difference between Type A and Type B had been observed or not).

Hypothetical scenario

Probably most of you know that in the Western Collaborative Group Study, no other factors seemed to explain the difference in CHD incidence between Type A and Type B groups. So here we will depart from the actual study in order to create a scenario in which the difference in the observed incidence for Type A and Type B participants is actually due to differences in cigarette smoking.

Suppose we had obtained the data in the Table 1. How could we see whether the difference in incidence between Type A and Type B groups should be attributed to differences in smoking rather than to behavior type? The traditional and most common approach to answering this question is to break down or stratify the data by cigarette smoking status of the participants. Table 2 shows the results of such a stratified analysis (with hypothetical data).

Table 2
Relationship of CHD to Behavior Pattern,
Stratified Analysis Controlling for Smoking Status [HYPOTHETICAL DATA]

	Smokers		Nonsmokers	
	Type A	Type B	Type A	Type B
CHD	168	34	10	45
$\overline{\text{CHD}}$	880	177	249	1,085
Total	1,048	211	259	1,130

This table shows the relationship between behavior type and CHD, stratified by smoking experience. Now we can compute the (cumulative) incidence of CHD among Type A nonsmokers and compare that to Type B nonsmokers, which will tell us the effect of behavior type when smoking could not possibly account for the results (not counting environmental tobacco smoke). We can also look at the incidence for Type A smokers and Type B smokers, where again we have (to some extent) created groups that are more comparable.

What do we see when we do these calculations? The incidence of CHD among Type A nonsmokers is 10/259 = 0.04, exactly the same as that among Type B nonsmokers (45/1130 = 0.04). We are therefore led to the conclusion that at least among nonsmokers, behavior pattern made no difference. Similarly, the cumulative incidence is the same (0.16) for Type A smokers and Type B smokers. Again, behavior pattern made no difference. Smoking, apparently, made a big difference. This key "extraneous" variable was apparently very unevenly distributed between the two behavior pattern groups and led to our observing a difference we nearly attributed to behavior pattern.

Confounding – a discrepancy between the crude and the controlled

This example illustrates confounding. In the uncontrolled or "crude" table, we saw an association (CIR of 2.3). When we controlled for smoking (which we will assume for the present is the only relevant extraneous variable), we find that there was no association (CIR of 1.0) between our study factor (behavior pattern) and the outcome (CHD). This discrepancy between the crude CIR (2.3) and the stratum specific CIR's (1.0) indicates that there is confounding by smoking status. Stratification is one method of controlling for the confounding effect of smoking. [Please let me emphasize here that the above example is **not** true to life. In the actual study by Friedman and Rosenman, Type A behavior was found to be associated with CHD even when the effects of smoking and other known CHD risk factors were controlled.] It may also be worthwhile to mention that confounding could also happen in the reverse manner, that is, we might see no association in the crude analysis but find that there is one when we stratify. So confounding can create an apparent association or mask a real one.

Confounding arises from unequal distribution of a risk factor

How can the phenomenon of confounding occur? As indicated above, the conditions needed to create confounding (in this rather simplified situation) are that a true risk factor for the health outcome is unevenly distributed between the groups being compared. To see this in the above example, I have rearranged the columns from Table 2. This rearrangement emphasizes that most of the Type A's were smokers and most of the Type B's were not.

Table 3
Relationship between CHD, Behavior Pattern, and Smoking Status
[HYPOTHETICAL DATA]

	Type A behavior pattern			Type B behavior pattern			Both
	Smokers	Non smokers	**Total**	Smokers	Non smokers	**Total**	Grand total
CHD	168	10	**178**	34	45	**79**	257
$\overline{\text{CHD}}$	880	249	**1,129**	177	1,085	**1,262**	2,391
Total	1,048	259	**1,307**	211	1,130	**1,341**	2,648

Although this table was created by rearranging columns in Table 2, it may be more revealing to think of it as providing the underlying story for the uncontrolled (crude) data in Table 1. Notice that Table 1 is contained in this table as the marginals for each of the two subtables (the bolded columns). The subtables show the composition of the Type A group and the Type B group. Clearly, the overwhelming majority (1048/1307 = 80%) of the Type A participants are smokers, whereas the overwhelming majority (1130/1341 = 84%) of the Type B participants are nonsmokers. With such a marked imbalance, it should not be surprising that a risk factor such as smoking could distort the overall (uncontrolled) association. The attributes of a confounder, then, are that it is an independent risk factor for the outcome and is associated with the study factor.

Confounding – misattribution of an observed association

The excess of cases in the Type A group is due, clearly, to the greater proportion of smokers in the Type A group than in the Type B groups. Were we to have gone with the crude value, we would have misattributed the observed difference between groups to behavior pattern rather than to smoking. Confounding can be defined as a distortion in the measure of association due to the unequal distribution of a determinant of the outcome.

Note, however, that the crude association is still "real". The type A participants *did* have a greater incidence of CHD. Confounding arises when we *attribute* that elevated incidence to their being type A, since the higher incidence is really due to their smoking (in this example). But the type A men as

a group did indeed have higher CHD incidence. There are situations where the crude association remains important to consider.

Another perspective – weighted averages

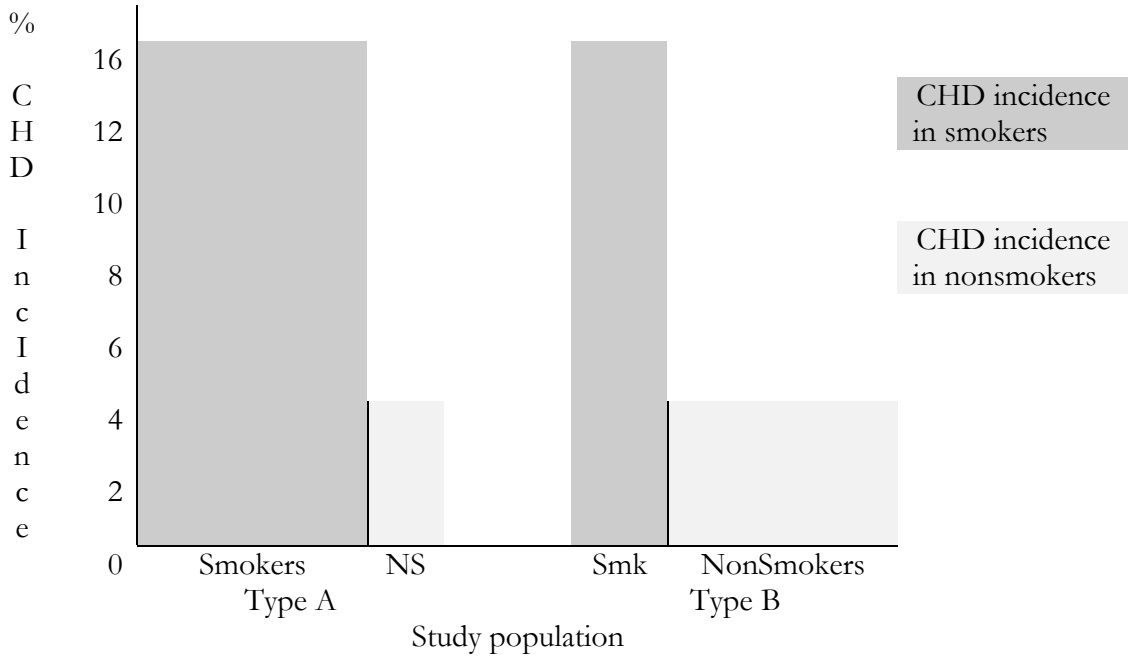
A summary table highlights the incidences and makes the pattern very evident.

Table 4
Incidence of CHD by Behavior Type and Smoking Status
[HYPOTHETICAL DATA]

	Smoking status			
Behavior pattern	Smoker	Nonsmoker	Total	
Type A	0.16	0.04	0.14	← (incidences
Type B	0.16	0.04	0.06	← from table 1)
Total	0.16	0.04		

Here it is very clear that when we hold smoking constant (i.e., look down either of the first two columns of incidences), there is no effect of behavior type. When we hold behavior type constant (i.e., look across either of the first two rows), we see that smoking is associated with a fourfold increase in incidence. The marginals of the table are, in effect, weighted averages of the incidences in the interior of the table. The incidences in the bottom row are the same as in the interior of the table – they have to be, because a weighted average of two identical numbers is always that number. The incidences in the rightmost column, however, could be almost any numbers between 0.16 and 0.04 – depending upon the weighting used in averaging 0.16 and 0.04. These concepts can be shown graphically.

CHD Incidence by Behavior Pattern and Smoking Status
[HYPOTHETICAL]



As the diagram shows, the study population can be viewed as consisting of four distinct subgroups, each with a different combination of behavior type and smoking status. If these were the only relevant subgroups, then the incidence rates for each would represent the irreducible "true" state in the study population. The rate for the study population as a whole and for any group in it, e.g., all Type A's, may be regarded as a weighted average of the incidences in the component subgroups, where the weights are the proportional sizes of the component subgroups. Thus the rate in the Type A's is:

$$0.14 = \frac{178}{1{,}307} = \frac{1{,}048}{1{,}307} \times \frac{168}{1{,}048} + \frac{259}{1{,}307} \times \frac{10}{259}$$

or symbolically,

$$CI_{CHD|A} = P_{S|A} \times CI_{SA} + \bar{P}_{S|A} \times \bar{CI}_{SA}$$

where:

CI is (cumulative) incidence

P is prevalence or proportion

$\bar{S}$ indicates smokers ($\bar{S}$ indicates nonsmoker)

A indicates behavior Type A

and the notation S|A means "smokers among (or given) Type A behavior".

Confounding – comparison of weighted averages using different weights

The incidence for any group (e.g., Type A's) can vary from the lowest incidence of any of its subgroups (e.g., nonsmoker Type A's) to the highest incidence of any subgroup (e.g., smoker Type A's). Where in this range the overall group's incidence falls is determined by the size of each subgroup (Type A smokers, Type A nonsmokers) as a proportion of the overall group (all Type A's). Confounding can result when these proportions differ for groups that are being compared.

Since there are many possible ways in which these proportions can differ, confounding can cause an overall (crude) measure of association to overstate, understate, completely obscure, or even invert the association that would be seen in comparisons carried out within the subgroups. As a familiar example, if two populations have different age distributions, then a comparison of their overall (crude) death rates can overstate or understate the picture seen by comparing within specific age groups, even to the point that the comparison of crude rates appears to favor the population that has higher (worse) death rates within each age stratum. Age standardization is a special case of the more general strategy called stratified analysis, which is one primary recourse for controlling confounding.

The limits to confounding

There are limits on the strength of the (secondary) association that can be produced by confounding. For example, given the data in Table 1, a strong effect for smoking and a striking imbalance between the two behavior type groups was required in order for smoking to account completely for the apparent effect of Type A behavior. That is one of the reasons why strength of association is a criterion for causal inference. The stronger the observed association between the disease and the study factor, the less likely that some completely extraneous factor could account for all of the observed association.

Case-control studies

So far in our discussion we have confined ourselves to cohort-type studies. When we turn to the issue of confounding in case-control studies, there are some additional complexities as a consequence of the way in which the base population is represented in the study population. To understand the characteristics of confounding in a case-control study, let us generate such a study from the cohort we considered earlier.

The original cohort consisted of 2,648 individuals with complete follow-up and yielded 257 cases. Ideally, our case-control study would detect all incident cases and would sample from non-cases as the cases occurred (called "density sampling"). To simplify our illustration, however, let us sample our controls from those individuals who were free from CHD at the end of the follow-up period. The following table shows the same cases, with the distribution of controls expected from obtaining a representative sample from the noncases, of size twice the number of cases (i.e., assume 514 controls with the same proportion of Type A's and smokers as found in all noncases in the cohort study). (The numbers in the "No CHD" row are obtained by multiplying the "No CHD" row in Table 1 (i.e., all the noncases) by 514/2391 (0.21) so that the 2,391 noncases become 514 controls.

In this way, the 1,129 Type A's without manifest CHD become 243 Type A controls, and the 1,262 Type B's without manifest CHD become 271 Type B controls.)

Table 5
Expected Results from Case-Control Study [HYPOTHETICAL]

	Behavior pattern			
	Type A	Type B	Total	
CHD cases	178	79	257	
No manifest CHD	243	271	514	← This row is simply 0.21 times the corresponding row in Table 1.
Total	421	350	771	

The odds ratio for this table is [2.5], slightly larger than the risk ratio in the cohort study. [The difference between the odds ratio and risk ratio reflects the CHD incidence in the cohort – the smaller the incidence, the closer the odds ratio would be to the risk ratio.]

Now let us generate, in the same manner, the expected table for smoking and behavior pattern in a stratified analysis:

Table 6
Expected Results for Case-Control Study, Stratified by Smoking Status [HYPOTHETICAL]

	Smokers		Nonsmokers		
	Type A	Type B	Type A	Type B	
CHD	168	34	10	45	
CHD	189	38	54	233	← This row is simply 0.21 times the corresponding row in Table 2.
Total	357	72	64	278	

The odds ratios for each table are 1.0, so confounding is again present. Here again we see that the confounding factor is associated with the outcome: the odds ratio for smoking and CHD in the Type B group is 4.6. We also find that smoking is associated with behavior type: the proportion of smokers among Type A noncases is 0.78 whereas among the Type B noncases it is only 0.14 [verify these numbers].

The reason for the above emphasis on **conditional** associations ("in the Type B group", "among noncases") rather than **unconditional** or crude associations is that a confounding variable must be associated with the exposure under study in the population from which the cases arise (see Rothman and Greenland). It is the control group that provides the estimate of exposure prevalence in the source population. Also, in a case-control study, the totals for different exposure groups (e.g., total

Type A smokers) are not very meaningful quantities, at least for comparison purposes. The reason is that the relationships among these totals largely reflect the (arbitrary) ratio of cases to controls. So the association of exposure that is relevant for confounding in a case-control study is the association between exposure and the potential confounder among the controls.

The reason for not looking within the Type A group is that an association in this group could reflect effect modification between the exposure (Type A behavior) and the covariable, rather than confounding as such. We will elaborate on this matter when we take up effect modification, in the next chapter.

Confounding – a characteristic of the study base

We have said that confounding requires two associations: (1) the confounder must be a risk factor for the outcome or its detection and (2) the confounder must be associated with the exposure. The latter association must exist within the study base (see Rothman and Greenland). This point merits elaboration.

Follow-up study

In a follow-up study, the study base, from which the cases arise, is simply the population being followed, the study population. For confounding to occur, the exposure and potential confounder must be associated in this population. Randomized assignment of an intervention tends to distribute potential confounders evenly across intervention and control groups. To the extent that randomized assignment succeeds, i.e., no extraneous variables will be associated with the intervention, so confounding cannot occur. If, however, the randomization does not "work" so that an imbalance exists for a particular potential confounder, then confounding with respect to that potential confounder can occur. The greater the number of participants, the less likely that any meaningful imbalance will occur by chance.

Case-control studies

In a case-control study, the study base is the underlying population that is being followed through the window of the case-control design. For confounding to occur, the exposure and potential confounder (risk factor) must be associated in that underlying population (source population from which cases arise). But since the investigator observes that population only indirectly, the matter is trickier. However, if there is no association between the potential confounder and exposure in the study base, then confounding does not occur even if we do find the potential confounder and exposure to be associated within the control group of our case-control study (Miettinen and Cook, cited in Rothman, page 93).

This somewhat surprising result is easily illustrated. Suppose we are observing a population over time to examine an association between a suspected occupational carcinogen and a cancer that is also strongly (IDR=10) related to cigarette smoking. Suppose also that the occupational exposure is in fact a carcinogen and that in this population smoking is not associated with the occupational

exposure. If we assume a baseline rate of 3 cases/1,000 person-years and an IDR of 3.3 for the occupational carcinogen, the follow-up of the population might produce the following table.

Incidence rates, population sizes, and number of cases for hypothetical data on an occupational exposure and smoking

	Smokers		Nonsmokers	
	Exposed	Unexposed	Exposed	Unexposed
	(1)	(2)	(3)	(4)
1. Number of cases	300	90	70	21
2. Population size (person-years)	3,000	3,000	7,000	7,000
3. Incidence density per 1,000 py	100	30	10	3
	IDR = 3.3		IDR = 3.3	

With a hypothetical 7,000 person-years of observation for nonsmokers who are also not exposed to the carcinogen, the assumed baseline incidence rate of 3/1,000 py will produce an expected 21 incident cases. If the amount of person-time among exposed nonsmokers is also 7,000 py, then we would expect 3.3 × 3/1,000 py × 7,000 py ≈ 21 cases for that group. If person time for exposed and unexposed smokers is 3,000 py for each group, then we expect 300 and 90 incident cases, respectively, if the IDR for the occupational exposure is the same among nonsmokers and smokers and the IDR for smoking is 10, regardless of occupational exposure.

Note that this hypothetical population has been constructed so that the proportion of exposed person-years is 50% among smokers (columns 1 and 2), among nonsmokers (columns 3 and 4), and overall, i.e., no association between smoking and the occupational exposure. Similarly, the proportions of person-years for smokers among exposed (columns 1 and 3) and unexposed (columns 2 and 4) are each 30% (3,000/[7,000+3,000]). The crude IDR for the occupational carcinogen is therefore 3.3 (be certain that you can derive this IDR), which is identical to the IDR for the exposure among smokers and among nonsmokers. Thus, confounding is not present.

Suppose now that we were to conduct a case-control study in this population during the same period of time. If there is a cancer registry we might hope to identify and include all 481 cases (see row 1 in the following table, which is identical to row 1 in the preceding table). If we obtain a 5% representative sample of the population as our control group, then the distribution of smoking and the occupational carcinogen in our control group (row 2 in the following table) will be the same as the distribution of these variables in the population-time in row 2 of the preceding table (30% smokers and 50% exposed to the occupational carcinogen, with no association between these two). The OR (be certain that you can calculate this) will be identical to the IDR of 3.3, above. In this case-control study with an (**unbiased**) control group that is directly proportional to the study base, there is no confounding.

Different control groups for hypothetical case-control study of an occupational exposure and smoking

	Smokers		Nonsmokers	
Row	Exposed	Unexposed	Exposed	Unexposed
#	(1)	(2)	(3)	(4)
1. Number of cases	300	90	70	21
2. Proportional controls	150	150	350	350
	(OR = 3.3)		(OR = 3.3)	
3. Biased controls	250	150	250	350
	(OR = 2.0)		(OR = 4.7)	

Suppose, however, that controls are selected in a biased fashion, producing a **biased control group** (row 3 in the second table) in which smoking and exposure **are** associated (verify this fact; try, for example, computing the OR for smoking in relation to exposure). Reflecting the biased control group, the stratum-specific IDR's are no longer 3.3. However, in this chapter our focus is the **crude association** and whether it accurately represents the true situation (which in this instance we constructed, rather than having to regard the stratified associations as the true situation). The crude OR from the above table, using the cases on row 1 and controls from row 3, is (do try computing this before reading the answer) (370 × 500) / (111 × 500) = 3.3.

Thus, even with this biased control group the crude OR remains unconfounded. Yet, the potential confounder (smoking, a causal risk factor for the outcome) is indeed associated with the exposure in the (biased) controls. [Several ways to see this association are:

The odds of exposure among smokers (cols. 1 and 2) are 250/150, quite different from the odds of exposure among nonsmokers (cols. 3 and 4: 250/350), producing an odds ratio between smoking and exposure of OR = 2.3).

Proportionately more smokers are exposed [250/(250 + 150) = 0.63] than are nonsmokers [250/(250 + 350) = 0.42].

The odds of smoking among exposed (cols. 1 and 3) are 250/250, quite different from the odds of smoking among the unexposed (cols. 2 and 4): 150/350), producing, of course, the same odds ratio, 2.3).

Proportionately more exposed are smokers [250/(250+250) = 0.5] than are unexposed [150/(150 + 350) = 0.3].

The potential confounder, smoking, is also associated with the outcome in the unexposed (e.g., IDR = 30 per 1,000py / 3 per 1,000py in the study base, OR = (90 × 350) / (21 × 150) in the case-control study with either control group. Thus, it is possible to have a risk factor that is associated with exposure in the noncases yet not have confounding.

Further insight can be gained by considering the mechanism that causes confounding, as illustrated in the Type A behavior example. Confounding results from an imbalance between exposed and unexposed groups in regard to a disease determinant. If the potential confounder increases disease risk and the potential confounder is associated with the exposure, then incidence of disease in the exposed will be boosted relative to that in the unexposed, due to the confounder. This disproportionate increase in incidence, and therefore in cases, will increase the odds of exposure in a representative group of cases. If the confounder is not controlled in the analysis, this increased odds will cause confounding of the exposure-disease association.

The (exposure) OR for the outcome is simply the ratio of the exposure odds in the case group divided by the exposure odds in the control group. The exposure odds in the case group is obviously not affected by anything that happens to the control group (including matching, incidentally). So a distortion of the crude OR will have to come from a change in the exposure odds in the control group. So long as the bias in the control group does not cause its **crude** exposure odds to differ from those in the source population (e.g., 0.5/0.5=1.0 in our occupational carcinogen example), the crude OR will remain the same as in the source population, i.e., unconfounded.

In most case-control studies we have little independent information about the study base, so the control group provides our window into the study base. If the control group is biased, then our view of the study base is distorted, and we may conclude that the condition for confounding (i.e., a risk factor for the disease is associated with the exposure in the noncases) is met. Due to such a biased control group, controlling for the potential confounder will introduce bias in this analysis (e.g., in the above example, the stratum-specific OR's are different from the correct value of 3.3). However, a weighted average of stratum-specific OR's may be close to the crude value.

Statistical tests for confounding

Since confounding requires an association between the potential confounder and the exposure, investigators sometimes present statistical tests of the differences in potential confounders between exposure groups. If the groups do not differ significantly, the investigators conclude that confounding will not occur. This practice will often yield a correct conclusion, though it is somewhat off the mark.

Statistical tests of significance address the question of whether or not there is an association between the exposure and the potential confounders beyond that likely to arise by chance alone. But confounding depends upon the magnitude of association (e.g., odds ratio, prevalence ratio), rather than on the strength of evidence that it did not arise by chance. So a large but "nonsignificant" difference can have more potential to cause confounding than a small but "highly significant" difference. The reason for this apparently paradoxical statement is that statistical significance depends upon the magnitude of the number of exposed and unexposed participants, so that nearly any association will be statistically significant if the study is sufficiently large and nonsignificant if it is sufficiently small. The presence or extent of confounding, however, is not affected by scaling up or down the number of participants.

Confounding, then, is a function of the magnitude of associations, rather than of their statistical significance. Since strong associations are likely to be statistically significant, statistical tests comparing exposed and unexposed groups can be a convenient device for identifying associations that may be strong enough to cause confounding, which is why the procedure often yields the correct conclusion about the need to control for confounding. Some (see Rothman and Greenland) have suggested using significance testing with a value for alpha (Type I error probability) of 0.20, to increase the power to detect differences that may be important in regard to confounding. But as a guide to likely confounding, statistical tests are somewhat beside the point. (There is a subtle but valuable distinction to be made between statistical tests to evaluate confounding and statistical tests to assess whether randomized allocation to treatment or control "worked". Since randomized allocation attempts to operationalize "chance", the number and size of observed differences between treatment and control groups should not often exceed what we expect from chance, which is precisely what statistical tests are designed to evaluate. If there are more differences than there "should be", that may indicate some problem in the implementation of the randomization. It would also be expected that control for these differences would be needed.)

Components of the crude relative risk

There are several other aspects of confounding that it will be instructive to consider. The first of these is a method, due to Miettinen (Miettinen OS: Components of the crude risk ratio. *Am J Epidemiol* 1972; 96:168-172) for allocating an observed association to a component due to confounding and a component due to the study factor of interest. According to Miettinen, the crude risk ratio (or odds ratio) may be regarded as the product of a "true" risk ratio and a component due to confounding. In the examples we have considered thus far, the whole of the observed association has been due to confounding, i.e., to the effect of smoking. But it is also possible to have an association that remains, though stronger or weaker, after the effects of a confounder have been removed.

The following hypothetical data illustrate Miettinen's concept. Suppose that you are carrying out a case-control study to investigate whether trihalogenated hydrocarbons that occur in chlorinated drinking water containing organic matter increase colon cancer incidence. You collect data on all cases in a multi-county region during several years and assemble a control group using random-digit dialing. You interview cases and controls about their source of drinking water (treated surface water versus well or bottled water) and, because other studies have suggested that some unknown factor in urban living increases colon cancer incidence, you also collect data on urban-rural residence. The crude analysis of your data yields the following table:

Table 7a
Colon cancer and drinking water (hypothetical case-control data)

	E	$\bar{E}$	Total
Colon cancer cases	170	80	250
Controls	80	170	250
Total	250	250	500

The crude OR for this table is (170 x 170) / (80 x 80) = 4.5. Is it confounded by rural-urban residence?

We can investigate confounding by stratifying the data by urban-rural residence and examining the stratum-specific OR's:

Table 7b
Colon cancer and drinking water (hypothetical case-control data)

	Rural			Urban			Crude	
	E	$\bar{E}$		E	$\bar{E}$		E	$\bar{E}$
D	20	30	D	150	50	D	170	80
$\bar{D}$	50	150	$\bar{D}$	30	20	$\bar{D}$	80	170

The OR's in both the rural and urban strata are 2.0, so we know that the crude OR is confounded – it overstates the "true" OR, making a moderate association appear as a strong one. How much of the crude OR can be attributed to confounding? Miettinen suggests that the OR due to confounding is the OR for the association that would be observed even if the exposure (trihalogenated hydrocarbons) had no effect on the outcome (colon cancer). If the exposure has no effect on the outcome, then whatever association remains in the crude analysis must be due entirely to confounding.

So to obtain the OR attributable to confounding, we can eliminate the true association between trihalogenated hydrocarbons and colon cancer. In the above example, we regard the stratum-specific tables as displaying the true relationship (i.e., we are assuming that there is no selection bias, or information bias and that the only potential confounder is rural-urban residence as a dichotomous variable measured without error). So we will "eliminate" the true association from the stratum-specific tables. Then we can combine the modified stratum-specific tables into a new crude table and compute a new crude OR. That OR must entirely reflect confounding, because the true association no longer exists.

Since the OR is the crossproduct ratio for the four cells of a table, we can change the OR by changing any cell of the table. By convention, we change the "a"-cell (exposed cases) to what it would contain if there were no association between the study factor and the disease. Here, if the D,E cell in the rural stratum contained 10 instead of 20, then the OR for the rural stratum would be 1.0, i.e., no association. Similarly, if the D,E cell in the Urban stratum contained ___ (your guess?) instead of 150, then the OR for that stratum would likewise be 1.0. The revised tables are shown below:

Table 7c
Modified tables for Colon cancer and drinking water

	Rural		Urban		Modified crude		Original crude	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	10	30	___	50	85	80	170	80
$\bar{D}$	50	150	30	20	80	170	80	170

The OR for the modified crude table, and therefore the component attributable to confounding, is 2.25. Interestingly, this figure is the same as the quotient of the original crude (4.5) and controlled odds ratios (2.0). Indeed, this relationship holds in general: the crude OR equals the product of the controlled OR and the component attributable to confounding:

$$\text{Crude odds (or risk) ratio} = \text{Component due to study factor} \times \text{Component due to confounding}$$

So the component (of the crude ratio) attributable to confounding is the degree of association "expected" from the distribution of the potential confounder (in this case, rural-urban residence), i.e., from the fact that the potential confounder is distributed differently in exposed and unexposed persons in the study base.

Another way to look at this relationship is that the component attributable to the effect of the study factor, i.e., the unconfounded association, can be written as:

$$\text{Component due to study factor} = \frac{\text{Crude odds (or risk) ratio}}{\text{Component due to confounding}}$$

So the component (of the crude ratio) attributable to the study factor, i.e., the unconfounded, or "true", association, can be regarded as the ratio of an "observed" association to an "expected" association. Expressing the relationship in this way is reminiscent of the standardized mortality ratio (SMR), which is also a ratio of an "expected" to an "observed". In fact, Miettinen refers to the controlled OR above (i.e., the component due to the study factor) as an "internally standardized odds ratio", which is simply the odds ratio version of the SMR. It is also interesting to note that the stratum-specific OR's are also ratios of "observed" to "expected", in that these OR's are equal to the

ratio of the observed number of exposed cases (the contents of the "a-cell") and the expected number in the absence of a true association.

By this point you may well be wondering how much of this you need to know to practice epidemiology or control for confounding. The answer is that this particular formulation is not essential, but seeing confounding from this perspective is another aid to understanding the closely interrelated concepts of confounding, stratified analysis, standardization, and even the counterfactual framework of causal inference. If the true causal comparison is between the experience in an exposed group and what their experience would have been in the absence of exposure, the SMR might be regarded as the most relevant adjusted measure of effect, since it is the ratio of the observed rate in the exposed group to the rate that would be expected for them if they were not exposed (assuming that the rates in the study population differ from those in the study population due only to the standardizing factor and the exposure).

Matched studies

Since confounding is a problem of comparison, a principal aim of study design is to obtain groups that are comparable with regard to determinants of the outcome. In experimental designs, this aim is perhaps the principal motivation for randomized assignment of the study factor. Since randomized allocation does not guarantee the equal distribution of all relevant factors (though in very large studies the probability of equal distribution is very high), **prestratification** (also called "**blocking**") may be employed to enforce identical distributions when sample size is small. Prestratification involves first placing participants into groups according to their configuration of risk factors and then performing separate randomizations within each group. The procedure generally increases statistical efficiency (degree of precision per trial participant) (see Rothman and Greenland, p161).

Follow-up studies

In a nonrandomized study, where the investigator does not have the opportunity to assign the study factor, the analogous procedure to prestratification is **matching**. In matching, the participants in the comparison group (i.e., the unexposed group in a follow-up study or the control group in a case-control study) are selected so as to resemble the **index group** (the exposed in a follow-up study or the cases in a case-control study) on one or more relevant factors. When the unexposed group in a follow-up study has been matched to the exposed group on all relevant factors, so that the two groups differ only in terms of exposure to the study factor of interest, then the incidences in the two groups can be compared with no danger of confounding by the matching variables. In practice, however, competing risks and/or loss to follow-up can introduce differences. For this and other reasons (see Rothman and Greenland, p160), matched cohort studies are not common.

In any case, neither prestratification nor matching is required to avoid confounding, since confounding can be controlled in the analysis of the study results – providing there is adequate overlap in risk factor distributions between groups. For this reason, the primary purpose of matching is to increase statistical efficiency by ensuring sufficient overlap (which therefore indirectly aids in controlling confounding).

Case-control study

In a case-control study the situation is, as usual, not as straightforward. Because of the nature of the case-control study design, matching does not avoid confounding by the matching factor(s). Moreover, by changing the composition of the control group, matching in a case-control study can even cause the crude (uncontrolled) analysis to be biased. How can this be?

Since a case-control study selects participants according to disease, matching means ensuring that the case and control groups are the same in respect to the potential confounders. However, as we saw earlier, confounding depends on the comparability of exposed and unexposed groups in the study base, not between cases and controls in the study population. Although ensuring that cases and controls are similar with respect to potential confounders may facilitate control for confounding (through greater statistical efficiency), matching controls to cases does not change the study base and thus cannot alter the exposure odds among cases. But confounding arises because the exposure odds in cases is influenced by a population imbalance in a cause of the outcome.

Furthermore, by selecting the control group in a way that makes it conform to the case group in age, sex, treatment facility, or other factors, the investigator can cause the overall control group to have a different prevalence of exposure than that in the study base, which the control group seeks to reflect. Of course, a matched control group can still provide a correct estimate of exposure prevalence within each configuration of risk factors. So there need be no problem as long as the analysis takes account of the matching. If the matched analysis and unmatched analysis yield the same results, then the unmatched analysis can be used, and for simplicity often is unless the matched analysis provides greater precision.

Example of matching in a case-control study

The following example may help to clarify these concepts. Consider another study of colon cancer and drinking water, presented in the following table. This time the stratum-specific population sizes and prevalences of exposure to chlorinated drinking water are presented, along with the number of cases and the prevalence of exposure among cases.

Colon cancer and drinking water (hypothetical data)

Residence	Population size	% of total pop. with chlorinated drinking water	# of colon cancer cases	% of cases with chlorinated water
Rural	400,000	20 %	30	40 %
Urban	600,000	80 %	90	90 %
Total	1,000,000	56 %	120	___ %

An investigator conducting a case-control study in this population and selecting community controls without matching, would observe an exposure prevalence of 56% (i.e., an average of the urban- and

rural-specific exposure prevalences, weighted by their respective population sizes: [0.20(400/1000) + 0.80(600/1000)]). In contrast, a control group matched to the distribution of cases would have an exposure prevalence of 65% [0.20(30/120) + 0.80(90/120)], since in this case the two prevalences are weighted by the proportions of rural and urban cases, rather than the proportions of rural and urban residents in the population.

The prevalence of exposure in the matched control group, 65%, is a distorted estimate of the overall prevalence of exposure in the population as a whole. But the estimate is not a problem when our analysis takes rural-urban residence into account, since the stratum-specific exposure prevalences are still correct and we know the proportions of rural and urban residents in the population. If the exposure prevalence (right-most column) is 40% in rural cases and 90% in urban cases, then the odds ratios are 2.67 (rural) and 2.25 (urban), 2.70 (crude, unmatched controls) and 1.85 (crude, matched controls). Thus neither the matched nor the unmatched controls give a correct OR for a crude analysis. In contrast, a stratified analysis that takes residence into account will yield a valid odds ratio estimate with either control group. (Suggestion: derive all of these OR's.)

For a fuller treatment of matching, see chapter 10 of Rothman and Greenland. According to these authors, though there are circumstances where it is clearly desirable or not desirable, the value of matching in case-control studies is a complex question.

Potential confounders versus actual confounders

An issue of considerable practical and theoretical importance is how to choose which variables to investigate as confounders? As we saw above, to be a confounder a variable must be associated with both the disease and the exposure. Thus when through matching in a cohort study we ensure that the distribution of potential confounders is identical in both exposure groups (i.e., there is no association between these variables and exposure), then the former cannot confound our results (assuming no bias from competing causes of death and other attrition mechanisms). Apart from that situation, we must control for potential confounders in the analysis of the study to see whether or not they have distorted the observed association (which implies that we have remembered to measure them!).

Investigation of whether a variable is a potential confounder or an actual confounder is thus generally a matter of empirical determination in our data. In practice, therefore, the question of whether or not variable X is a confounder is a side issue. Our primary concern is to obtain a valid estimate of the relationship between study factor and outcome. If we have to control we do; if we do not need to, we may not. In either case we are not particularly concerned, ordinarily, about concluding that such-and-such a variable is a confounder.

But which variables to regard as potential confounders, i.e., which variables must be measured and possibly controlled in order to obtain a valid estimate of the association between study factor and outcome, is a matter of first importance. Our decisions here depend upon our understanding of which variables other than our study factor might explain or account for an observed relationship (or lack thereof). Thus, the decision about whether a variable ought to be considered for control as a potential confounder rests first and foremost on our conceptual model.

First and foremost, a potential confounder must have some relationship to the occurrence of the disease or other outcome. The potential confounder must increase the probability that the disease will occur or must shorten the time until the disease occurs. If not, why should we attribute an observed association to that variable rather than to our study factor? (Since disease occurrence must be observed, a factor that affects disease detection may also qualify.) Furthermore, if the relevant variable occupies an intermediate position in the hypothesized causal chain linking the study factor to the disease, then again, how could that variable rather than the study factor be the "true" cause of an observed association? (If I persuade George to rob a bank and the police find out, can I persuade the judge to set me free because apart from what George did I did not rob anything?) Thus, in stratifying on smoking status in our Type A - CHD example, we are assuming that the association between Type A behavior and smoking arises due to a common antecedant cause (e.g., inadequate coping skills in a high-pressure occupational environment) or due to an effect of smoking status on behavior pattern, but not due to an effect of behavior pattern on smoking status, which would make smoking an intervening variable and therefore not appropriate for control in this way (Kaufman and Kaufman, 2001).

In practice, however, it is often difficult to make definite decisions about which variables are true risk factors, which are intervening variables, and so on, so that a cautious approach is to obtain data on as many potentially relevant variables as possible, explore the effects of controlling them in the analysis of the study, and then attempt to make sense out of the results. Consider, for example, a study of the effect of overweight on CHD incidence. Since overweight increases cholesterol and blood pressure levels, both of which are causal risk factors for CHD, then the crude association between overweight and CHD will reflect some combination of:

1. a direct effect of overweight on CHD if such exists,
2. an indirect of overweight on CHD due to the effect of overweight on cholesterol and blood pressure, which in turn increase CHD risk
3. possible confounding, if cholesterol and blood pressure are higher in people who are overweight not because of an effect of overweight but due to some other reason (e.g., diet, sedentary lifestyle, genetic factors).

Should we control for blood pressure and cholesterol when estimating the association between overweight and CHD? If we do not, then our measure of association will be distorted to the extent that confounding is present. If we do control by the usual methods, however, our measure of association will be distorted to the extent that overweight has its effects on CHD through increases on blood pressure and cholesterol.

For another example, consider the problem of studying whether sexually transmitted diseases such as gonorrhea increase the risk of acquiring HIV and whether condom use decreases the risk. Should the relationship between STD and HIV seroconversion be controlled for condom use? Should the relationship between condom use and HIV incidence be controlled for STD? Both condoms and STD appear to affect the risk of acquiring HIV infection, but condoms are also a means of preventing STD, which in that sense can be regarded as a variable located on the causal pathway from condoms to HIV. Furthermore, an obligatory causal factor for both sexually-acquired STD and/or HIV is sexual contact with an infected partner. "Risky" partners have a higher probability of

being infected, and the more of them, the greater the risk of exposure to the infection. Should we control for the number of risky partners in investigating the relationship among condoms, STD, and HIV? But risky partners are also a risk factor for STD, so that STD can be regarded as an intermediary variable between sex with risky partners and HIV. Thus, thinking through which variables to control for in a web of causation can itself be confounding! Greenland, Pearl, and Robins (1999) present a system of causal diagrams for describing and analyzing causal pathways to identify what variables must be controlled. Among other points, they explain that controlling for a variable can in some situations create confounding which would otherwise not occur. In general, control for confounding (and interpretation of data in general) is founded on assumptions of causal relationships involving measured and unmeasured variables. Data alone are inadequate to resolve questions of causation without these assumptions. Methodological understanding in this area is expanding (see the articles by Greenland, Pearl and Robins and Kaufman and Kaufman). However, limited knowledge of causal relationships in addition to the one under study and the likely existence of unmeasured but important variables will remain fundamental stumbling blocks for observational research..

Controlling for sociodemographic variables

Nearly all epidemiologic investigations control in some way or other for sociodemographic variables (e.g., age, gender, race, socioeconomic status). As we saw in the chapter on Standardization, comparisons that do not control for such variables can be very misleading. However, there are significant issues of interpretation of adjustment for sociodemographic factors, partly because associations with sociodeomographic factors likely reflect the effects of factors associated with them, and some of these factors may be intervening variables. For example, studies of ethnic health disparities often attempt to control for differences in socioeconomic status. However, it has been argued that socioeconomic status is an intervening variable between ethnicity and health outcomes, so that its control by the usual methods is problematic (Kaufman and Kaufman, 2001). The problem of interpretation is compounded when the persistence of an association with ethnicity, despite control for other factors, prompts the investigator to make an unwarranted inference that a genetic factor must be at work. It is also worth noting that the crude association presents the situation as it exists. Even if the causal explanation indicates other factors as responsible, the fact of disproportionate health status remains an issue to be dealt with. Moreover, a remedy may not require dealing with the "real" cause.

"Collapsibility" versus "comparability"

Although the problem of confounding and the need to control for it have long been a part of epidemiology and other disciplines, theoretical understanding of the confounding has been developed largely since Miettinen's articles in the mid-1970's. Two opposing definitions or perspectives have been debated during that time, one called "comparability" and the other called "collapsibility".

In the **comparability** definition, advocated by Sander Greenland, James Robins, Hal Morgenstern, and Charles Poole, among others (see bibliography for article "Identifiability, exchangeability, and epidemiological confounding" and correspondence "RE: Confounding confounding"), confounding

is defined in relation to the counterfactual model for causal inference, described in the beginning of this chapter. Confounding results from noncomparability, i.e., a difference between the distribution of outcomes for the unexposed group to what would have been observed in the exposed group if it had not been exposed. Since the latter value is hypothetical and unobservable, the comparability definition cannot be directly applied, though it has some theoretical advantages as well as practical implications.

In the **collapsibility** definition, advocated by D.A. Grayson (*Am J Epidemiol* 1987;126:546-53) and others, confounding is present when the crude measure of association differs from the value of that measure when extraneous variables are controlled by stratification, adjustment, or mathematical modeling. If 2 x 2 tables for different strata of a risk factor (as in the Type A example above) produce measures of association (e.g., RR's) that are essentially equivalent to the measure of association for the "collapsed" 2 x 2 table (disregarding the risk factor used for stratification), then there is no confounding in regard to that measure of association. The collapsibility definition is readily applied in practice and is widely used. Disadvantages for this definition are that it makes confounding specific to the measure of association used and the particular variables that are being controlled.

Fortunately for practicing epidemiologists, the two definitions generally agree on the presence or absence of confounding when the measure of effect is a ratio or difference of incidences (proportions or rates). The major practical problem arises when the measure of association is the odds ratio (unless the situation is one where odds ratio closely estimates a risk or rate ratio, e.g., a rare outcome). Further explanation of this and related issues are presented in the Appendix (and in Rothman and Greenland).

Controlling confounding

Now that we are all impressed with the importance and value of taking into account the effects of multiple variables, what are some of the analytic approaches available to us? The principal ones are the following:

- Restriction
- Matching
- Stratified analysis
- Randomization
- Modeling

Restriction

When we adopt restriction as an approach, we are in effect opting not to attempt a multivariable analysis – we simply restrict or confine our study to participants with particular characteristics (e.g., male, age 40-50, nonsmokers of average weight for height, with no known diseases or elevated blood pressure) so that we will not have to be concerned about the effects of different values of those variables. Restriction of some sort is nearly always a part of study design, since virtually all

studies deal with a delimited geographical area, specific age range, and so on, though the motive may be feasibility rather than avoidance of confounding. If it is known or suspected that an association is strongest in a particular population subset, then it may make sense to focus studies in that group. Or, if there are few data available that apply to a particular population, it may make sense to restrict study participants to persons in that population. Restriction is also useful when an important variable (particularly a strong suspected risk factor) is very unevenly distributed in the target population, so that it will be difficult and expensive to obtain enough participants at the less common levels of that variable.

Considerations such as these have often been cited as the reason why so many studies in the United States have enrolled only white participants, often only white males. For example, in many potential study populations (defined by geography, employment, or membership), there are (or were) too few members of minority groups to provide sufficient data for reliable estimates. Reasoning that in such a situation stratified analysis by race/ethnicity is essentially be equivalent to restriction to whites only, investigators often simply limited data analysis or data collection to whites. The reasoning behind limiting studies to males has been that the very different disease rates in men and women, so that studies of, for example, CHD in middle-aged persons, require many, many more women than men, in order to obtain a given number of cases (and therefore a given amount of statistical power or precision). The fact that until about the 1980's the number of women epidemiologists and their representation in policymaking were fairly small may also have had some influence.

The reasons for restricting study participants according to race/ethnicity are more complex. If race/ethnicity (the term "race" virtually defies precise definition) is not a risk factor for the outcome under study, then there is no need to stratify or restrict by race/ethnicity in order to control for confounding. But the United States' ever-present racial divide and its accompanying pervasive discrimination, widespread exploitation, frequent injustices, recurrent atrocities, and continuing neglect by the dominant society have created intellectual, attitudinal, political, and logistical barriers to race-neutral research (see bibliography; the states of the American south maintained legally-enforced apartheid well into the 1960's, and extra-legally enforced apartheid continues to this day). Many of these issues have also arisen for other United States populations with ancestry from continents other than Europe.

The concept of race as a powerful biological variable capable of confounding many exposure-disease associations has its historical roots in 19th century "race science", where various anatomical, physiological, and behavioral characteristics, assumed to be genetically-based, were interpreted as demonstrating the relative superiority/inferiority of population groups and justifying the subordination by whites of colored peoples (Bhopal, Raj. Manuscript in preparation, 1996; see bibliography for additional references). Various diseases and conditions were linked to racial groups (including "drapetomania", the irrational and pathological desire of slaves to run away, and "dysaethesia Aethiopica" ["rascality"]). One reads in medical books from the period that blacks are "an exotic breed".

Most of these ideas are now widely discredited, though by no means extinct (see Carles Muntaner, F. Javier Nieto, Patricia O'Campo "The Bell Curve: on race, social class, and epidemiologic research". *Am J Epidemiol*, September 15, 1996;144(6):531-536). But until recently the vast majority of epidemiologic study populations have been white, English-speaking, urban or suburban, and not

poor. A scientific basis linking race itself to health outcomes has emerged for only a handful of conditions (most prominently skin cancer and sickle cell trait and disease). But the suspicion that race could be a risk factor is difficult to dispel, in part because it is reinforced by the many race-related differences in health outcomes. These differences presumably arise from differences in diet and nutrition, physical and social environment, early life experiences, economic resources, health care, neighborhood characteristics, social interactions, experiences of discrimination, lifestyle behaviors, and the host of other factors that affect health and wellbeing, but race is a much more easily measured (if not defined) surrogate for risk.

In addition, many of these differences present logistical challenges (e.g., unfamiliarity of [primarily white, middle class] researchers and staff in studying persons and communities from other backgrounds, distances from research institutions, limited infrastructure, scarcity of questionnaire and other measurement tools that have been validated on multiple racial/ethnic groups, among others). The practical aspects of epidemiologic studies typically demand a great deal of time, effort, and cost, so it is natural to seek ways to reduce these.

Whatever the motivations and their merits, the overall impact of focusing research on white, English-speaking, urban/suburban, nonpoor populations is a scarcity of knowledge, research expertise, data collection tools, and ancillary benefits of participation in research (e.g., access to state-of-the-art treatments, linkages between health care providers and university research centers) for other populations – even for conditions for which these populations have higher rates or for which there are scientific or public health reasons for questioning the applicability of findings for Americans of European extraction to people of color or Latino ethnicity.

Since about the mid-1980's, partly in response to prodding from Congressionally-inspired policies of the National Institutes of Health and Centers for Disease Control and Prevention that now require all grant applicants to provide strong justification for not including significant numbers of women and minorities in proposed studies, research on understudied populations has increased substantially. These policies and the measures taken to enforce them have created new challenges for epidemiologists and in many cases have increased the complexity of epidemiologic studies. However, epidemiology cannot on the one hand claim that it is an essential strategy for improving public health and on the other hand largely ignore one-fourth (minorities) or five-eighths (women plus minority men) of the population.

Several years ago the American College of Epidemiology issued a "Statement of Principles on Epidemiology and Minority Populations" (*Annals of Epidemiology*, November 1995;5:505-508; commentary 503-504; also under "policy statements" on the College's web site, www.acepidemiology.org) recognizing the importance of minority health for public health, of improving epidemiologic data on minority populations, and of increasing ethnic diversity in the epidemiology profession. The Statement has been endorsed by the governing bodies of various epidemiology and public health organizations, including the Council on Epidemiology and Prevention of the American Heart Association, North American Association of Central Cancer Registries, Association of Teachers of Preventive Medicine, American College of Preventive Medicine, American Statistical Association Section on Epidemiology in Statistics, American Public Health Association, and the epidemiology faculties at numerous institutions (e.g., Harvard, UNC,

University of Massachusetts at Amherst, and University of Texas Health Sciences Center). In January 2000, the U.S. Department of Health and Human Services announced the goal of eliminating racial/ethnic disparities in health by the year 2010. This challenge and the related one of bringing racial/ethnic diversity to the epidemiology profession are fundamental to public health in the United States at least.

Matching

As discussed earlier, confounding by a risk factor(s) can be avoided in a follow-up study by ensuring, through matching, that the various exposure groups have the same (joint) distributions for those risk factor(s). Thus in a cohort study or an intervention trial, we can select participants at one exposure level and then select participants for another exposure level (including "unexposed") from a larger candidate population according to the distribution of selected risk factors in the first group.

For example, consider a retrospective cohort study to investigate whether players in collegiate revenue sports (e.g., football, basketball), when they reach age 60, are more likely to have altered evoked potentials in response to auditory stimuli, suggestive of differences in neurologic function. The exposed cohort might consist of former basketball players from the team rosters of several universities, the comparison (unexposed) cohort of former students from the same universities during the same years.

Since measurement of evoked potentials is a lengthy and expensive process, we want each participant to be as informative as possible, in order to minimize the total number of participants needed for the study. If we choose unexposed participants completely at random, it is likely that they will differ from the basketball players in a number of ways (measured during their college years) that might affect evoked potentials – height, physical health, strength, agility, coordination, age (for example, the basketball players are unlikely to be mature students returning to complete a degree after taking time off to support a family), parental education, SAT (Scholastic Aptitude Test) scores (because athletes may be recruited for their talent even if their academic records are less competitive), and sex (revenue sports are, or at least were, all male). Some of these characteristics may affect evoked potentials. Thus, comparisons of evoked potentials at age 60 between the basketball players and the other alumni could be confounded by different distributions of these and other variables.

When we attempt to control for these differences, we may find that they are so large that there are basketball players (e.g., those taller than 6-feet) for whom there are very few or no comparison subjects and comparison subjects (e.g., those shorter than 5-feet, 8-inches) for whom there are very few if any basketball players. But strata with few basketball players or with few comparison subjects provide less information for comparing evoked potentials than do strata where the two groups are present in approximately equal numbers. The findings from the analysis that controls for confounding will therefore be less "efficient", in terms of information per subject, than if basketball players and comparison subjects had similar distributions of the risk factors being controlled. With the same total number of subjects and the same risk ratio, the study with more similar comparison groups will yield a narrower confidence interval (greater statistical precision) as well as a smaller p-value (greater statistical significance).

One way to obtain a better balance in risk factors between the basketball players and the comparison group is to match the comparison group to the basketball player group on the most important risk factors. For example, we could stratify the basketball players by height and GPA (grade point average) during college. A two-way stratification might have a total of 16 strata. We could then select comparison subjects so as to have the same distribution across these 16 strata. Choosing the comparison group in this way is called frequency or category matching. (This study might also be a logical place to use restriction, e.g., to include only males, aged 18-22 years, without any medical or physical impairments.)

The above method of frequency matching requires knowing the risk-factor distribution of the index group before enrollment of comparison subjects, so that the latter can be chosen to have the same distribution. Another method of accomplishing frequency matching is paired sampling. With paired sampling, a comparison subject is chosen sequentially or at random from among potential comparison subjects having the same covariable values as each index subject. For example, every time a former basketball player enrolls in the study, we find a comparison subject belonging to the same height-GPA stratum as the basketball player. Whenever we stop enrolling subjects, the two groups will have identical distributions across the 16 strata.

Similar to paired sampling is pair matching. In pair matching, we choose each comparison subject according to characteristics of an index subject, characteristics that not widely shared with other index subjects (i.e., strata are very small, possibly containing only one index subject each). For example, we might decide to use as comparison subjects the brothers of the index subjects. Or, we might decide that we wanted the joint height-GPA distribution to be so similar between player and comparison groups that we did not want to have to categorize the variables. In this case we would choose each comparison subject to have his height within a certain range (e.g., 2 centimeters) of the index subject's height and GPA within a certain small range of the index subject's GPA (pair matching in this way is called "caliper matching", though it has been criticized (see Rothman and Greenland).

What differentiates pair matching from paired sampling and other forms of frequency matching is the tightness of the link between index and comparison subjects. If there are multiple index-comparison subject pairs in each stratum, so that the pairs could be dissolved, shuffled, and reformed, with no effect as long as all subjects stayed in their strata, then the situation is one of frequency matching. If, in contrast, comparison subjects are for the most part not interchangeable with other comparison subjects, if each comparison subject is regarded as fully comparable only to the index subject with whom he is paired, then the situation is one of pair matching. (For a discussion of paired sampling versus pair matching, see MacMahon and Pugh, 1970, pp. 252-256. Also, although the present discussion has focused on pairs, all of these concepts apply to triplets, quadruplets, and "n-tuplets", as well as to variable numbers of comparison subjects for each index subject, e.g., the index subject's siblings.)

In case-control studies, as we saw earlier, the study architecture prevents us from ensuring that exposure groups are similar with respect to other risk factors even in the study population, and certainly not in the study base. Therefore, matching in a case-control studies does not prevent confounding. Matching can be beneficial, though, since if important potential confounders are

similarly distributed in cases and controls, the comparison of these two groups can be more statistically efficient – with the same number of participants, the confidence interval for the odds ratio estimate will be narrower (i.e., the estimate will be more precise).

Unfortunately, the issue of whether or not it is beneficial to match controls to cases turns out not to have a simple answer, since in some cases matching can lead to reduced statistical efficiency. If the matching variable(s) are strongly associated with the exposure, then the exposure prevalence in matched controls will be more similar to that in cases than would occur for an unmatched control group, thereby diminishing the observed strength of association between exposure and disease. If the matching factors are not strong risk factors for the disease, then "overmatching" has occurred and a true association may be completely obscured.

The current advice for case-control studies is to match only on strong determinants of the outcome under study, especially if they are likely to be very differently distributed in cases and controls. Also, of course, do not match on a variable whose relationship to the outcome is of interest. Once you have matched cases to controls on a variable, its odds ratio will be one. Although matching in a follow-up study does not incur the problems that can arise in case-control studies, in any study design the use of matching can present practical and logistical difficulties, particularly if the pool of potential comparison subjects is small or if identifying or evaluating potential matches is costly.

Randomization

Randomization, the random assignment of participants to "exposed" or "treatment" and comparison groups, is available only in intervention trials. Randomization will ensure that, on the average, index and comparison groups will have similar proportions and distributions of all factors. Of course, in any particular study the groups may (and often will) differ in one respect or another (i.e., the randomization will not "work", though in a more precise sense, it does work – it just does not accomplish all that we would like it to). So often intervention and control groups will be constrained to be similar (through matching, also called "pre-stratification") or will be analyzed using stratified analysis.

An important consideration regarding randomization – and its decisive advantage over any of the other methods available – is that on the average randomization controls for the effects of variables that cannot be measured or are not even suspected of being risk factors. Unless a variable has been identified as relevant and can be measured, none of the other approaches described above (or below) can be used. With randomization, we have the assurance that at least on the average we have accommodated the influence of unknown and unsuspected risk factors.

Stratified analysis

Stratified analysis involves the breaking down of an aggregate into component parts so that we can observe each subcomponent individually. If smoking is a relevant factor for the disease under study, we simply say, "very well, we will look at the smokers and then we will look at the nonsmokers." Most of the examples of confounding and effect modification we have examined have been presented in terms of stratified analysis.

Stratified analysis is intuitively meaningful and widely used. It is particularly suited to the control of nominal variables (variables whose values have no ordered relation to one another, such as, geographical region [north, east, west]) and ordinal variables that have few categories (e.g., injury severity [minor, moderate, severe]). Stratified analysis gives a "picture" of what is going on in the data, is easily presented and explained, and requires no restrictive assumptions about a statistical model.

On the other hand, stratified analysis requires that continuous variables be categorized, which introduces a degree of arbitrariness and causes the loss of some information. It is not possible to control for more than a few variables at the same time because as the number of strata grows large, understanding and interpreting the results may present a major challenge, especially if the results vary from one stratum to another without any obvious pattern. Despite these drawbacks, stratified analysis is a mainstay of epidemiologic analysis approaches.

When there are multiple strata, it may be difficult to describe and to summarize the results, particularly since many strata will contain relatively few participants, so differences might readily be due to random variation. In such a case, various summary measures – generally different forms of weighted averages of the stratum-specific measures – are available. A summary measure is a single overall measure of association over all strata (or over a subgroup of the strata), controlling for the variables on which stratification has taken place. The standardized risk ratio (SRR) presented in the section on age standardization is one such summary measure. Others will be presented in the chapter "Data analysis and interpretation". Of course, as with any summary measure, if there are important differences across strata an overall average may not be meaningful.

Modeling

Given an unlimited number of participants, and an unlimited amount of time, patience, and capacity to interpret data, we could approach any multivariable analysis problem by means of stratification. But consider the dimensions of the challenge: if we have three variables, each dichotomous, there are eight possible unique strata; if we have six variables, each dichotomous, there are 64; if we have six dichotomous variables and three variables having three levels each, the number of strata soars to 1728! Imagine trying to interpret 1728 odds ratios, even assuming that we have enough participants for each one.

Since we often have more than a few variables we wish to accommodate, and variables (e.g., age, blood pressure, body weight) are often continuous so that we stand to lose information by categorizing them into any small number of levels, there is an obvious need for some more sophisticated approach that does not require us to examine every possible combination of factor levels in order to uncover the effects of each variable. There is such an approach – mathematical modeling – but its use involves a price, in terms of certain assumptions we make in the interests of simplifying the situation. Another price we pay is that the data themselves are hidden from view. In the words of Sir Richard Doll (interview with Beverly M. Calkins printed in the American College of Epidemiology Newsletter for Fall 1992):

"There have been many important steps along the way: larger scale studies, more powerful statistical techniques, and the development of computers that allow these techniques to be applied. I fear, however, that the ease of applying statistical packages is sometimes blinding people to what is really going on. You don't have a real close understanding of what the relationships are when you put environmental and all of the other components of the history together in a logistic regression that allows for fifteen different things. I am a great believer in simple stratification. You know what you are doing, and you really want to look at the intermediate steps and not have all of the data in the computer."

Limitations in the ability to control potential confounders

Typically, epidemiologists do not know all of the determinants of the health conditions they study. Other determinants may be known but cannot be measured, either in general or in the circumstances under study. Unknown and unmeasured potential confounders can be controlled only through randomization. This unique advantage randomized designs is a primary reason for their particular strength.

Even for potential confounders that are controlled through restriction, matching, stratified analysis, or modeling, limitations or errors in the conceptualization, measurement, coding, and model specification will compromise the effectiveness of control. Such incomplete control results in "residual confounding" by the potential confounder. Residual confounding, like uncontrolled confounding, can lead to bias in any direction (positive or negative, away from the null or towards the null) in the adjusted measure of effect between the study factor and outcome. Even if measurement error in the potential confounder is nondifferential (i.e., independent of the study factor and outcome), the bias in the association of primary interest can be in any direction.

It is important to be aware of these limitations, but they are not grounds for discouragement. Notwithstanding these and other obstacles, epidemiology has provided and continues to provide valuable insights and evidence. The limitations derive primarily from the subject matter – health - related phenomena in free-living human populations – rather than from the discipline. Remaining aware of limitations, minimizing them where possible, and insightfully assessing their potential impact in interpreting data are the mark of the well-trained epidemiologist.

Confounding and effect modification

As noted in the chapter on Causal Inference, epidemiology's single variable focus, the one-factor-at-a-time approach that underlies the evolution of epidemiologic understanding, is the basis for the concepts of "confounding" and "effect modification". There are also some similarities in the way that they are investigated in data analysis. To make the distinction clear, we will contrast these two different implications of multicausality.

If we observe an association between a disease and some new factor - but fail to adequately account for possible effects of known causes of the disease - we may erroneously attribute the association we observe to the new factor when in fact we may be seeing the effects of known factors. "Confounding" refers to a situation in which an observed excess of disease can be mistakenly

attributed to the exposure of interest when, in fact, some other factor – related to both the outcome and the exposure – is responsible for the observed excess. For example, the crude death rate in Florida is higher than in Alaska. If we attribute the higher death rate in Florida to the effect of citrus fruit industry, then we have fallen afoul of confounding. For the underlying "true" reason for the higher Florida death rates is the older age distribution of the Florida population.

When considering confounding, we are asking the question "Is the observed association between oral contraceptive use and myocardial infarction risk due to an effect of oral contraceptives or is the association actually due to the effects of other MI risk factors, such as cigarette smoking, elevated blood pressure, elevated blood cholesterol, and diabetes, that happen to be associated with oral contraceptive use?" To answer that question, we will attempt to ascertain that the groups being compared are the same with regard to these "potential confounders" and/or we will examine the OC-MI relationship within categories of the "potential confounders" in an attempt to "hold other factors constant".

"Effect modification" refers to variation in the relationship between exposure and outcome, variation that is due to the actions of some other factor (called an effect modifier). For example, the relationship between exogenous estrogens and endometrial cancer appears to be weaker in the presence of obesity. The relationship between oral contraceptives and myocardial infarction appears to be stronger in women who smoke cigarettes than in those who do not.

When considering effect modification, we are asking the question "Is the observed association between oral contraceptive use and MI risk importantly influenced by other MI risk factors, such as cigarette smoking, elevated blood pressure, elevated cholesterol, or even by factors which, by themselves, do not affect MI risk?" To answer that question, we will examine the OC-MI relationship within categories of these "potential modifiers". We will also seek biological and/or behavioral explanations for possible modifying influences.

With confounding, we are concerned with determining whether a relationship between our exposure and our outcome does or does not exist. With effect modification, we are concerned with defining the specifics of the association between the exposure and the outcome. That is, we are interested in identifying and describing the effects of factors that modify the exposure-outcome association. The question about confounding is central in establishing risk factors. The question about effect modification has important implications for defining disease etiology and for intervention. Confounding is a nuisance. Effect modification, though for statistical reasons it may be difficult to assess, is of considerable potential interest.

A mnemonic aid that may be helpful is the following. An evaluation of confounding is an investigation into "guilt" or "innocence". An evaluation of effect modification is an investigation into "conspiracy".

MAIN POINTS

- Confounding is a distortion or misattribution of effect to a particular study factor. It results from noncomparability of a comparison group.
- A confounder is a determinant of the outcome or its detection, or possibly a correlate of a determinant, that is unequally distributed between groups being compared.
- A determinant of the disease should appear as an independent risk factor, i.e., not one whose association with disease results from its association with the study factor.
- A potential confounder (i.e., a disease determinant) need not be an actual confounder – an actual confounder must be associated with the study factor.
- Confounding can be controlled in the study design and/or analysis.
- Control through the study design is accomplished through restriction, matching (prestratification), or randomization.
- Control in the analysis is accomplished through stratified analysis and/or mathematical modeling.
- Adequacy of control is compromised by errors in the conceptualization, measurement, coding, and model specification for potential confounders.
- Confounding deals with "guilt" or "innocence"; effect modification deals with "conspiracy".
- Discovery that an association arises from confounding does not make it less "real", but does change its interpretation.
- The crude association is real and for some purposes is the relevant measure.

Bibliography

Rothman and Greenland (see index); Rothman, *Modern epidemiology*, pp. 177-181 and 226-229.

W. Dana Flanders and Muin J. Khoury. Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. *Epidemiology* 1990; 1:239-246.

Greenland, Sander; Hal Morgenstern, Charles Poole, James M. Robins. Re: "Confounding confounding". Letter and reply by D.A. Grayson. *Am J Epidemiol* 1989; 129:1086-1091

Savitz, David A.; Anna E. Baron. Estimating and correcting for confounder misclassification. *Am J Epidemiol* 1989; 129:1062-1071.

Mickey, Ruth M; Sander Greenland. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989; 129:125-37.

Stellman, Steven D. Confounding. *Preventive Medicine* 1987; 16:165-182 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

Greenland, Sander and James M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 1986; 15:412-418. (advanced)

Schlesselman, J.J.: Assessing effects of confounding variables. *Am J Epidemiol* 108(1):3-8, 1979.

Schlesselman, J.J. *Case-control studies*. Pp. 58-63.

Kleinbaum, Kupper and Morgenstern. *Epidemiologic research: principles and quantitative methods*. Chapter 13, Confounding.

Boivin, Jean-Francois; Sholom Wacholder. Conditions for confounding of the risk ratio and of the odds ratio. *Am J Epidemiol* 1985; 121:152-158.

Greenland, Sander; James M. Robins. Confounding and misclassification. *Am J Epidemiol* 1985; 122:495-506.

Kupper, Larry L. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol* 1984; 120:643-8.

Greenland, Sander. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; 112:564-9.

Greenland, S. and Neutra R.: Control of confounding in the assessment of medical technology. *International J Epidemiology* 9:361-367, 1980.

Race/ethnicity

Amott T, Matthaei J. *Race, gender, and work: a multicultural economic history of women in the United States.* Boston, South End Press, 1991.

Crow JJ, Escott PD, Hatley FJ. *A history of African Americans in North Carolina*. Raleigh, Division of Archives and History, N.C. Department of Cultural Resources, 1992.

Franklin, John Hope. *From slavery to freedom: a history of African Americans*. 7th ed. NY, McGraw-Hill, 1994.

Freeman HP. The meaning of race in science – considerations for cancer research. *Cancer* 1998;82:219-225.

Gamble, Vanessa N. Under the shadow of Tuskegee: African Americans and health care. *Am J Public Health* 1997;87:1773-.

Greenland, Sander; Judea Pearl, James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37-48.

Hacker A. *Two nations: black and white, separate, hostile, unequal*. NY: Macmillan; 1992.

Kaufman, Jay S.; Sol Kaufman. Assessment of structured socioeconomic effects on health. *Epidemiology* 2001;12:157-167.

Krieger ND, Rowley DL, Herman A. Racism, sexism and social class: implications for studies of health, disease, and wellbeing. *Am J Prev Med* 1993;9(6[Suppl]):82-122.

Morbidity/mortality gap: is it race or racism? American College of Epidemiology Tenth Annual Scientific Meeting, *Ann Epidemiol* 1993;3:119-206.

Moss, Nancy. What are the underlying sources of racial differences in health? Editorial. *Ann Epidemiol* 1997;7:320.

Osborne NG, Feit MD. The use of race in medical research. *JAMA* 1992;267:275-279.

Stepan N. *The idea of race in science*. London, Macmillan, 1982.

Senior P, Bhopal RS. Ethnicity as a variable in epidemiologic research. *British Med J* 1994;309:327-330.

Varma J. Eugenics and immigration restriction: lessons for tomorrow. *JAMA* 1996;275:734.

Warnecke, Richard B., et al. Improving question wording in surveys of culturally diverse populations. *Ann Epidemiol* 1997;7:334-.

Williams, David R. Race and health: basic questions, emerging directions. *Ann Epidemiol* 1997;7:322-333.

Appendix

The following discussion is for the more advanced student (either now or when you are taking a more advanced methodology course) – others can skip this section.

Confounding: "Comparability" versus "collapsibility"

As presented earlier, the comparability definition labels as confounding a situation where the value of an outcome for the unexposed group differs from the (contrafactual) value for that outcome in the exposed group if it could be observed without the exposure. The collapsibility definition sees confounding as a situation where the crude measure of association differs from the value of that measure when extraneous variables are controlled (by stratification, adjustment, or mathematical modeling). The two definitions yield the same judgment in many situations, a major exception being those where the measure of association is an odds ratio which does not estimate a risk ratio (rare disease assumption not met) or a rate ratio (assumptions for estimating the IDR not met).

The reason the odds ratio is different from the rate and risk ratios in this respect is related to the fact that unlike proportions and rates, the odds for a group are not a simple average of individual members' odds (Greenland S. *Am J Epidemiol* 1987;125:761). Stratified analysis simply places the individual members of a group into a handful of strata. Since incidence for a group does equal the simple average of the risks (or "hazards") for the individual members, the overall incidence will also equal the average of the stratum-specific risks or rates, now weighted by the stratum size (or number exposed, or number unexposed) as a proportion of the total (i.e., the distribution of participants across the strata).

For risk or rate, therefore, the comparison (by means of a ratio or difference) of overall incidence in the exposed to overall incidence in the unexposed is a comparison of weighted averages. If the weights in the exposed and unexposed groups are the same, then the comparison is valid (i.e., no confounding). In this case, the overall incidence ratio (difference) is a weighted average of the incidence ratios (differences) across the strata, a condition for nonconfounding proposed by Boivin and Wacholder (*Am J Epidemiol* 1985;121:152-8) and implies collapsibility. Since the weights are the distributions of exposed and unexposed participants across the strata, equal weights mean identical distributions, which in turn means that exposure is unrelated to the risk factor used to create the strata.

If the distributions of exposed and unexposed participants across strata differ (i.e., the exposure is related to the stratification variable), then the overall incidence in exposed and in unexposed participants are averages based on different weights, so their ratio and difference will not be equal to a weighted average of the stratum-specific incidence ratios and differences. Comparability and collapsibility are therefore not present, and the comparison between overall incidences is confounded by the stratification factor. However, since the odds for the group is not a simple average of the odds for individual members, none of the above holds for the odds ratio unless it is sufficiently rare that it approximates the risk ratio or has been obtained from a design that causes the odds ratio to estimate the incidence density ratio.

Some of the relationships just presented can be readily demonstrated using simple algebra. Let a_i, b_i, c_i, d_i, n_{1i}, and n_{0i} in each stratum take on the values implied by the table below, and let their respective totals across all strata by a, b, c, d, n1, and n0 (i.e., a = all exposed cases, b = all unexposed cases, c = all exposed noncases, d = all unexposed noncases, n_1 = all exposed persons, n_0 = all unexposed persons).

	Exposure			
Disease	Yes	No	Total	
Yes	a_i	b_i	m_1	$(a_i + b_i)$
No	c_i	d_i	m_2	$(c_i + d_i)$
Total	n_{1i}	n_{0i}	n_i	
	$(a_i + c_i)$	$(b_i + d_i)$		

The incidence in exposed persons is a_i/n_{1i} within each stratum and a/n_1 when the strata are ignored (i.e., the total, or crude table). The (weighted) average incidence in the exposed across the strata is:

$$\Sigma \left(\frac{n_{1i}}{n} \times \frac{a_i}{n_{1i}} \right) = \Sigma \left(\frac{a_i}{n} \right) = \frac{a}{n}$$

where the summation goes over all strata. a/n is simply the crude incidence in the exposed. Similarly, the weighted average of the stratum-specific risk ratios can be expressed as the sum across all strata of:

$$\frac{w_i}{W} \times \frac{a_i / n_{1i}}{b_i / n_{0i}} = \frac{w_i}{W} \times \frac{a_i\, n_{0i}}{b_i\, n_{1i}}$$

where w_i are the weights for each stratum and W is the sum of the w_i. If we let $w_i = b_i n_{1i} / n_{0i}$, then we have the sum across strata of:

$$\frac{b_i\, n_{1i} / n_{0i}}{W} \times \frac{a_i\, n_{0i}}{b_i\, n_{1i}} = \frac{a_i\, w_i}{W}$$

Meanwhile, W is the sum across all strata of:

$$w_i = \frac{b_i n_{1i}}{n_{0i}}$$

If exposure is unrelated to the stratification variables, so that the distribution of exposed $n_{1i}/(a + c)$ is the same across strata as the distribution of the unexposed n_{0i}/n_0, then the ratio of exposed to unexposed in all strata must be the same as in the overall table, n_1/n_0. Therefore

$$w_i = \frac{b_i n_{1i}}{n_{0i}}, \quad \text{whose sum is simply} \quad W = \frac{b\, n_{1i}}{n_{0i}}$$

$$\text{Thus, the sum of } \frac{a_i}{w_i} \text{ is } \frac{a}{bn_1 / n_0}, \text{ which equals}$$

$$\frac{a / n_1}{b\, n_0}, \text{ the overall risk ratio.}$$

So, when there is no confounding, the following three summary measures are all equal:

$$\text{Overall risk (or rate) ratio} = \frac{\text{Overall incidence in exposed}}{\text{Overall incidence in } \underline{\text{un}}\text{exposed}}$$

$$= \frac{\text{Weighted average of incidence in exposed, across strata}}{\text{Weighted average of incidence in } \underline{\text{un}}\text{exposed, across strata}}$$

$$= \text{Weighted average of stratum-specific risk (or rate) ratios}$$

With incidence odds and odds ratios, however, the above does not apply. The overall incidence odds are simply a/c. In contrast, the average of the stratum-specific odds, weighted by the number of exposed, is the sum over all strata of:

$$\frac{n_{1i}}{n_1} \times \frac{a_i}{c_i}$$

It is possible to construct an incidence odds ratio that is a weighted average of the stratum-specific incidence *odds ratios*, and therefore a summary incidence odds ratio. However, this summary incidence odds ratio will not be equal to a ratio of average stratum-specific incidence *odds* for exposed and average stratum-specific incidence *odds* for unexposed.

Multicausality: Confounding - Assignment

1. Some years ago several studies were published showing an association between reserpine (a drug used to lower blood pressure) and breast cancer in women. Since obesity is associated both with breast cancer and with hypertension (elevated blood pressure), the suspicion arose that the association between reserpine and breast cancer could be secondary to the effect of obesity. Assume that a cohort study had been conducted to address this question and produced the following data:

Annual age-adjusted incidence of breast cancer per 100,000 women by body weight and reserpine status

	Reserpine use		
	Yes	No	Total
Obese	12.50	8.30	8.72
Not Obese	6.40	4.10	4.22
Total	10.47	6.14	

Answer the following questions on the basis of the above data (ignore considerations of statistical significance and precision). For each answer cite the most relevant figures from the table, allowing for the possibility that one factor affects the observed relation between the other factor and breast cancer risk.

a. Is reserpine a risk factor for breast cancer?

b. Is obesity a risk factor for breast cancer?

c. Is reserpine use associated with obesity?

d. Is the association between reserpine and breast cancer attributable to obesity?

2. A 20-year retrospective cohort study of the incidence of chronic obstructive pulmonary disease (COPD) was performed in two occupational cohorts with different levels of $S0_2$, copper smelters (high SO_2) and truck maintenance workers (low SO_2). In 1961, when the cohort was defined, 55% of the smelter workers and 55% of the truck shop workers were smokers. The relative risk for COPD due to smoking was 10.5 among the smelters and 3.0 among the truck shop workers. Pulmonary function data taken in 1980 showed that 75% of the smelter workers had low FEV_1 values (<90% predicted) and 33% of the truck shop workers had low FEV_1 values. COPD and low FEV_1 were strongly associated in each cohort. [FEV_1 is forced expiratory volume in one second.]

   a. In the above study, is smoking a likely confounder of the association between COPD and SO_2 exposure (i.e., in smelters vs. truck shop workers)? Briefly discuss (1-3 sentences).

   b. The best reason for not controlling for low FEV_1 as a potential confounder is:

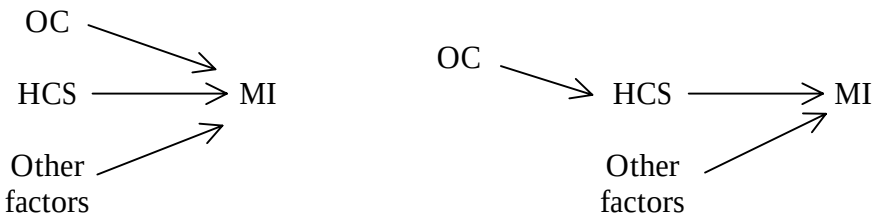
      A. Low FEV_1 is not associated with SO_2 exposure according to the data.

      B. Low FEV_1 is not associated with COPD according to the data.

      C. Low FEV_1 is not an independent risk factor for COPD.

      D. Low FEV_1 is not associated with smoking according to the data.

3. Diagrammed below are two possible causal models involving oral contraceptive use (OC), plasma homocysteine level (HCS) and myocardial infarction (MI). Briefly discuss the implications of the two models with respect to whether HCS would need to be considered as a potential confounder of the relationship between OC and MI.



[arrows show hypothesized causal pathways]

4. The following table, published in the Oxford Family Planning Association Contraceptive Study (Vessey et al.), shows characteristics of individuals at the time of recruitment in to the study. Based on the data presented in the table, discuss three potential sources of bias apparent from the characteristics of the three contraceptive groups. How would these factors be expected to influence the appearance of a causal association between oral contraceptive use and circulatory deaths if no adjustment for the factors were carried out?

Some characteristics of subjects in the three contraceptive groups at time of recruitment

Characteristic	Method of Contraception in use on Admission: Oral	Diaphragm	IUD
Percentage aged 25-29 years	56	35	35
Percentage in Social Classes I or II*	39	49	34
Percentage smoking 15 or more cig./day	17	7	12
Mean Quetelet's Index**	2.25	2.26	2.31
Percentage*** with history of:			
Hypertension	0.91	0.67	0.50
Pre-eclamptic toxaemia	12.58	16.26	16.07
Stroke	0.03	0.04	0.30
Rheumatic fever	0.76	0.66	1.04
Rheumatic heart disease	0.09	0.26	0.32
Congenital heart disease	0.12	0.31	0.16
Venous thromboembolism	0.87	4.30	7.96

* Registrar General's classification [Social Class I is highest]

** Weight (g) / height $(cm)_2$.

*** Standardized by indirect method for age and parity. See Vessey, et al.

5. The following questions (from the 1985 EPID 168 second midterm exam) are based on data from Kantor AF, Hartge P, Hoover RN, et al. Urinary tract infection and risk of bladder cancer. *Am J Epidemiol* 1984;119:510-5). In that study, 2982 newly-diagnosed bladder carcinoma patients identified through the U.S. National Cancer Institute SEER (Surveillance, Epidemiology and End Results) Program during a one-year period beginning in December 1977 were interviewed. 5782 population controls from the same geographic areas covered by SEER were selected using an age- and sex-stratified random sample of the general populations, with 2:1 frequency-matching of controls to cases. Information on physician-diagnosed urinary tract infections (UTI) more than one year before interview (and many other factors) was obtained through personal interviews using structured questionnaires in respondents' homes. The following data are from Table 1 in Kantor, Hartge, Hoover et al.:

Table 1
Relative risks (RR) of bladder cancer associated with history of urinary tract infection,* by number of infections; 10 geographic areas of the United States, 1978

	Males				Females			
No. of urinary tract infections	Cases	Controls	RR	95% confidence interval	Cases	Controls	RR	95% confidence interval
0	1758	3642	1.0+		398	979	1.0+	
1 or 2	309	423	1.5	(1.3-1.8)	176	296	1.2	(0.9-1.5)
3+	146	152	2.0	(1.6-2.6)	145	206	2.1	(1.6-2.7)

* Maximum likelihood estimate of relative risk adjusted for race, age, smoking status (never smoked, ex-smoker, current smoker) from stratified analyses.

+ Reference category

a. The sex-specific relative risks for bladder cancer (BC) shown in Table 1 are adjusted for race, age, and smoking status. Which one of the following could be the relative risk of BC risk for 3+ urinary tract infections (UTI) adjusted for race, age, smoking status, and gender? [Choose one best answer]

   A. 1.33

   B. 1.92

   C. 2.05

   D. None of the above.

b. Using the data in Table 1, construct a labeled 2 by 2 table for estimating the crude (with respect to race, age, smoking status, and gender) relative risk for BC in men and women who have a history of 3+ UTI compared to men and women with no history of UTI. Your answer should show the correct formula and substitution.

c. Is gender associated with history of 3+ UTI? Support your answer with the most relevant numbers from Table 1.

d. Is gender a confounder of the association between BC risk and history of 3+ UTI? Support your answer with data from Table 1 and/or your answers to the two preceding questions._

Multicausality: Confounding - Assignment solutions

1.

a. Reserpine is a risk factor. Overall, the incidence of breast cancer is 10.47 per 100,000 women-years in reserpine users and 6.14 per 100,000 women-years in nonusers. Moreover, among the non-obese, the rate ratio is: 6.40/4.10 = 1.6. Looking in the non-obese women avoids potential confounding by obesity if it is a risk factor.

b. Obesity is also a risk factor. The overall incidence rates for obese and non-obese women, respectively, are 8.72/100,000 women-years and 4.22/100,000 women-years. Among the women who did not take reserpine, the rate ratio is 8.30/4.10 = 2.0. Looking in the group of reserpine non-users avoids potential confounding by reserpine.

c. Reserpine use <u>is</u> associated with obesity, though that fact cannot be deduced from the stratum-specific breast cancer rates <u>alone</u>. The direct approach is to remember that the crude rates are weighted averages of stratum-specific rates, with the weights consisting of the population prevalence of the risk factors. So:

$$12.50\ P_{RO} + 8.30\ (1-P_{RO}) = 8.72 \quad \text{and}$$

$$6.40\ P_{\overline{RO}} + 4.10(1-P_{\overline{RO}}) = 4.22$$

where P_{RO} is the prevalence of reserpine use in obese subjects,
and

$P_{\overline{RO}}$ is the prevalence of reserpine use in nonobese subjects.

Solving these two equations gives $P_{RO} = 0.1$ and $P_{\overline{RO}} = 0.05$, so reserpine use is more prevalent in obese women (presumably because they are more likely to have hypertension). The relative prevalence is 2.0; the odds ratio of association between obesity and reserpine use is [(.1)(.95)]/[(.9)(.05)] = 2.1. Such an association might be characterized as "moderate."

Note that the above procedure involving weighted averages can be equally well carried out on the basis of the column rates, rather than the row rates. The odds ratio will be effectively the same.

d. The association between reserpine and breast cancer is <u>not</u> attributable to obesity (in the data for this problem). The most relevant rates to demonstrate that the association is not

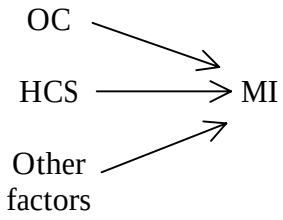
completely attributable to obesity are those comparing reserpine users and nonusers among the nonobese (6.40/4.10=1.6). Since the crude rate ratio (10.47/6.14=1.7) is ever so slightly greater than each of the stratum-specific rate ratios (1.6 in the nonobese, 1.5 in the obese) it can be argued that a slight amount of the crude association is attributable to obesity.

2.

a. Smoking is <u>not</u> likely to be a confounder, because both of the compared groups (smelters and truck stop workers) have identical proportions of smokers (55%). Smoking could therefore not account for a difference in lung disease between the two groups compared.

b. The best reason for not controlling low FEV_1 as a potential confounder is C, low FEV1 is not an independent risk factor for the development of COPD, but is rather a manifestation of COPD.

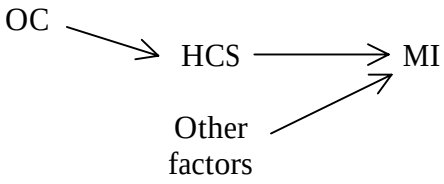
3. In the first causal model:



[arrows show hypothesized causal pathways]

HCS is a causal risk factor for MI, through a pathway independent of OC. Therefore, HCS must be controlled as a potential confounder of an OC-MI association.

In the second causal model:



[arrows show hypothesized causal pathways]

HCS is an intermediate in the causal pathway from OC to HCS. Therefore, HCS cannot logically be a confounder of the OC-MI relationship. If one controls for the effect of HCS, no residual effect will be found for OC. A more useful approach would be to investigate the

link between HCS and MI to ensure that it is causal. Then the link between OC and HCS should be explored while other influences on HCS are controlled.

4. Some possible answers:

Age: The oral contraceptive users were much younger than either the diaphragm or IUD users. In fact, there were 21% more women in the age range 25-29 using oral contraceptives than in either of the other groups -- or almost one-and-a-half times more younger women using OC's than other methods. This would bias any associations between OC use and circulatory deaths downward, since younger women are less likely to develop circulatory disease than older women.

Cigarette Smoking: The oral contraceptive users were also more likely to smoke 15 or more cigarettes a day than either diaphragm or IUD users. There were almost two-and-one half times (17% vs 7%) more 15+/day smokers among OC uses than among women using the diaphragm, and almost one-and-one-half times (17% vs 12%) more smokers among OC users than among IUD user. These differences would likely increase the risk of circulatory deaths among OC as compared to non-OC users, since cigarette smoking is significantly related to death from circulatory disease.

History of Hypertension: Although the percentages were low, the oral contraceptive users were more likely to have a history of hypertension than were other contracepting women (.91% vs .67% and .50%). This slight excess of hypertension among OC users might increase their risk of developing circulatory disease as compared to the non-users, since hypertension is one of the most significant risk factors for circulatory death.

Venous Thromboembolism: The percentage of OC users with a history of venous thromboembolism was much lower than among women using the diaphragm or the IUD. There were 5 times (4.30% vs. .87%) more women with a thromboembolism problem among diaphragm users than among OC users, and 9 times (7.96% vs. .87%) more women with a history among IUD than OC users. This difference would likely bias the risks of OC use downward, since the non-OC user group had more prevalent circulatory disease, which is more likely to lead to circulatory death.

History of Rheumatic Heart Disease: Although the percentages were small, the OC users were less likely to have had rheumatic heart disease than the non-OC users (.09% vs .26% and .32%). Since a history of rheumatic heart disease increases risk of circulatory death, the risks for OC users would be lowered by this difference.

5.

a. C -- the adjusted relative risk would be a weighted average of 2.0 and 2.1, so only 2.05 is a possible value for the adjusted relative risk.

b. Bladder cancer

	Cases	Controls	Total
3 + UTI	146 + 145 = 291	152 + 206 = 358	649
No UTI	1758 + 398 = 2,156	3642 + 979 = 4,621	6,777
Total	2,447	4,979	7,426

$$OR = \frac{ad}{bc} = \frac{(291)(4{,}621)}{(358)(2{,}156)} = 1.74$$

c. Yes, strongly. 13.9% [206/(206+296+979)] women have 3+ UTI, compared to only 3.6% [152/(152+423+3642)] men. [The question did not ask about the relative risk of bladder cancer.]

d. Yes, there is some confounding, since 1.7 is below both 2.0 and 2.1.

12. Multicausality: Effect Modification

Issues in characterizing the combined effect of two or more causes of a disease (or, equivalently, the effect of one factor in the presence or absence of other factors).

Multicausality

The rise of the germ theory of disease brought with it the paradigm of specificity of disease causation, in which diseases were specific entities and each specific disease had a specific cause. Since identifiable microorganisms could be linked to specific clinical syndromes and natural histories, this paradigm contributed to the dramatic progress in medical microbiology and development of antibiotics, which have transformed human vulnerability to infectious disease. The doctrine of specific causation proved something of a hindrance, however, in the study of noninfectious disease, notably in appreciating the health effects of tobacco smoke.

Now that the concept of multifactorial disease is fully accepted, we should perhaps adopt a more relativist perspective, in which specificity of causation varies according to the "disease" (used hereinafter to refer to any outcome of interest) and its definition, the type of causal agents or factors we wish to consider, and on the stage in the causal process. John Cassel invited this way of thinking when he described tuberculosis, a hallmark of the revolution in bacteriology, as a multifactorial disease in regard to various characteristics of the host and his/her social environment. The resurgence of mycobacterial tuberculosis in the United States during the 1980's, as a result of such factors as the spread of HIV, the rise in homelessness, and the reduction of funding for tuberculosis control, illustrates the importance of host and environmental factors for this disease.

The upsurge in syphilis in the southeastern region of the U.S. during a similar period provides another example. In the chapter on Phenomenon of Disease, syphilis served as an example of a disease defined by causal criteria and thus definitionally linked to a specific microorganism, *Treponema pallidum*. Syphilis infection can induce a great variety of symptoms and signs, so great that it has been called the "Great Imitator" (by Sir William Osler, I believe, who I think also wrote "Know syphilis and you will know all diseases"). Given the diversity of ways in which syphilis manifests, it is fortunate that we do not need to rely on manifestational criteria to define syphilis. Nevertheless, although defined in relation to its "cause", syphilis can also be considered a multifactorial disease, since risk of syphilis is related to personal and contextual factors such as number and type of sexual partners, use of condoms, exchange of sex for drugs or money, use of crack cocaine, access to care, proficiency of clinicians, effectiveness of public health services, degree of social stigma, racism, and limited resources devoted to developing a vaccine. Since syphilis is not transmitted in every unprotected exposure, there may be transmission and immune factors to add to this list.

Similarly, coronary heart disease is a classic multifactorial disease, with an ever-growing list of risk factors that includes at least atherogenic blood lipid profile, cigarette smoke, elevated blood pressure, sedentary lifestyle, diabetes mellitus, elevated plasma homocysteine, and insufficient intake

of dietary antioxidants. However, coronary artery disease is a clinically-defined entity that develops from a composite of changes in the coronary arteries. As our understanding of the pathophysiology and pathogenesis of coronary heart disease becomes more refined, researchers may eventually decide that it is more useful to subdivide this complex disease entity into its specific pathogenetic processes, which include certain types of injury to the coronary endothelium, growth of atheromas, and thrombus formation. These different pathologies could be defined as separate diseases, even though clinical manifestations usually require more than one to be present.

The one-variable-at-a-time perspective

Epidemiologists, however, typically focus on a single putative risk factor at a time and only sometimes have the opportunity to focus on specific pathogenetic processes. One reason for this is that epidemiology is in the front lines of disease control, and it is often possible to control disease with only a very partial understanding of its pathophysiology and etiology. Once it was demonstrated that cigarette smoking increased the risk of various severe diseases, including lung cancer, coronary heart disease, and obstructive lung diseases, many cases could be prevented by reducing the prevalence of smoking even though the pathophysiologic mechanisms were largely unknown. Once it was found that AIDS was in all probability caused by an infectious agent and that unprotected anal intercourse greatly facilitated its transmission, effective preventive measures could be taken even before the virus itself was identified and the details of its pathogenicity unravelled.

Thus, epidemiologists often find ourselves taking a "one-variable-at-a-time" approach to diseases of unknown and/or multifactorial etiology. Lacking the knowledge needed to work from a comprehensive model of the pathophysiologic process, epidemiologists attempt to isolate the effects of a single putative risk factor from the known, suspected, or potential effects of other factors. Thus, in the preceding chapter we examined how the effects of one factor can be misattributed to another factor ("guilt by association") and considered ways to control for or "hold constant" the effects of other risk factors so that we might attribute an observed effect to the exposure variable under investigation.

Another consequence of the one-variable-at-a-time approach is the phenomenon that an association we observe may vary according to the presence of other factors. From our ready acceptance of multicausation, we have little difficulty entertaining the idea that some disease processes involve the simultaneous or sequential action of more than one factor or the absence of a preventive factor. Indeed, with the growth of genetic knowledge all disease is coming to be regarded as a product of the interaction of genetic and environmental (i.e., nongenetic) factors.

But from the one-variable-at-a-time perspective, our window into these interdependencies comes largely from measures of association and impact for each particular risk factor-disease relationship. Thus, if two factors often act in concert to cause disease, we will observe the risk difference for one of the factors to differ depending upon the level of the other factor. It may therefore be important to control for factors that may modify a measure of effect of the exposure of primary interest. Control may be necessary even if the susceptibility factor cannot itself cause the disease and so would not qualify as a potential confounder.

Interdependent effects

The preceding chapters have largely dealt with situations involving a single exposure and a single outcome. The chapter on standardization of rates and ratios and the chapter on confounding concerned the need to control for a variable, such as age or a second exposure, so that comparisons could focus on the exposure of primary interest. We referred to the interfering variable as a confounder or potential confounder – essentially a nuisance variable – that threatened to interfere with our investigation of the primary relationship of interest.

We now want to consider another role for a second exposure variable. That role is involvement in the pathophysiologic process or in detection of the outcome in concert with or in opposition to the study factor (an exposure of primary interest). One of the factors may be regarded as a co-factor, a susceptibility factor, a preventive factor, or something else whose effect is entwined with that of the study factor.

Confounding, as we saw in the preceding chapter, results from an association between the exposure and the confounder. But the effects of these two exposures on the disease can be independent of one another. In fact, in the (hypothetical) Type A example, the exposure had no effect at all. In this chapter we are interested in exposures whose effects on the outcome are interdependent.

There are innumerable scenarios we can think of where such interdependence occurs. One entire category of interdependence involves genetic diseases whose expression requires an environmental exposure. For example, favism is a type of anemia that is caused by consumption of fava beans in people with reduced glucose-6-phosphate dehydrogenase (GPDH) activity. The anemia develops only in response to a constituent of fava beans, but people with normal GPDH activity are unaffected.

Another category of interdependence is that between exposure to infectious agents and immune status. Measles occurs only in people who have not already had the disease and rarely in people who have received the vaccine. People whose immune systems have been weakened by malnutrition or disease are more susceptible to various infectious agents, and HIV infection can render people vulnerable to a variety of infections called "opportunistic" because they occur only in immunocompromised hosts.

Causal chains that involve behaviors provide many illustrations of interdependency in relation to outcomes. Condoms reduce STD risk only when the sexual partner is infected. Airbags provide lifesaving protection to adult-size passengers involved in frontal crashes but can harm small passengers and provide less protection to persons not wearing a lap belt. Handguns are probably more hazardous when in the possession of people with poor anger management skills.

Since very few exposures cause disease entirely by themselves (rabies virus comes close), nearly every causal factor must modify the effect of other causal factors and have its effect modified by them. When these other factors are unidentified, they are generally regarded as part of the background environment, assumed to be uniformly distributed, and hence disregarded. Part of the challenge of

epidemiologic research is to identify major modifying factors that are not uniformly distributed, so that differences in findings across studies can be understood.

The terminology thicket

Even more than other areas of epidemiology, learning about how epidemiologists approach interdependent effects is complicated by a two decades old controversy about definitional, conceptual, and statistical issues and by a terminology that is as heterogeneous as the enrollment in a large class in introductory epidemiology! The terms epidemiologists have used to discuss interdependent or "joint" effects include: "synergy", "synergism", "antagonism", "interaction", "effect modification" (and "effect modifier"), and most recently "effect measure modification".

"**Synergy**" or "**synergism**" is the term applied to a situation in which the combined effect of two (or more) factors is materially greater than what we would expect from the effect of each factor acting alone. "**Antagonism**" refers to the reverse situation, where the joint effect is materially less than what we would expect. Synergism and antagonism are both types of "interaction".

The factors involved in an interdependent relationship can be regarded as having their effects modified by each other, which gives rise to the terms "effect modification" and "effect modifier". Sometimes the adjectives "quantitative" and "qualitative" are employed to distinguish between situations where the modifying variable changes the direction of the effect of the primary exposure or changes only the magnitude of effect. In **quantitative effect modification**, the modifier may strengthen or weaken the effect of the primary exposure, but the direction of effect does not change. In **qualitative effect modification**, the exposure either (1) increases risk in the presence of the modifier but reduces risk in its absence or (2) increases risk in the absence of the modifier but reduces risk in its presence. Although I first heard this distinction in a seminar presented by Doug Thompson, he more recently has referred to qualitative effect modification as a **crossover effect** (Thompson 1991).

Somewhere I picked up (or made up) the term "absolute effect modification" to refer to situations where the effect of at least one factor occurs only in the presence (or absence) of another factor. In such cases the first factor has no independent effect. In contrast, "relative effect modification" refers to situations where both factors have independent effects on risk regardless of the presence or absence of the other, but their joint effect is different from what one expects from their individual effects.

[Since more than two factors are generally involved, that means that, for example, variable A can be an absolute modifier of the effect of variable B (B has no effect without A) and a relative modifier of the effect of variable C (C has an effect without A, but its effect is stronger [weaker] in the presence of A). Whether B and/or C are absolute or relative modifiers of depends, in turn, on whether or not A has an (independent) effect on risk without B and/or C. But we are getting ahead of ourselves here.]

All of this terminology would be simply a matter of memorization were it not for one central difficulty. That difficulty arises in operationalizing the above concepts through the use of

epidemiologic data. Put simply, there is no simple connection between the concepts expressed above and the epidemiologic measures we have been using. Partly because of this disconnect, the terms "interaction" and "effect modification" have been employed with different meanings at different times by different authors (and sometimes by the same author). Thompson (1991:p221) says that the two terms have different "shades of meaning" but (wisely) uses the two terms interchangeably.

Previous editions of this chapter attempted to reduce terminology confusion by following the usage in the first edition of Rothman's text *Modern Epidemiology*. Rothman used the term "biological interaction" to refer to synergy or antagonism at the level of biological mechanisms, such as that in the favism example. He used the term "effect modification" to refer to data that give the appearance of joint effects that are stronger or weaker than expected (statistical interaction falls into this category). The second edition of *Modern Epidemiology* introduces a new term, "**effect measure modification**", with the purpose of reducing the tendency to link data and biology through the use of the same word. Kleinbaum, Kupper, and Morgenstern used the terms "**homogeneity**" and "**heterogeneity**" to indicate similarity or difference in a measure across two or more groups. These neutral terms, which carry no connotation of causation, may be the safest to use.

Statistical interaction

The term "interaction" has an established and specific meaning in statistics, where it is used to characterize a situation where effects are not additive. (Statisticians have the significant advantage of being able to use the term "effects" without a causal connotation.) For example, **analysis of variance** is used to compare the means of one variable (e.g., blood pressure, BP) between two or more populations. If we are concerned that BP is influenced by another variable (e.g., body mass index, BMI) and that the two populations have different BMI distributions, we may want to adjust the BP comparison for BMI. (The idea is similar to our computation of a standardized rate difference to compare mortality rates in two populations.) If the relationship between BP and BMI is linear, then the method of adjustment is called **analysis of covariance** and can be illustrated as two lines on a pair of axes (see left side of figure).

The vertical distance between the lines represents the **adjusted difference** in mean BP between the two populations. Unless the two lines are parallel, however, the distance between them will vary with the level of BMI. The lines will be parallel when the slope of the relationship between BP and BMI is the same in the two populations, i.e., the strength of the association between blood pressure and BMI is the same in the two populations.

When the two lines are parallel, the blood pressures in the two populations can be represented by an equation with three terms on the right-hand side – a constant (**a**), a variable (POP) indicating the population in which the relationship is being estimated, and BMI, e.g.,

$$BP = a + b_1\,POP + b_2\,BMI$$

in which a, b_1, and b_2 will be estimated through a procedure called "**linear regression**".

Since the indicator variable (POP) is usually coded as 0 for one population and 1 for the other, the equations representing blood pressures are:

$$BP = a + 0 + b_2\,BMI \qquad (POP=0)$$

in one population and:

$$BP = a + b_1 + b_2\,BMI \qquad (POP=1)$$

in the other.

b_1 is then the vertical distance between the two lines, which corresponds to the adjusted difference in mean blood pressure between the populations. b_2 is the slope of the relationship between BP and BMI, i.e. the number of units increase in BP associated with a one-unit increase in BMI. This term accomplishes the adjustment needed for BMI. "a" is a constant that is usually needed to move the lines to their correct vertical position.

In the right side of the figure, the two lines are not parallel – there is **interaction**. Since the distance between the lines varies according to the level of BMI, the distance cannot be stated as a single number. In the presence of interaction, the linear model for blood pressure requires the addition of an "**interaction term**" to represent the varying distance between the lines:

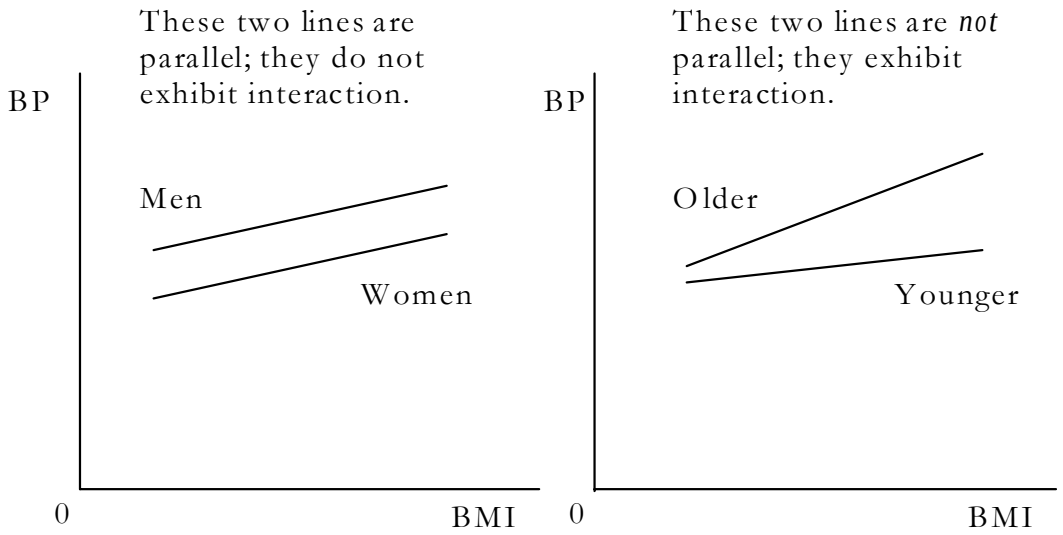
$$BP = a + b_1\,POP + b_2\,BMI + b_3\,(POP)\,(BMI)$$

With POP coded as 0 or 1, the first population will still have its blood pressures modeled by: $BP = a + b_2\,BMI$. However, the data in the second population will be modeled as:

$$BP = a + b_2 + b_2\,BMI + b_3\,BMI \qquad (POP=1)$$

b_3 represents a further adjustment to account for the lack of parallelism and thus the inability of b_1 alone to represent the distance between the lines. The difference between the two populations will be stated as $(b_1 + b_3\,BMI)$, so that it will be different for different levels of BMI.

Illustration of statistical interaction



If the figure on the left represents the relationship between blood pressure (BP) and body mass index (BMI) in men (upper line) and women (lower line), then the graph shows that the association of body mass and blood pressure is equally strong in both sexes – a one-unit increase in body mass index in men and a one-unit increase in women both are associated with the same increase in blood pressure. Therefore there is no (statistical) interaction.

In contrast, if the figure on the right represents the relationship in older people (upper line) and younger people (lower line), then the graph indicates an interaction between body mass index and age – a one-unit increase in body mass index in older people is associated with a larger increase in blood pressure than is a one-unit increase in younger people.

Statisticians use the "interaction" to refer to the latter situation, where the equations for different groups differ by a variable amount on a given scale (e.g., interaction may be present on the ordinary scale but not on the log scale).

Biological interaction

Epidemiologists are more interested in what Rothman and Greenland call "biological interaction". Biological interaction refers to interdependencies in causal pathways, such as those discussed at the beginning of this chapter. Such interdependencies – situations where one factor may potentiate or inhibit the effect of another – have implications for understanding of disease etiology or effectiveness of treatments or interventions. Laboratory researchers can readily observe such interdependencies, but epidemiologists must content ourselves with analyzing clinical or population data.

Over two decades ago (Causes, *Am J Epidemiol*, 1976), Rothman introduced a diagrammatic representation of multicausality in this biological or other mechanistic sense. He has continued to elaborate this schematic model and uses it to illustrate and explain relationships between epidemiologically-perceived relationships and "biological relationships".

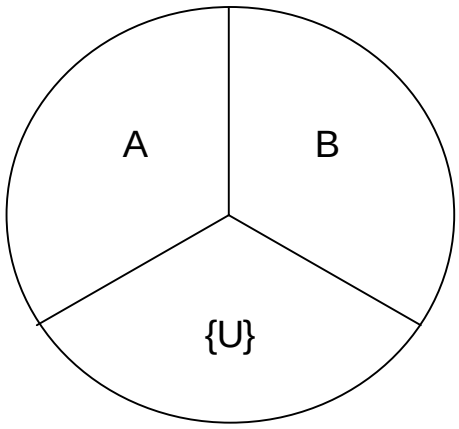
Rothman's model envisions causal pathways ("sufficient causes") as involving sets of "component causes". A "sufficient cause" is any set of component causes that simultaneously or sequentially bring about the disease outcome. "Component causes" are the individual conditions, characteristics, exposures, and other requisites (e.g., time) that activate the available causal pathways. Since there are always causal components that are unknown or not of interest for a particular discussion, sufficient causes include a component to represent them. Let us explore the way Rothman's model works.

"Cause" - (1) an event, condition or characteristic that plays an essential role in producing the occurrence of the disease (this is a "component cause"); or (2) a constellation of components that act in concert.

"Sufficient cause" - Set of "minimal" conditions and events that inevitably produce a disease; none of the conditions is superfluous; most of the components are unknown.

"Necessary cause" - A causal component that must be present for the disease to occur.

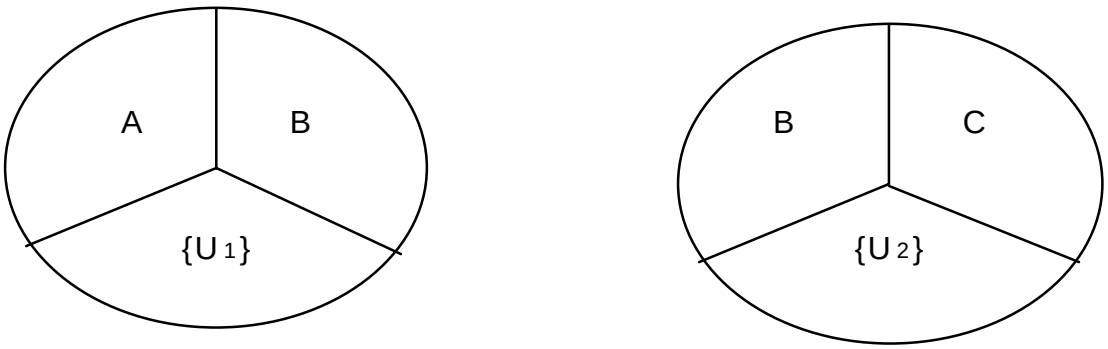
The circle below represents a sufficient cause, e.g., a pathway, chain, or mechanism that can cause a particular disease or other outcome. If all components are present, then the disease occurs (on analogy with the game Bingo). A and B represent component causes. For this sufficient cause to come into play, both A and B must be present. {U} represents the unknown background factors that also must be present for this sufficient cause to operate.



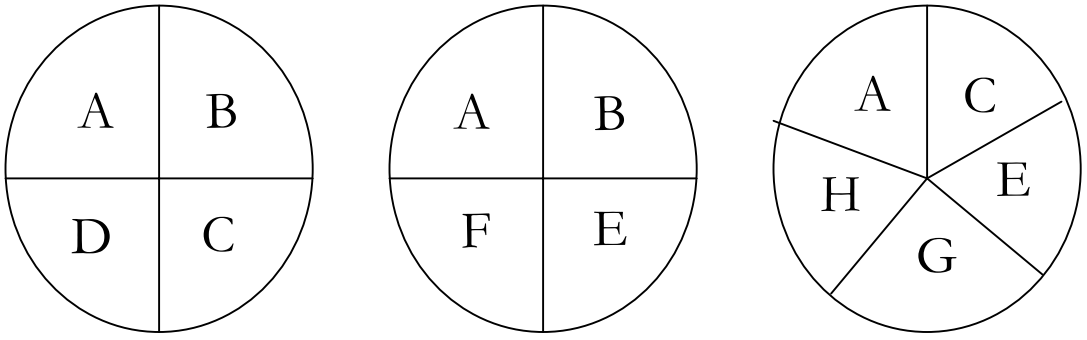
If this diagram (model of biological, chemical, physical, psychological, etc. reality) represents the primary or only pathway to the outcome, then component causes A and B have interdependent effects. Each component cause must be present for the other to have its effect. We could say that they are synergistic. The favism situation could be represented in this way, with A representing fava

bean intake and B representing genetically-determined reduced glucose-6-phosphate dehydrogenase activity. If this sufficient cause is the only causal pathway by which the disease can occur, then this synergism is absolute: without A, B has no effect; with A, B does if the remaining components {U} are present; without B, A has no effect; with B, A does (when {U} are present). (If either factor is preventive, then A or B represents its absence.)

If there were additional causal pathways containing B but not A, then the absence of A would not completely eliminate the effect of B. The latter situation, illustrated below, might be characterized as intermediate, partial, or relative synergism. B can now affect disease risk even in the absence of A.



A has thus become a relative modifier of the effect of B. B, however, remains an absolute modifier of the effect of A, because A has no effect in the absence of B. We may also note that component cause B is a necessary cause, since there is no sufficient cause (causal pathway) that can operate unless B is present.



In this diagram, G and H exhibit absolute synergy, since neither has an effect in the absence of the other. B and C exhibit partial synergy with respect to each other, since their combined effect exceeds what would be expected from knowledge of their separate effects.

Applying Rothman's model to epidemiologic concepts (induction period) and measures

In our discussion of natural history of disease, we defined the induction period as the time between initial exposure and the development of the disease. Since causal pathways involve multiple component causes, though, in Rothman's model the concept of induction period applies to component causes, rather than to the disease. The induction period in respect to a particular component cause is the time usually required for the remaining component causes to come into existence. If necessary, one component cause can be defined as a period of time (e.g., for a microbial pathogen to multiply). By definition, the induction period for the last component cause to act has length zero.

Another and even more fundamental issue is that in multicausal situations, disease occurrence, extent, association, and impact all depend upon the prevalence of the relevant component causes in the populations under study. While we have previously acknowledged that the incidence and/or prevalence of a disease or other phenomenon depends upon the characteristics of the population, we have not examined the implications of this aspect for other epidemiologic measures. For example, we have generally spoken of strength of association as though it were a characteristic of an exposure-disease relationship. But though often treated as such, strength of association is fundamentally affected by the prevalence of other required component causes, which almost always exist.

Rothman's model helps to illustrate these relationships in situations where biological interdependency (used as a general term to signify any causal interdependency) is present. A basic point is that disease incidence in persons truly unexposed to a study factor indicates the existence of at least one sufficient cause (causal pathway) that does not involve the study factor. If exposed persons have a higher prevalence of the component causes that constitute this sufficient cause, their disease rate will be higher. This process is the basis for confounding to occur.

Second, since very few exposures are powerful enough to cause disease completely on their own, the rate of disease in exposed persons will also depend upon the prevalence of the other component causes that share pathways (sufficient causes) with the exposure. Measures of association and impact will therefore also depend upon the prevalence of other component causes, since these measures are derived from incidence rates.

Third, if two causal components share a causal pathway, then the rarer of the two component causes will appear to be a stronger determinant of the outcome, especially if the remaining component causes are common. As in economics, the limiting factor in production experiences the strongest upward pressure on price.

Fourth, proportion of disease attributable to a component cause (i.e., its ARP) depends upon the prevalence of the other component causes that share the causal pathway(s) to which it contributes. This result is so because if the strength of association depends upon prevalences, then so must the ARP. However, the ARP's for the various component causes are not additive and will often sum to more than 1.0. For example, if two component causes are in the same causal pathway, then the entire risk or rate associated with that pathway can be attributed to each of the two components. The absence of either component prevents the occurrence of the outcome.

Phenylketonuria example

An example of these relationships, from the article referred to earlier (Causes, *Am J Epidemiol* 1976; 104:587-92), is the causation of phenylketonuria (PKU), a condition that, like favism, is linked to a dietary factor (phenylalanine, an amino acid) and a genetic defect. Infants with the PKU gene who ingest more than a minimal amount of phenylalanine develop serious neurologic effects including mental retardation. The "causal pie" for this example would be the same as the first one in this chapter, with A representing the PKU gene and B representing dietary phenylalanine.

Since Western diets typically contain phenylalanine, in the absence of specific preventive measures (universal screening of newborns and institution of a special diet) nearly all infants with the PKU gene develop clinical manifestations. The risk ratio for the PKU gene is therefore enormous; the PKU gene is a "strong" cause. In contrast, phenylalanine is a "weak" cause, since nearly all infants are exposed to it and only a tiny proportion develop clinical PKU. However, in a society in which a large proportion of the population have the PKU gene and infant diets rarely contain phenylalanine, then dietary phenylalanine will appear as the strong cause and the PKU gene as the weak cause! (Recall: "any measure in epidemiology is a weighted average . . .".).

Numerical example - favism

To explore these ideas further, let us construct a numerical example. Suppose that in a population of size 10,000,000, there are two sufficient causes of favism, one that involves both GPDH deficiency and fava bean intake, and a second that involves neither of these factors. Assume:

- 1% of the population (100,000 persons) have GPDH deficiency;
- 20% (2,000,000) of the population consume fava beans;
- These two factors are distributed independently of one another, so that 20,000 people have both factors (20,000 = 1% of the 2,000,000 fava bean = 20% of the 100,000 GPDH deficient persons).
- All remaining component causes {U} needed to lead to favism through the first sufficient cause are simultaneously present in 10% of persons, independent of their other risk factors;
- The sufficient cause that does not involve fava beans or GPDH deficiency occurs in 0.03% of the population, again independent of other factors/component causes. (We are assuming that the definition of favism does not require involvement of fava beans themselves.)

In this situation, the first sufficient cause will act in the expected 1% x 20% x 10% = 0.02% of the population in whom all these components are present, i.e., 2,000 cases. The second sufficient cause will operate in 3,000 persons, regardless of GPDH deficiency and/or fava beans. The table below shows what we can expect to observe in various subsets of the population.

Incidence of favism by population subgroup

Sub-population	N	Incidence	Cases
People who do not eat fava beans and do not have GPDH deficiency; [N = 80% x 99% x 10,000,000; cases come only from the 2nd pathway]	7,920,000	0.03%	2,376
People who eat fava beans but do not have GPDH deficiency [N = 20% x 99% x 10,000,000; cases come only from the 2nd pathway]	1,980,000	0.03%	594
People with GPDH deficiency who do not eat fava beans [N = 1% x 80% x 10,000,000; cases come only from the 2nd pathway]	80,000	0.03%	24
People with GPDH deficiency who eat fava beans N= 1% x 20% x 10,000,000; 10% (2,000 cases) occur in the 10% with the remaining component causes; also, 0.03% of the 20,000 (6 cases) get favism through the second pathway; (0.6 people would be expected to have both pathways acting so are subtracted from the above total)]	20,000	10.03%	2,005.4
Total	10,000,000	0.05%	4,999.6

From this table we can compute (crude) incidences and incidence ratios for each exposure:

Incidence and incidence ratios of favism (crude)

GPDH deficiency		
Present (2,030 cases / 100,000 people)	2.03%	
Absent (2,970 / 9,900,000 people)	0.03%	
Incidence ratio		67.67
Eat fava beans		
Yes (2,600 cases / 2,000,000)	0.13%	
No (2,400 / 8,000,000)	0.03%	
Incidence ratio		4.33

So indeed, the scarcer factor (GPDH deficiency) has the greater incidence ratio. If we increase the prevalence of GPDH deficiency without changing other parameters, the incidence ratio for fava bean consumption will rise. A spreadsheet is a convenient way to see the effect on incidence ratios from varying the prevalences (check the web page for a downloadable Excel spreadsheet).

Bottom line – what we observe as strength of association is greatly dependent upon prevalence of other component causes.

The above example also illustrates the non-additivity of the attributable risk proportion [ARP=(RR-1)/RR]:

$$\text{ARP for GPDH deficiency} \quad \frac{67.67 - 1}{67.67} = 98.5\ \%$$

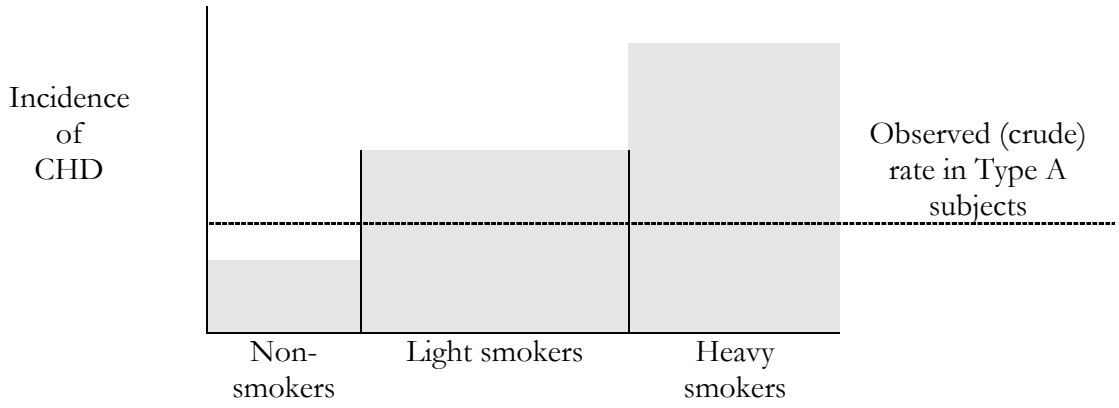
$$\text{ARP for Fava bean consumption} \quad \frac{4.33 - 1}{4.33} = 76.9\ \%$$

Clearly, these ARP's do not sum to 100%, nor, when we think about it, should they.

Before continuing with Rothman's diagrams, we need to revisit an old friend, weighted averages.

Crude rates as weighted averages

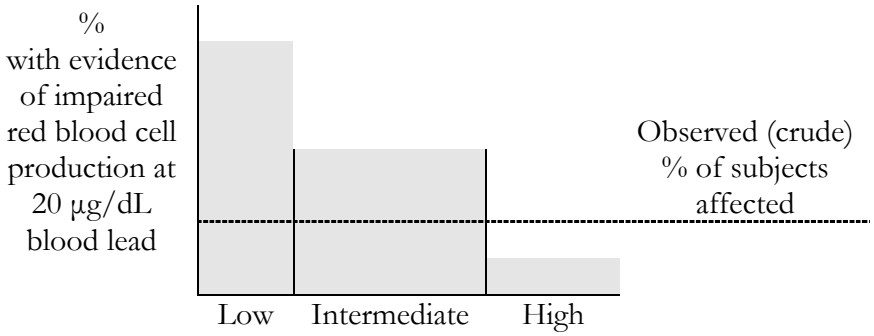
Recall the example of Type A behavior and CHD incidence with which we began the chapter on Confounding. In that example, smokers had a much higher incidence of CHD than did nonsmokers. Since the Type A group consisted mainly of smokers, its CHD incidence was greater than the Type B group, which consisted mainly of nonsmokers. If there were three smoking status groups, then the Type A incidence would be a weighted average of the rates for each of the three smoking status groups (see diagram).



So whenever we compare groups, it is important to pay attention to their distributions of risk factors. In the chapter on confounding, though, we considered only subgroups defined by other (independent) risk factors. We will now see that we must widen our view to include subgroups defined by variables that may influence the effect of the exposure even if those variables have no effect in its absence.

Since every rate we observe in some population is a weighted average of the rates for its component subgroups, this principle must apply to a group of exposed persons as well. Thus, the incidence in the exposed group depends on the composition of the group in regard to factors that are in the same causal pathways as the exposure. A prominent example is genetic factors, which thanks to the molecular biological revolution we are learning a great deal more about.

For example, it has been asserted that susceptibility to impairment of red blood cell production by low-level lead exposure varies according to the genetically-controlled level of the enzyme amino levulanate dehydratase. If that is the case, then in a group of children with a given level of blood lead (e.g., 20 micrograms/dL), the proportion with evidence of impaired red blood cell production would reflect a weighted average of the proportions in each subgroup defined by enzyme level:



Another example is that LDL cholesterol levels reflect both dietary intake of saturated fat and cholesterol and ApoE genotype (from Shpilberg *et al.*, 1997). Compared to persons with the most common allele (E3), those with the E2 allele have lower average cholesterol and those with the E4 allele have higher levels. Therefore, serum cholesterol levels associated with a given population distribution of dietary fat intake will depend on the distribution of these three genotypes.

Yet another example is incidence of venous thromboembolism. A strong effect of oral contraceptives (OC) on venous thromboembolism was one of the first hazards to be recognized for OC. Recent data from Vandenbroucke *et al.* (Factor V Leiden: should we screen oral contraceptive users and pregnant women? *Bio Med J* 1996;313:1127-1130) show an overall incidence of 0.8 per 10,000 women-years that rises to 3.0 per 10,000 women years with OC use. But among OC users who are also carriers of factor V mutation (associated with activated protein C (APC) resistance), the incidence rises to 28.5 per 10,000 women years (from Shpilberg *et al.*, 1997). So the incidence of venous thromboembolism in a population and the effects of OC will be greatly influenced by the population prevalence of factor V mutation.

So whatever phenomenon we are investigating, we need to take account of both independent risk factors for it and factors that may only appear to modify the effect of an exposure of interest (which we will subsequently refer to as an "effect modifier"). This is one reason why we typically stratify data by sociodemographic factors. Factors that affect susceptibility may well covary with demographic characteristics such as age, sex, geographic region, and socioeconomic resources, even if they do not have a role of their own in causation.

Since the distribution of effect modifiers may affect disease rates, it will also affect comparisons between rates in exposed and nonexposed subjects. But if the effect modifier is not itself a risk factor for the disease – i.e., if in the absence of the exposure of interest the effect modifiers is not associated with disease risk – then the modifier can confound associations only among groups with different levels of exposure, not between an exposed and an unexposed group.

Several examples will help to clarify these points. Assume for the moment, that asbestos has no effect on lung cancer incidence independent of smoking, but that smoking has an effect both alone and synergistically with asbestos. A study of the two factors might produce the following data:

Lung cancer rates by smoking and asbestos exposure (per 100,000 person years)

Smokers	
Exposed to asbestos	602
Not exposed to asbestos	123
Nonsmokers	
Exposed to asbestos	11
Not exposed to asbestos	11

From these data we would conclude (leaving aside all issues of statistical significance, bias, and so on) that (a) smoking increases lung cancer risk and (b) asbestos does so only in smokers. Smoking emerges as a risk factor, and asbestos as a modifier of the effect of smoking. Smoking could also be said to be an absolute modifier of the effect of asbestos, since the effect of the latter is null without smoking and dramatic in its presence. The rate ratios for lung cancer in smokers versus nonsmokers are 55 among those exposed to asbestos and 11 among those not exposed.

If we had not analyzed our data separately according to asbestos exposure, the lung cancer rate in nonsmokers would still be 11 per 100,000 person-years. But the rate in smokers would be somewhere between 123 and 602. The actual value would depend on the proportion of smokers exposed to asbestos. Similarly, the rate ratio for lung cancer and smoking would range between 11 and 55. So the crude rate ratio for lung cancer and smoking would always lie within the range of the stratum specific rate ratios.

The fact that the crude rate ratio differs from the stratum-specific rate ratios does not mean that confounding is present. Regardless of the proportion of subjects exposed to asbestos, the relationship between smoking and lung cancer cannot be due to asbestos exposure, though the strength of that relationship will depend on the degree of asbestos exposure. If the crude rate ratio can be expressed as a weighted average of the stratum-specific ratios, then confounding is not present.

The above results will always hold when the effect modifier has no effect in the absence of the exposure and the comparison of interest is between exposed and unexposed groups. A point of theoretical interest is that it was the above type of situation that led us in our discussion of confounding to focus on the question of an association between the potential confounder variable and the disease among the unexposed. An association among the exposed could reflect effect modification rather than independent causation (i.e., among exposed persons, disease rates are higher among those also exposed to a modifier, even if that is not the case among unexposed persons).

Since an effect modifier with no independent effect on the outcome does alter the risk or rate in the presence of exposure, however, an effect modifier can confound comparisons between groups exposed to different degrees. Suppose, for example, that we have divided the smokers in the previous table into light smokers and heavy smokers. Suppose further that most light smokers are exposed to asbestos and most heavy smokers are not. Then we might well observe a higher lung

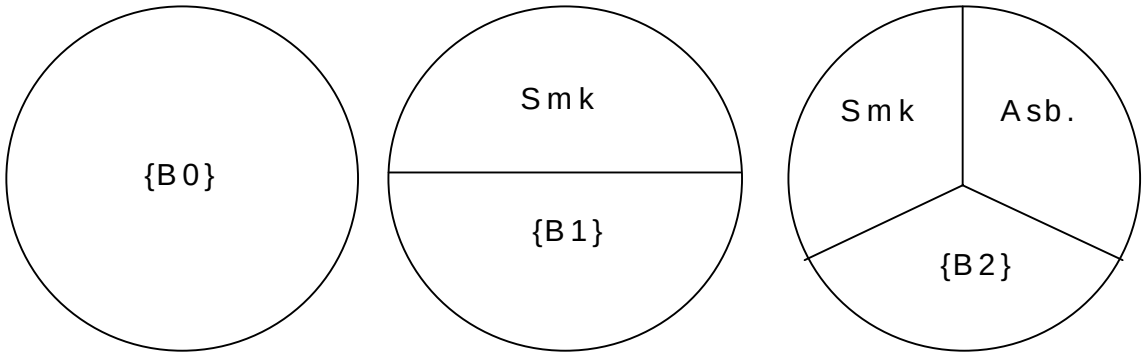
cancer rate among the light smokers (due to their greater asbestos exposure) than among the heavy smokers (where the rate has not been increased by asbestos). The following table gives a numerical illustration of such a situation.

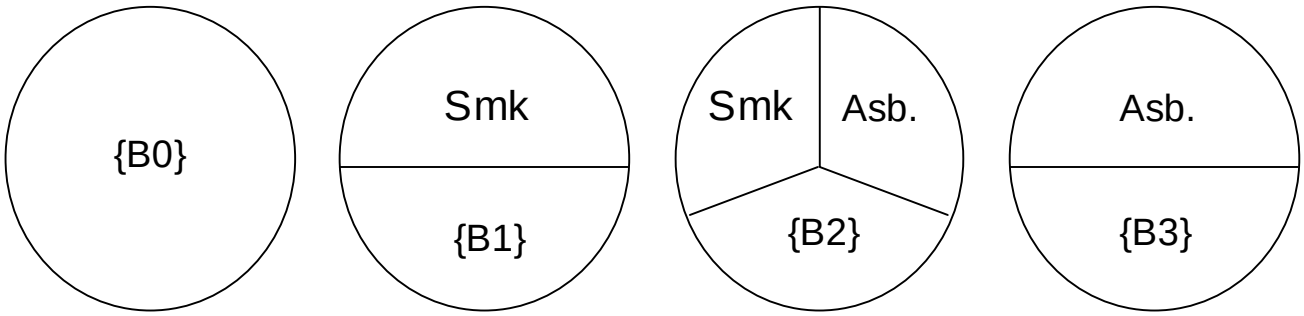
Lung cancer rates by level of smoking and asbestos exposure (per 100,000 person years)

Heavy smokers - overall	(200-1,000)
Exposed to asbestos	1,000
Not exposed to asbestos	200
Light smokers - overall	(100-500)
Exposed to asbestos	500
Not exposed to asbestos	100
Nonsmokers - overall	(11)
Exposed to asbestos	11
Not exposed to asbestos	11

Here, asbestos alone has no effect, heavy smoking in the absence of asbestos has rates twice that for light smoking, and asbestos increases lung cancer rates in smokers fivefold. If 60% of light smokers but only 10% of heavy smokers are exposed to asbestos, then the overall lung cancer rate in light smokers (340 = {500 x .60 + 100 x .40}) will exceed that in heavy smokers (280 = {1,000 x .10 + 200 x .90}).]

While the above situations may at first appear to be complex, they simply reflect different aspects of weighted averages, so with some practice the mystery evaporates. Additional complexity does enter the picture, however, when we turn to effect modification by a variable that has an effect on the outcome by a pathway that does not involve the exposure of interest, i.e., an independent effect. Compare these two causal schema:





where $\{B_0\}$, $\{B_1\}$, $\{B_2\}$, and $\{B_3\}$ are probably overlapping sets of (unidentified) background factors that are needed because (1) people exposed to neither cigarette smoke nor asbestos do get lung cancer, albeit at a low rate; (2) not all people who smoke get lung cancer; etc. [Note: these are the same as Rothman's {U}. I prefer to use different subscripts to make clear that different causal pathways generally involve different required background factors. Otherwise all persons susceptible to developing the disease through a causal pathway involving an exposure (e.g., smoking) would get the disease regardless through the "unexposed" pathway even if not exposed, so the exposure could not be associated with an increased rate of disease.]

The three-pathway configuration represents the situation we have just seen, where asbestos has no effect in nonsmokers. If we apply the data from the preceding numerical example into the upper configuration of causal pathways, we see that the rate that corresponds to the first causal pathway ($\{B_0\}$) is 11/100,000 py. The rate that corresponds to the second causal pathway (Smk|$\{B_1\}$) is 112/100,000 py (123 - 11: the incidence density difference, since people who smoke and can therefore get disease through the second causal pathway are also at risk of developing the disease through the first causal pathway). Similarly, the rate that corresponds to the third causal pathway (Smk|Asb|$\{B_2\}$) is (602-112-11)/100,000 py (since we observe 602 for people who have both exposures, but they could have developed disease from either of the first two causal pathways). These different disease rates presumably correspond to the prevalences of $\{B_1\}$, $B_2\}$, and $\{B_3\}$.

The four-pathway configuration does show an <u>independent effect</u> of asbestos. In this configuration, we see that confounding by asbestos can occur, since the risk in nonsmokers may be elevated by the effect of asbestos. Moreover, it now becomes more difficult to assess effect modification as "a combined effect greater than we expect from the effects of each variable acting alone". The problem is: if each variable has an effect on its own, what do we expect for their combined effect so we can say whether we have observed something different from that?

Consider, for example, actual data on the relationship of smoking and asbestos to lung cancer death rates (from E. Cuyler Hammond, Irving J. Selikoff, and Herbert Seidman. Asbestos exposure, cigarette smoking and death rates. *Annals NY Acad Sci* 1979; 330:473-90).

Age standardized lung cancer rates by smoking and asbestos exposure (per 100,000 person years)

	Smokers	Nonsmokers
Exposed to asbestos	602	58
Not exposed to asbestos	123	11

When we calculate the disease rates that correspond to each of the four causal pathways in the lower configuration of causal "pies" above, the two leftmost pathways have the same rates as in the upper configuration. The rate corresponding to the rightmost pathway (Asbestos|{B3}) is 58-11 = 47/100,000 py. The rate that corresponds to the third causal pathway (Smk|Asb|{B2}) is now reduced since some of cases with both exposures could be due to the effect of asbestos. So the rate that corresponds to the third pathway is now (602-112-11-47)/100,000 py = 410/100,000 py.

We might take these rates and reason as follows:

Increase due to smoking	123 - 11 = 112
Increase due to asbestos	58 - 11 = 47
Total increase expected due to both	112 + 47 = 159
Total observed increase	602 - 11 = 591 !

Since the increase due to the combined effect greatly exceeds that expected from our (additive) model, we would conclude that the effect is synergistic.

Alternatively, we might reason in relative terms:

Relative increase due to smoking	123 / 11 = 11.2
Relative increase due to asbestos	58 / 11 = 5.3
Total increase expected due to both	11.2 x 5.3 = 59.4
Total observed increase	602 / 11 = 54.7

This time the observed increase and that expected from our (multiplicative) model are quite close, so we conclude that there is no effect modification. We are thus faced with a situation where the decision about effect modification depends upon what model we employ to arrive at an expected joint effect to compare with the observed joint effect (or equivalently, upon the scale of measurement, hence the term "effect measure modification").

Before pondering this dilemma further, we should first state the additive and multiplicative models explicitly. To do so we introduce a notation in which "1" indicates presence of a factor, a "0" indicates absence of a factor, the first subscript represents the first risk factor, and the second subscript represents the second risk factor (see below).

Notation for joint effects

R_1	risk or rate in the presence of a factor, ignoring the presence or absence of other identified factors
R_0	risk or rate in the absence of a factor, ignoring the presence or absence of other identified factors
R_{11}	risk or rate when both of two factors are present
R_{10}	risk or rate when the first factor is present but not the second
R_{01}	risk or rate when only the second factor is present
R_{00}	risk or rate when neither of the two factors is present (i.e., risk due to background factors)
RD_{11}	difference between the risk or rate when both factors are present and the risk or rate when neither factor is present
RD_{10}	difference between the risk or rate when only the first factor is present and the risk or rate when neither factor is present
RD_{01}	difference between the risk or rate when only the second factor is present and the risk or rate when neither factor is present
RR_{11}	ratio of the risk or rate when both factors are present divided by the risk or rate when neither factor is present
RR_{10}	ratio of the risk or rate when only the first factor is present divided by the risk or rate when neither factor is present
RR_{01}	ratio of the risk or rate when only the second factor is present divided by the risk or ratio when neither factor is present

The use of two subscripts implies a stratified analysis. The first subscript indicates presence or absence of the first factor; the second subscript, presence or absence of the second factor. For example, R_{10} refers to the rate for persons exposed to the first factor but not to the second. That rate can be referred to as the rate for the exposed (to factor 1) in the stratum without factor 2; equivalently, R_{10} can be referred to as the rate for the unexposed (to factor 2) in the stratum where factor 1 is present. In contrast, a single subscript (R_1) means the factor is present, with other factors present or not present (i.e., crude with respect to other factors). "Background" factors and the risk R_{00} associated with them are assumed to be uniformly distributed across all strata.

Additive model

Under an additive model, the increase in rate or risk from a combination of factors equals the sum of the increases from each factor by itself. We can express this statement algebraically, using the rate (or risk) difference:

$$R_{11} - R_{00} = R_{10} - R_{00} + R_{01} - R_{00} \quad \text{(A1)}$$

$$RD_{11} = RD_{10} + RD_{01} \quad \text{(A2)}$$

Using elementary algebra and the definition of the rate difference, we can also write the additive model as:

$$R_{11} = R_{00} + RD_{10} + RD_{01} \quad \text{(A3)}$$

i.e., the expected rate where both factors are present is the baseline rate (R_{00}, neither factor present) plus the rate difference associated with the first factor plus the rate difference associated with the second factor. Another equivalent expression is:

$$R_{11} = R_{10} + R_{01} - R_{00} \quad \text{(A4)}$$

Since $RR_{11} = R_{11}/R_{00}$, $RR_{10} = R_{10}/R_{00}$, and $RR_{01}=R_{01}/R_{00}$, we can express the additive model in terms of the risk (or rate) ratio, by dividing each term in expression A1 by the baseline risk, R_{00}.

$$RR_{11} - 1 = RR_{10} - 1 + RR_{01} - 1 \quad \text{(A5)}$$

An advantage of this formulation is that we can use it even when we do not have estimates of specific risks or risk differences. The expression $(R_1-R_0)/R_0$, or RR – 1, is sometimes referred to as the (relative) **excess risk**. The additive model, expressed in terms of excess risk, is therefore:

Excess risk for A and B together = Excess risk for A + Excess risk for B

i.e., the joint excess risk equals the sum of the excess risk for each factor alone. With this expression we can evaluate the additive model even from case-control data.

More than two factors

Where there are three factors, we have, analogously:

$$RR_{111} - 1 = RR_{100} - 1 + RR_{010} - 1 + RR_{001} - 1 \quad \text{(A6)}$$

$$RD_{111} = RD_{100} + RD_{010} + RD_{001} \quad \text{(A7)}$$

$$R_{111} = R_{000} + RD_{100} + RD_{010} + RD_{010} \quad \text{(A8)}$$

and

$$R_{111} = R_{100} + R_{010} + R_{001} - 2\,R_{000} \quad \text{(A9)}$$

So the additive model can be regarded as based on 1) additivity of excess risks, 2) additivity of risk differences, and/or 3) additivity of the risks themselves. The reason that we need to subtract the baseline risk in the last of these forms is that risk in the presence of any of the factors includes, necessarily, the ever-present background risk. So when we add the risk for one factor to the risk for another factor, the background risk is added twice. Thus, when we refer to R_{ijk} as the risk (or rate) for a factor "by itself", the **"by itself" really means "with no other specified factors"**, since the **baseline risk is, by definition, always present**.

Multiplicative model

In parallel fashion, the multiplicative model assumes that the relative risk (risk ratio, rate ratio) for the factors operating together equals the product of their relative risks:

$$RR_{11} = RR_{10} \times RR_{01} \quad \text{(M1)}$$

Multiplying through by baseline risk (R_{00}) gives:

$$R_{11} = R_{00} \times RR_{10} \times RR_{01} \quad \text{(M2)}$$

and

$$R_{11} = R_{10} \times R_{01} / R_{00} \quad \text{(M3)}$$

i.e., the joint risk equals the product of 1) the baseline risk multiplied by the relative risk for each factor and/or 2) the individual risks and the reciprocal of the baseline risk. For three factors, the model becomes:

$$RR_{111} = RR_{100} \times RR_{010} \times RR_{001} \quad \text{(M4)}$$

and

$$R_{111} = R_{000} \times RR_{100} \times RR_{010} \times RR_{001} \quad \text{(M5)}$$

and

$$R_{111} = R_{100} \times R_{010} \times R_{001} / (R_{000})^2 \quad \text{(M6)}$$

Again, there is a baseline risk or rate in the denominator of each relative risk, so when the relative risks are converted to risks, the R_{000} in the numerator eliminates one of the resulting three R_{000}'s, leaving two remaining in the denominator. As before, "by itself" means without other specified factors, but including baseline risk.

Note, however, that the multiplicative model can also be written as an additive model on the logarithmic scale (because addition of logarithms is equivalent to multiplication of their arguments):

$$\ln(R_{111}) = \ln(R_{100}) + \ln(R_{010}) + \ln(R_{001}) - 2 \times \ln(R_{000}) \quad \text{(M7)}$$

For this reason, the difference between the additive and multiplicative models can be characterized as a transformation of scale. So "effect modification" is scale-dependent.

Optional aside – It can also be shown that a multiplicative model can be expressed as an additive model on the natural scale plus an interaction term. For two factors: $(R_{10} - R_{00})(R_{01} - R_{00})/R_{00}$, or equivalently, $(R_{00})(RR_{10}-1)(RR_{01}-1)$ – essentially, we add a "fudge factor".

Additive model:

$$R_{11} = R_{10} + R_{01} - R_{00}$$

Additive model with interaction term:

$$R_{11} = R_{10} + R_{01} - R_{00} + R_{00} \times (RR_{10}-1) \times (RR_{01}-1)$$

Multiplying out the interaction term:

$$R_{11} = R_{10} + R_{01} - R_{00} + R_{00} \times RR_{10} \times RR_{01} - R_{00} \times RR_{10} - R_{00} \times RR_{01} + R_{00}$$

Dividing both sides by R_{00}:

$$RR_{11} = RR_{10} + RR_{01} - 1 + RR_{10} \times RR_{01} - RR_{01} - RR_{10} + 1$$

Simplifying:

$$RR_{11} = RR_{10} \times RR_{01} = \text{the multiplicative model}$$

[End of aside]

The choice of model – additive, multiplicative, or other – is not a settled affair and involves a variety of considerations. One consideration is to choose the simplest model that can represent the data. Recall the example from an earlier lecture:

Relative versus Absolute Effects example Incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years

Age	Cigarettes/day	OC*	$\overline{OC}$*	RR**	AR***
30-39	0-14	6	2	3	4
	15 +	30	11	3	19
40-44	0-14	47	12	4	35
	15 +	246	61	4	185

Notes:
* Rate per 100,000 women-years
** RR=relative risk (rate ratio)
*** AR=attributable risk (rate difference)
Source: Mann *et al.* (presented in a seminar by Bruce Stadel)

Here, we saw that the rate ratio was a more stable index of the strength of association between OC and MI across the various combinations of age and smoking. In fact, the MI rates for many combinations of the three risk factors – age, smoking, and OC – are not far from those expected based on the multiplicative model. To see this, use the additive and multiplicative models just presented with the data in the above table to fill in the rightmost two columns of the following table. If we write the rates for the three risk factors as R_{100}, R_{010}, and R_{001}, with the background rate defined as R_{000}, then joint rates for several combinations of risk factors would be:

First and third factors present (row 6):

$$R_{101} = R_{100} + R_{001} - R_{000} \quad \text{(additive model)}$$

$$R_{101} = R_{100} \times R_{001} / R_{000} \quad \text{(multiplicative model)}$$

First and second factors present (row 7):

$$R_{110} = R_{100} + R_{010} - R_{000} \quad \text{(additive model)}$$

$$R_{110} = R_{100} \times R_{010} / R_{000} \qquad \text{(multiplicative model)}$$

All three factors present (row 8):

$$R_{111} = R_{100} + R_{010} + R_{001} - 2\,R_{000} \quad \text{(additive model)}$$

$$R_{111} = R_{100} \times R_{010} \times R_{001} / (R_{000})^2 \quad \text{(multiplicative model)}$$

where the three factors are 1) age, 2) cigarette smoking, and 3) oral contraceptives. For example, suppose R_{101} is the rate of MI in women who are in the older age group, smoke less than 15 cigarettes/day or not at all, and use oral contraceptives. The multiplicative model says that the rate for any combination of the three factors (with cutpoints defined as in the table) equals the product of the rates for each of the three factors when neither of the other two is present, divided by the square of the rate for those who have none of the three factors (i.e., only unidentified background factors are present). Here is a "test" of the model (one line is left incomplete, to give you the satisfaction of figuring it out):

Home-made multiplicative model of Incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years

Row		Age	Cigarettes /day	OC*	Observed Rate	Expected (Multiplic)	Expected (Additive)
1	R_{000}	0: 30-39	0: 0-14	0: no	2	-	-
2	R_{001}	0: 30-39	0: 0-14	1: yes	**6**	-	-
3	R_{010}	0: 30-39	1: 15 +	0: no	**11**	-	-
4	R_{100}	1: 40-44	0: 0-14	0: no	**12**	-	-
5	R_{011}	0: 30-39	1: 15 +	1: yes	30	___	___
6	R_{101}	1: 40-44	0: 0-14	1: yes	47	36	16
7	R_{110}	1: 40-44	1: 15 +	0: no	61	66	21
8	R_{111}	1: 40-44	1: 15 +	1: yes	246	198	25

Notes: 0: and 1: indicate the coding for each risk factor level. Rates for single factors in the absence of the other two are shown in bold.

[Thanks to Jim McDougal (1996) for spotting my longstanding errors in the 3-factor interaction in this table and its explanation.]

Certainly the multiplicative model yields expected rates that are closer to the observed rates for various combinations of the factors than does the additive model. The better fit for the

multiplicative model supports the use of the rate ratio as the measure of association for each risk factor and each risk factor combination in these data. If Mann *et al*. want a summary measure for the effect of OC on MI rates, controlling for age and smoking, a weighted average of the rate ratios (3, 3, 4, 4) for OC use across the four age and smoking categories would be a good choice. But then what happened to effect modification?

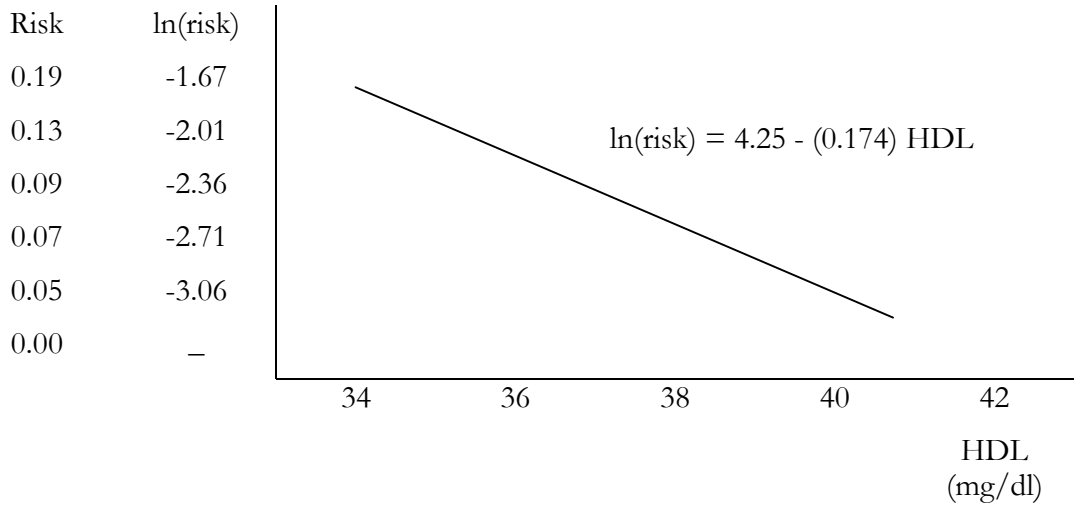
The "natural" scaling

The additive model has been put forth by Rothman as the "natural" scaling. Risks are probabilities, and the probability that either of two independent and mutually exclusive events will take place (e.g., smoking causes MI or OC causes MI) is the sum of the probabilities for each. Therefore if the risk (probability of disease) in people with both exposures exceeds the sum of the risks for each exposure separately, then some non-independence (i.e., interaction) must exist between these two disease events. Rothman's proposition appears to have become the consensus in terms of evaluating impact on public health and/or individual risk (see below). Our earlier suggestion that the risk or rate difference serves more often as a measure of impact than as a measure of strength of association in respect to etiology is distinctly parallel.

When our interest is the relationship of the mathematical model or scaling to possible biological mechanisms, however, the issue becomes more problematic. Kupper and Hogan (Interaction in epidemiologic studies. *Am J Epidemiol* 108:447-453, 1978) demonstrated how two factors having biologically equivalent modes of action, so that either factor can be regarded as a different concentration of the other, can appear to be synergistic in their joint effect if the dose-response curve is nonlinear. (This example harks back to the fact that additivity on the logarithmic scale is equivalent to multiplicativity on the natural scale.) Therefore, a departure from additivity can occur even in the absence of biological interaction.

Data from a study published in that year provides an illustration. Bradley DD, *et al*. (Serum high-density-lipoprotein cholesterol in women using oral contraceptives, estrogens and progestins. *New Engl J Med* 299:17-20, 1978) suggested that smoking and oral contraceptives (OC) may each increase myocardial infarction risk by reducing levels of HDL cholesterol. The effects of smoking and oral contraceptives on HDL appear to be additive. But if the relationship between HDL level and myocardial infarction risk is exponential, with the logarithm of risk increasing in linear fashion with declining HDL, then the effects of the two behavioral risk factors on myocardial infarction risk will be multiplicative.

In the figure below, the natural logarithm of heart attack risk is a linear function of HDL level, so that risk rises exponentially as HDL decreases. The risk function comes from Bradley *et al*.'s paper.



If smoking causes a reduction in HDL of 6 mg/dL, and oral contraceptives cause a reduction of 2 mg/dL, then the changes in ln(risk) [from the formula in the figure] and in the RR's for smoking and oral contraceptives separately and for both together are shown in the following table:

Factors	**HDL reduction**	**Increase in ln(risk)**	**RR**
Smoking	6	1.044	2.84
OC only	2	0.348	1.42
Smoking and OC	8	1.392	4.02

Smoking is associated with a 6 mg/dL lower HDL level, corresponding to an increase in ln(risk) of 1.044, which in turn corresponds to a relative risk of 2.84. Although (in this conceptual model) the biological effects of smoking and OC on HDL are additive, because the dose-response curve is not linear, this additivity of dose does not imply additivity of response.

This point has been elaborated by Thompson (1991), who makes the point that pathogenetic processes are likely to include factors that intervene between the variables in our simplified causal models. Such intervening factors are generally unknown or unmeasured by epidemiologists. Yet as illustrated above, the form of the functional relation between two variables can change the appearance of a risk function. The actions of two factors may be additive on their immediate target, but their effect on risk of a downstream effect could be additive, multiplicative, or anything else. Only in the case of a crossover effect (a.k.a. qualitative interaction, which to be certain that it exists should be demonstrated by confidence intervals that lie wholly below the null value in one stratum and wholly above the null value in the other stratum – see Thompson 1991) do we have a basis for inferring that something of biological interest is occurring (after excluding other non-mathematical explanations). Another situation where interpretation is unambiguous – what I have called "absolute effect modification", where one factor has no effect in the absence of the other – is in practice just as problematic as other non-crossover situations, since it is rarely possible to exclude the presence of at least a weak effect (Thompson 1991).

Effect modification as a reflection of information bias:

Another consideration that arises in interpreting apparent effect modification in epidemiologic data relates to the question of the actual dosage received by subjects. Suppose that data from a study of lung cancer and smoking yielded these results:

Lung cancer rates per 100,000 person-years

	Males	**Females**
Smokers	300	500
Nonsmokers	50	50

The rate ratios for males and females are 6 (300/50) and 10 (500/50), respectively, which might suggest that women are more susceptible to the carcinogenic properties of tobacco smoke. But what if women smokers inhale more deeply and therefore receive a larger dose of carcinogenic substances, the actual exposure? So whereas effect measure modification in epidemiologic data may suggest the need for a more detailed understanding of the phenomenon under study, an interpretation in terms of biological synergism involves causal inference and needs to be approached from that perspective.

Consensus

Rothman, Greenland, and Walker (1980) presented four perspectives on the concept of interaction:

1. The biologic perspective is concerned with elucidating how various factors act at the biological (mechanistic) level.
2. The statistical perspective treats interaction as "leftovers", i.e., the nonrandom variability in data that is not accounted for by the model under consideration. Statisticians often try to reformulate the model to eliminate these leftovers, i.e., to find the simplest model that fits the data adequately.
3. The public health perspective should regard interaction is a departure from additivity, if one assumes that costs are proportional to the number of cases. If effects are more than additive, then a greater than proportional payoff can be obtained by intervening against a factor involved in an interaction.
4. The individual decision-making perspective should also regard interaction as a departure from additivity, again assuming a linear relationship between costs and, in this case, risk. For example, if the combined effect of smoking and hypertension on CHD risk is greater than additive, someone with hypertension can reduce his risk even more by quitting smoking than someone with normal blood pressure.

These perspectives appear to be widely accepted. The term "effect modification" is generally used to refer to a meaningful departure from a given mathematical model (i.e., additive, multiplicative, or whatever) of how risks or rates combine. ("Meaningful" means that the departure is large enough to

have clinical or public health significance and thought not to be due to random variability, measurement inadequacy, or confounding.) The additive model appears to be accepted as the indicator of "expected joint effects" for policy or decision-making considerations.

Summary

In view of the foregoing, we may attempt to summarize the relevance of interaction and effect modification in terms of four implications:

1. Increasing the precision of description and prediction of phenomena under study;
2. Indicating the need to control for the factors that appear as modifiers;
3. Suggesting areas for developing etiologic hypotheses; and
4. Defining subgroups and factor combinations for special attention for preventive approaches.

Elaboration

1. Increasing precision of description:

In our smoking in men and women illustration, the different strength of the smoking-lung cancer association between men and women may lead to an appreciation of the need to be more precise in the measurement and specification of the exposure variable.

2. Indicating the need to control for modifiers:

Since an effect modifier changes the strength of the association under study, different study populations may yield different results concerning the association of interest. Unlike potential confounders, modifying variables cannot create the appearance of an association (for exposed versus unexposed) where none exists. But the proportion of the study population that has a greater susceptibility will influence the strength of the association. Therefore, to achieve comparability across studies, it is necessary to control for the effect of the modifying variables, generally by carrying out a separate analysis at each level of the modifier.

3. Developing etiologic hypotheses:

Attention to interactions in the data may lead to the formulation of etiologic hypotheses that advance our understanding of the pathogenetic processes involved. Although the linkage between mechanisms and relationships in data is uncertain, a strong interaction might suggest that a shared mechanism is involved. For example, the interaction of smoking and asbestos might suggest a scenario such as impairment of lung clearing processes and/or of mechanical injury from asbestos particles increases susceptibility to carcinogens in cigarette smoke.

4. Defining subgroups for preventive approaches:

To observe that the OC-MI association is particularly strong among smokers and/or women over 35 carries evident preventive implications in terms of health education warnings, contraindications to prescribing, targeting of messages, and so forth. The synergistic relationship between smoking and asbestos in the etiology of lung cancer suggests the value of extra efforts to convince asbestos workers not to smoke. If the cost of helping a smoker to quit smoking is the same for asbestos workers and others, then the benefit-cost ratio will be greater for a cessation program with smokers who work with asbestos because more cases of lung cancer will be avoided for the same number of quitters.

The rationale for viewing effect modification as a departure from an additive model of disease risks, at least for public health purposes, is that if an additive model holds, then removal of one agent can only be expected to eliminate the risk that arises from that agent but not the risk from other agents. If there is positive interaction, however, removal of any one of the agents involved will reduce some risk resulting from the other as well. In such a situation, the impact of removing a risk factor is greater than that expected on the basis of its effect on baseline risk.

A "real-life" example

The following table comes from a randomized, controlled trial of a self-help smoking cessation intervention using brief telephone counseling. Quit rates for smokers in the intervention group and the other groups suggested that participants with certain baseline characteristics were more or less likely to benefit from the telephone counseling intervention. For example, the telephone counseling intervention was associated with a 14 percentage point (31%–17%) higher quit rate for participants who were <u>not</u> nicotine dependent but with only a 3 percentage point (17%–14%) higher quit rate for participants who were nicotine dependent. The intervention was associated with a 12 percentage point (29%-17%) higher quit rate for participants who <u>had not</u> previously undergone an intensive cessation program but with only a 2 percentage point (17%–15%) higher quit rate for participants who <u>had</u>. The observed differences appeared to be consistent with the fact that the intervention was a minimal treatment (so would not be of much help to a smoker who had already experienced an intensive treatment program) that incorporated nicotine-fading/brand-switching (which has limited applicability for a smoker who is already smoking a low-nicotine brand).

Baseline Characteristics Associated with a Significantly Different Telephone Counseling Effect on 7-day Abstinence at 16-months Follow-up in 1,877 Smokers at Group Health Cooperative of Puget Sound, Washington, 1985-1987

	Quit rate			
	with characteristic		without characteristic	
Baseline characteristic	**Counseling**	**No Counseling**	**Counseling**	**No Counseling**
Nicotine dependent	17	14	**31**	**17**
Intensive treatment	17	15	**29**	**17**
Brand nicotine > 0.7 mg	**24**	**12**	22	20
VIP better role model	**28**	**15**	19	16
Close friends/relatives	21	17	**29**	**14**
Nonsmoking partner	19	19	**25**	**14**

Note: For each characteristic, the difference in quit rates between counseling and no-counseling groups among those with the characteristic is significantly ($p<0.05$) greater or less (by about 10 percentage points) than the quit rate difference among those without the characteristic. Bolding denotes the greater telephone counseling effect.

Reference: Schoenbach VJ, Orleans CG, Wagner EH, Quade D, Salmon MAP, Porter CQ. Characteristics of smokers who enroll and quit in self-help programs. *Health Education Research* 1992;7:369-380, Table 3.

Bibliography

Rothman and Greenland, *Modern epidemiology*.

Hertz-Picciotto I, Neutra RR. Resolving discrepancies among studies: the influence of dose on effect size. *Epidemiology* 1994;5:156-163.

Koopman, James S. and Douglas L. Weed. Epigenesis theory: a mathematical model relating causal concepts of pathogenesis in individuals to disease patterns in populations. *Am J Epidemiol* 1990; 132:366-90.

Marshall, Roger J. Confounder prevalence and stratum-specific relative risks: implications for misclassified and missing confounders. *Epidemiology* 1994;5:439-44

Koopman, J.S.: Interaction between discrete causes. *Am J Epidemiol* 113:716-724, 1981.

Khoury, Muin J.; W. Dana Flanders, Sander Greenland, Myron J. Adams. On the measurement of susceptibility in epidemiologic studies. *Am J Epidemiol* 1989; 129:183-90.

Rothman, KJ. Causes. *Am J Epidemiol* 1976; 104:587-92.

Rothman, K.J.: Synergy and antagonism in cause-effect relationships. *Am J Epidemiol* 99:385-388, 1974.

Rothman, K.J., Greenland, S., and Alexander M. Walker: Concepts of interaction. *Am J Epidemiol* 112:467-470, 1980.

Shpilberg O, Dorman JS, Ferrell RE, Trucco M, *et al*. The next stage: Molecular epidemiology. *J Clin Epidemiol* 1997;50:633-638.

Siemiatycki, Jack and Duncan C. Thomas. Biological models and statistical interactions: an example from multistage carcinogenesis. *International J Epidemiol* 10:383-387, 1981.

Thompson, W. Douglas. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991;44:221-232.

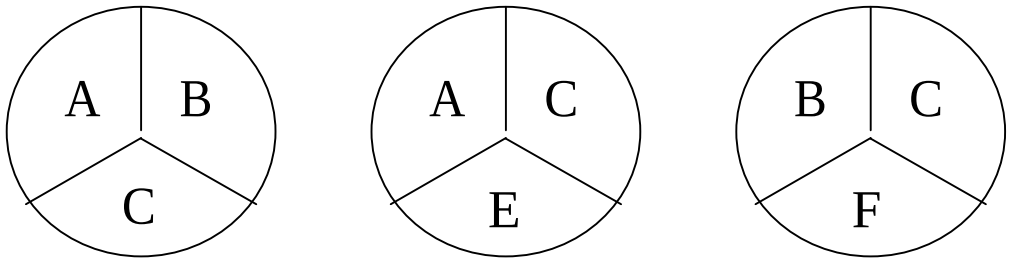
Walker, Alexander M. Proportion of disease attributable to the combined effect of two factors. *International J Epidemiology* 1981; 10:81-85.

Weed, Douglas L.; Michael Selmon, Thomas Sinks. Links between categories of interaction. *Am J Epidemiol* 1988; 127:17-27.

Weiss, Noel S. Accounting for the multicausal nature of disease in the design and analysis of epidemiologic studies. *Am J Epidemiol* 1983;117:14-18.

Multicausality: Effect modification - Assignment

1. Answer the following questions based on the Rothman-style diagram for the etiology of "incidensititis"



   a. Which component cause is a necessary cause?

   b. Which two component causes have both modifying and independent (of each other) effects?

   c. If the population prevalence of C is very high, prevalence of E and F are very low, and prevalence of A is less than that of B, which will be perceived as the "stronger" cause, E or F?

   d. What are some implications of a situation like the above for studying the incidence of "incidensititis"?

2. Evaluate the joint effect of cigarette smoking and asbestos exposure on the lung cancer rate of the following groups of workers.

Incidence of Lung Cancer Per 100,000 person-years

Asbestos exposure	Nonsmokers	Smokers
No	10	100
Yes	20	400

Select <u>one best interpretation</u> based on these data, and write a brief justification for your choice.

A. Synergism apparently exists in these data because the joint effect of smoking and asbestos exposure is greater than the effect of smoking alone or of asbestos alone.

B. Smoking appears to have a synergistic effect and also to be a confounder, since smoking is associated with asbestos exposure and is a proven risk factor for lung cancer.

C. Smoking appears to have a synergistic effect because the rate ratio of lung cancer among smoking asbestos workers is greater than what would be expected on the basis of the individual rate ratios of smoking alone and of asbestos alone.

D. Smoking appears to have synergistic effect because the excess rate (RR-1) for smoking and asbestos together is greater than the sum of the excess rates for smoking alone and asbestos alone.

E. From these data, it is not possible to evaluate synergism since we do not know the distribution of smoking habits among smokers who are exposed to asbestos and among those who are not.

3. Consider the following data based on the Royal College of General Practitioners Oral Contraceptive Study (1977).

Mortality rates per 100,000 women-years from cardiovascular disease (ICD 390-458) by smoking habit at entry and oral contraceptive use (Standardized by age, social class, and parity).

Cigarette smoking status	Oral Contraceptive Status	
	User	Non User
Non-Smoker	13.8	3.0
Smoker	39.5	8.9

a. Formulate an expression for the joint effect of oral contraceptive use (OC) and smoking on cardiovascular disease mortality, based on an <u>additive</u> model, and determine whether the rates in the above table fit such a model (do not do any statistical tests).

b. Formulate an expression for the joint effect of oral contraceptive use and smoking on cardiovascular disease mortality, based on a <u>multiplicative</u> model, and determine whether the rates in the above table fit such a model (do not do any statistical tests).

c. In commenting on the mortality rates for OC and smoking, a prominent epidemiologist remarked that "the relative risk for oral contraceptive users, compared to non-users, is the same for smokers and non-smokers." Other observers have characterized the relationship as synergistic. Briefly discuss the issues underlying the assessment of synergism in the above data.

4. Several studies have shown a synergistic effect between smoking and drinking in their relation to oral cancer. Consider these hypothetical data:

Yearly Incidence Rates per 100,000 at Risk

	Drinker	Non-Drinker
Smoker	100	40
Non-smoker	15	10

   a. Draw a diagram using Rothman's "causal pies" to show pathways by which oral cancer occurs.

   b. Assuming that there are 100,000 smoker/drinkers, 100,000 smoker/non-drinkers, 100,000 drinker/non-smokers and 100,000 non-drinker/non-smokers, how many cases of oral cancer would be prevented in one year if (only) smoking were eliminated?

   c. How many cases of oral cancer would be eliminated if (**only**) drinking were eliminated?

   d. How many cases of oral cancer would be prevented if **both** smoking and drinking were eliminated?

   e. How many cases of oral cancer can be attributed to **each** causal pathway you have identified in part a.?

   f. Explain why the answers to b. and c. do not sum to the answer in d.

5. Walker (*International Journal of Epidemiology* 1980; 10:81) suggests a measure to estimate the proportion of cases due to the synergism between two factors, which he calls the etiologic fraction due to interaction $EF_{(AxB)}$.

$$EF_{(AxB)} = \frac{\text{Observed rate for A and B together} - \text{Expected rate if there were no synergy}}{\text{Observed rate for A and B together}}$$

   a. For the data in question 9, what is the observed rate for oral cancer among the smokers and drinkers?

   b. What rate would you expect to see if there is no synergism between smoking and drinking?

   c. Calculate the $EF_{AxB)}$.

   d. Suggest a public health application for this result.

   (Thanks to Stephen Kritchevsky, Ph.D., for questions 4 and 5.)

Multicausality: Effect modification - Assignment solutions

1.

a. C - because it is a component of all three sufficient causes of "incidensititis." C is a necessary cause since "incidensititis" cannot occur in the absence of C.

b. A and B - Modification implies that two component causes are members of the same sufficient cause for "incidensititis." Independence implies that two component causes are members of different sufficient causes for "incidensititis." Both A and B fulfill these requirements.

c. F will appear as the stronger cause. Since C will be present in most people, and the prevalence of B is greater than the prevalence of A, people with F will be more likely to develop incidensititis than will people with E.

d. The most important implication for our purposes is the need to control the other factors when studying the effect of A. If our A and not-A groups differ with respect to B, C, and E, then the disease rates observed could be due to the latter factors, rather than to A. Moreover, the effect of A will appear to differ from study to study unless these other factors are taken into account.

It is also interesting to note that the multiplicity of sufficient causes imply different "etiologic routes" to incidensititis. So, for example, a person could acquire incidensititis through the first sufficient cause and never have either component cause E or F. Thus, cases of incidensititis will be heterogenous with regard to the etiology of their disease. The only common (necessary) cause is C, which must be present for disease occurrence.

2. D. Smoking appears to have synergistic effect because the excess rate (RR-1) for smoking and asbestos together is greater than the sum of the excess rates for smoking alone and asbestos alone.

3.

a. Under an additive model, we expect that the joint effect of the two factors will be equal to the sum of the excess risk from each factor separately, i.e.,

Expected Rate Difference (RD) of OC and SMK together =

$$\text{Expected } RD_{OC,SMK} = RD_{OC,\overline{SMK}} + RD_{\overline{OC},SMK}$$

(or equivalently, the rate for persons exposed to both factors together is expected to be equal to the rate for those exposed to neither plus the increase associated with the first factor alone plus the increase associated with the second factor alone):

$$\text{Expected } R_{OC,SMK} = R_{neither} + (R_{OC,\overline{SMK}} - R_{neither}) + (R_{\overline{OC},SMK} - R_{neither})$$

$$= R_{OC,\overline{SMK}} + R_{\overline{OC},SMK} - R_{neither}$$

In the data from the table,

Expected $R_{OC,SMK} = 13.8 + 8.9 - 3.0 = 19.7$

Observed $R_{OC,SMK}$ = 39.5 per 100,000 women-years.

Or,

$$\underline{\text{Expected}} \text{ excess risk } (RR - 1) \text{ (of OC alone + SMK alone)} = \left(\frac{13.8}{3.0} - 1\right) + \left(\frac{8.9}{3.0} - 1\right)$$

$$\underline{\text{Observed}} \text{ excess risk } (RR - 1) \text{ (of OC alone + SMK alone)} = \left(\frac{39.5}{3.0} - 1\right)$$

The large discrepancy between expected and observed rates indicates that the data do not fit an additive model.

b. Under a multiplicative model, we expect the joint effect of the two factors to be equal to the product of the risk (or rate) ratios for each factor separately, i.e.:

Expected Rate Ratio (RR) for OC and SMK together,

$$RR_{OC,SMK} = (RR_{OC,\overline{SMK}})\,(RR_{\overline{OC},SMK})$$

or equivalently, the risk or rate (R) for OC and SMK together is:

$$\text{Expected } R_{OC,SMK} = \frac{(R_{OC,SMK})(R_{OC,SMK})}{R_{OC,SMK}}$$

In these data,

$$\text{Expected } RR_{OC,SMK} = \frac{(13.8)\,(9.9)}{3.0} = 40.9$$

$$\text{Observed } RR_{OC,SMK} = 39.5$$

or,

$$\text{Expected } RR_{OC,SMK} = RR_{OC,SMK} \times RR_{OC,SMK}$$

$$= \frac{13.8}{3.0} \times \frac{8.9}{3.0} = 13.6$$

$$\text{Observed } RR_{OC,SMK} = \frac{39.5}{3.0} = 13.2$$

The very close agreement for the observed rate and that expected under a multiplicative model suggests that the relationship among OC, SMK, and cardiovascular mortality is multiplicative.

c. Both positions can be supported. It is correct that the relative risk for OC users is the same for smokers and nonsmokers, indicating that the data fit a multiplicative model. An analysis stratified by smoking status will show no effect modification of the association between OC and CVD.

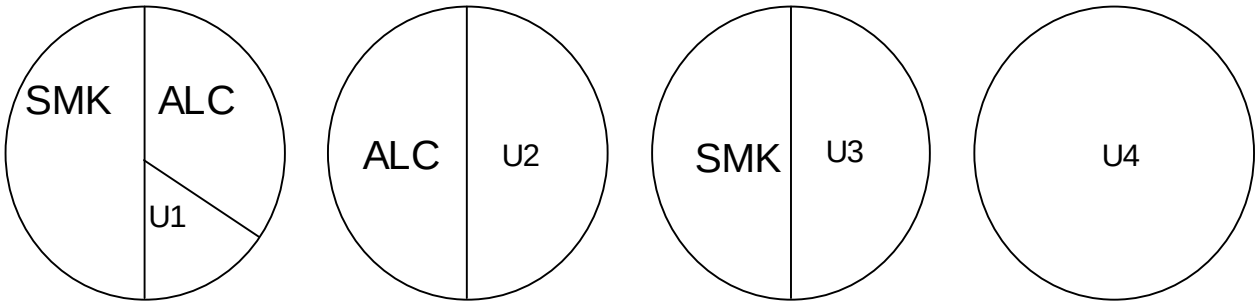
On the other hand, the additive model is more appropriate for assessing public health impact (and individual decision-making). The fact that the joint effect of OC and smoking substantially exceeds the sum of the effects (risk differences) for OC and smoking individually indicates that there the relationship is synergistic in terms of public health

impact. Synergism in this sense implies that if per-person intervention costs are equal, a greater reduction in disease rates will result from focusing on women who both smoke and use OC.

The multiplicative nature of the relationship might suggest that smoking and OC operate on some common element in the pathogenetic process, so that the effects of the one potentiate the effects of the other. However, assessment of biological synergism requires knowledge of biological mechanisms beyond that generally obtainable from epidemiologic data.

4.

a. I II III IV



b. In the Non-drinking group the number of cases would be expected to drop from 40 to 10 (the rate among the non-smokers). The number of cases among the drinkers would be expected to drop from 100 to 15 cases. Thus, (100 - 15) + (40 - 10) = 115 cases would be expected to be prevented through smoking cessation.

c. In the Non-smoking group the number of cases would be expected to drop from 15 to 10 cases, and in the smoking group the number of cases would be expected to drop from 100 to 40 cases. Thus, (15 - 10) + (100 - 40%) = 65 cases could be prevented by abstinence.

d. If both drinking and smoking were eliminated then each cell could be expected to have the same number of cases as in the non-smoking and non-drinking cell. So, (100 - 10) + (40 - 10) + (15 - 10) = 125 cases are prevented by the elimination of smoking and drinking.

e. Ten cases can be attributed to unidentified background factors in pathway IV. For the Smokers-Nondrinkers, 30 cases can be attributed to smoking (pathway III) since 10 of the 40 cases would have occurred in the absence of smoking. Similarly, 5 cases can be attributed

to drinking in the absence of smoking (pathway II). For the Smoker-Drinkers, of the 100 cases, 45 would have been expected to occur either from smoking alone, drinking alone or through unidentified causes (the background rate). Therefore, 55 cases can be attributed to the synergy between smoking and drinking represented by pathway I.

f. In removing either drinking or smoking we prevent not only those cases attributable to the factor alone but also those cases caused by the synergy between the two. Therefore, by removing smoking we prevent 55 of the deaths due to synergy and by removing drinking we prevent the same 55 deaths due to synergy. Of course if we remove both factors we do not prevent the same 55 cases twice. What you have worked through above is an example of non-additivity of attributable risks, which is equivalent to interaction on an additive scale.

5.

a. 100 cases per 100,000 per year.

b. This would be the rate due to smoking + the rate due to drinking + the rate due to unidentified factors. There were 30 cases due to smoking (40 due to the combination of smoking and unidentified factors), 5 due to drinking (15 due to the combination of drinking and unidentified factors), and 10 due to unidentified factors. The expected rate would be: 30 + 5 + 10 = 45 cases per 100,000 people per year. [Note that Rothman's model is based on (or implies) an additive model for combining risks.]

c. EF(Smoking x Drinking) = [(100 - 45) / 100]= .55

d. Since there is such a strong synergism between smoking and drinking health education, physician counseling, and warning labels on both substances should give some attention to the combined effect.

13. Multicausality – analysis approaches

Concepts and methods for analyzing epidemiologic data involving more than two variables; control of confounding through stratified analysis and mathematical modeling.

Multivariable analysis

In the preceding two chapters we delved into the kinds of issues and situations that arise in a multivariable context. We introduced the additive and multiplicative models for joint effects of multiple exposure variables and employed stratified analysis to examine the effects of one variable whle controlling for the values of others. In this chapter we consider analytic approaches for examining the relationships between an outcome and multiple explanatory variables. The latter may consist of a study factor and potential confounders, a study factor and potential modifiers, or several exposures all of which are of interest.

Confounding:

To restate briefly, confounding is a situation where a factor or combination of factors other than the study factor is responsible for at least part of the association we observe for the association between the study factor and the outcome. If we do not control for confounding, then we may misattribute an effect to the study factor when the association really reflects the effect of another variable. In a situation of confounding, the crude data give us the wrong picture of the relationship between the exposure and outcome. Other factors may be exaggerating the strength of the relationship or obscuring some or all of it. To see the correct picture, we need to take into account the effects of other factors.

In a law enforcement analogy, the exposure is the suspect in a bank robbery and the other factors are known offenders with whom he associates. We need to establish the suspect's guilt. The suspect may be completely innocent, may have had some role in the crime, or may have had a greater role than at first appears. In order to determine the suspect's guilt, we need to examine the total picture of the actions of all of the individuals. In this analogy, confounding would occur if we charge the suspect with a crime he did not commit or with a role in the crime greater or smaller than accords with his actions. For example, it would be confounding to charge the suspect with bank robbery if he was just passing by and one of the robbers called him in. Confounding would also occur if we charged the suspect as an accomplice when in fact he was the principal organizer of the robbery.

The most common method of deciding whether or not confounding exists is to compare the crude (uncontrolled) results with the controlled results. If these two sets of results are meaningfully different, if they send a different "message" or suggest a different conclusion about the association under study, then confounding is present; the crude results are "confounded". The conclusion about the presence of confounding, however, is secondary to our main purpose, which is to obtain a valid estimate of the existence and strength of association between the exposure of interest and the disease outcome. When

we determine that confounding is present, we either present the stratum-specific findings or compute an adjusted measure of association (e.g., a standardized rate ratio) that controls for the effects for the confounding variables.

Effect modification

Effect modification is a situation where neither the crude and nor the adjusted measure provides an adequate picture of the relationship under study. The picture is **not wrong**, but it could nevertheless mislead. We may not have fulfilled our responsibility to present the full picture. Effect modification means that there are important differences between groups (or at different levels of some modifying variable) in the relationship between the exposure and the disease on our scale of measurement. Where effect modification is present, then the relationship between exposure and disease is not susceptible to being stated in such a simple formulation as "D and E are associated with a relative risk of about 2". Rather, an answer to the question "what is the relative risk for D given E?", must be "It depends." For example, any discussion of the heart disease risks for women talomg oral contraceptives would be seriously incomplete if it did not explain that the situation is quite different for women who smoke cigarettes or not, especially at ages above 35 years.

Where effect modification is present, the summary measure is an average of disparate components, so that the summary is too uninformative by itself. If Carlos is 90 cm tall, Shizue is 120 cm tall, and Rhonda is 150 cm tall, it may be useful to know that their average height is 120 cm, but probably not a good idea to buy three medium size (120 cm) school uniforms.

Analytic approaches

There are two primary approaches to analyzing data involving more than two variables: stratified analysis and modeling. We have already encountered both. In stratified analysis we divide the observations into one group for each level or combination of levels of the control variables. We analyze the association between the study factor and outcome separately within each group. In this way we may be able to observe the association involving the study factor without interference from the stratification variables.

Comparison of the crude measure of association to the stratum-specific measures or their weighted average will disclose whether the crude measure of association is confounded. Examination of the data within the individual strata will reveal if the measure of association varies so greatly that a summary measure by itself may mislead. For a fuller exploration we can stratify by each of the covariables and by various combinations of them. Stratified analysis gives us a full picture that we can examine in detail.

At some point, however, detail becomes an obstacle instead of an advantage. Modeling is a strategy for submerging the detail and focusing on relationships. Viewing the data through the framework of the model we gain analytic power and convenience. Rather than confuse ourselves and our audience by presenting a plethora of tables, we employ the elegant simiplicity of the model and its parameters, through which we can estimate the measures of association we seek. If we have chosen our model well and evaluated its suitability, we can obtain an optimal analysis of the data. But just as a pilot can fly

very far on instruments but needs to see the runway when landing, a modeling analysis should be supplemented with some stratified analyses. On the other hand, in a stratified analysis, computation of summaries across strata generally involves at least an implicit model framework.

Whichever approaches we use, there's no escaping the fact that how we proceed and how we interpret the results we observe depend on our conceptual model of the relationships among outcome, exposure, and stratification variables. If nothing is known about the factors under study, we may have to proceed in a completely empirical manner. But if there is some knowledge, it will serve as a guide. For example, suppose we see an association involving our study factor and outcome, but when we control for another factor the association disappears. Whether we conclude "confounding" and dismiss the crude association as an artifact or not depends upon whether or not we think of the stratification variable as a "real" cause of the outcome rather than the study factor. If the stratification variable is an intermediate factor in the causal pathway between the study factor and the outcome, then the situation is not one of confounding even though it can be numerically identical.

Stratified analysis — interpretation

Stratified analysis is conceptually simple. It involves disaggregating a dataset into subgroups defined by one or more factors that we want to control. For example, in studying the effect of reserpine use on breast cancer risk, we could stratify by obesity. Analyses within each strata can then be regarded as unconfounded by that risk factor, to the degree that the strata are sufficiently narrow. (If the strata are broad, e.g., "body mass index of 2.2 through 3.2" or "blood pressure greater than 95 mmHg", we may have "residual confounding" due to heterogeneity of the stratification variable within one or more strata.)

We have already encountered stratified analyses, notably in the chapters on confounding and effect modification. In this chapter we will gain a more indepth understanding of stratified analysis and how it relates to other concepts we have learned. We will also see when and how to obtain an overall summary measure that takes account of the stratification.

Example

Suppose that four case-control studies have investigated a possible association between reserpine and breast cancer (a question that arose in the 1970s) and that each controlled for obesity by dividing the data into two strata. The table below shows the crude and stratum-specific odds ratios from these four (hypothetical) studies. How would we describe the results of each study?

Association between reserpine and breast cancer controlling for body weight (odds ratios)
Hypothetical data

Study	Obese	Nonobese	Summary (adjusted)	Total (crude)
A	2.0	2.2	2.1	4.0
B	4.0	2.2	3.1	3.0
C	2.0	2.2	2.1	2.0
D	4.0	2.2	3.1	1.5

In study A, we see that the OR within each body weight category is about 2.0, whereas the crude OR is 4.0. Study A, therefore, illustrates a situation of **confounding**: the crude measure of association lies outside the range of the stratum-specific measures. The crude OR is meaningfully different than the adjusted OR and no other method of adjustment would change that, since any weighted average of the stratum-specific OR's would have to lie between 2.0 and 2.2.

In studies B and C, on the other hand, the crude OR could equal (or nearly equal) a weighted average of the stratum-specific measures (as is in fact the case for the adjusted OR's shown), because it (nearly) lies within the range of those measures. Therefore, confounding is not a feature of the data in either of these studies. In study B, if the numbers of participants in each stratum are large enough for us to regard the difference between the stratum-specific OR's as meaningful (not simply due to "noise"), then the difference indicates effect modification of the OR. It was important for the study to report the stratum-specific OR's and not rely completely on the crude or adjusted measures.

If the strata were large enough and the OR's were regarded as reasonably free of bias, we might wonder whether in some way obesity could potentiate the effect of reserpine (at least on the odds ratio scale). If the relationship is judged to be causal and these OR's the best estimates of the strength of relationship, then the stronger OR for obese patients suggests that they especially should avoid taking reserpine if they cannot lose weight (the usual criterion for "public health interaction" and "individual risk management interaction" are departure from the additive model of expected joint effect. However, if the observed association is "supra-multiplicative" [stronger than that expected from multiplicative model], it will also be "supra-additive" [stronger than expected form an additive model]). In study C, on the other hand, the slight difference between the two strata, even if not attributable to random variation, is insufficient to warrant attention. Any weighted average of the two stratum-specific measures would be a satisfactory summary.

Study D illustrates both confounding and effect modification, since the crude OR lies outside the range of the stratum-specific ORs and therefore could not equal any weighted average of the two. At the same time, the stratum-specific ORs appear to be importantly different (assuming adequate stratum sizes). It would not be sufficient to provide only a summary measure (on the OR scale).

Summarizing the relationships

Often we are interested in obtaining an overall assessment of the role of the study factor, controlling for other risk factors. The usefulness of an overall measure of association will obviously differ in these four studies. In studies A and C, a single overall measure could adequately summarize the OR's in the two strata so that it would not be essential to present them as well. In studies B and D, however, we clearly need to present the stratum-specific OR's, though for some purposes a summary measure may also be useful.

The most convenient overall estimate, if it is not confounded, is the measure based on the aggregate data, the crude estimate. The stratified analysis in study C above indicates no confounding by obesity. If that is the only variable we need to control for, then we can use the crude OR to summarize the relationship.

In both study A and study D, however, confounding is present. Relying on the crude OR as the summary of the stratified results will clearly mislead. Therefore, we require a summary measure that "adjusts for" obesity. The summary measure we derive is a weighted average of the stratum-specific measures. The summary measures we encountered in the chapter on standardization (the SMR and the SRR) are examples of such summary measures.

Relationship between stratified analysis and models for joint effects

The additive and multiplicative models introduced earlier express the joint incidence or effect of two (or more) factors in terms of the separate incidence or effect of each. The multiplicative model, for example, expresses the joint RR as:

$$RR_{11} = RR_{10} \times RR_{01}$$

and the joint risk (or rate) as:

$$R_{11} = \frac{R_{10} \times R_{01}}{R_{00}}$$

where the first and second subscripts indicate presence (1) or absence (0) of the first and second factors, respectively. It turns out that if the data fit this model, then in a stratified analysis controlling for either factor the stratum-specific RR's for the other factor will be equal to each other.

To see this, simply divide both sides of the second form of the model by R_{01}:

$$\frac{R_{11}}{R_{01}} = \frac{R_{10} \times R_{01}}{R_{00} \times R_{01}} = \frac{R_{10}}{R_{00}}$$

Let's examine the term on the left and the term on the right. In both of these terms, the first factor is present in the numerator rate but absent from the denomator rate. Thus, each of these terms is a rate ratio for the effect of the first factor.

$$\begin{array}{c} \text{RR for 1st factor} \\ \text{(2nd factor present)} \end{array} = \begin{array}{c} \text{RR for 1st factor} \\ \text{(2nd factor absent)} \end{array}$$

Meanwhile, the second factor is present in both numerator and denominator rates on the left, and absent from both rates on the right. Since each rate requires a number of cases and a person or person-time denominator, then each RR must come from a 2 x 2 table containing exposed cases, unexposed cases, exposed noncases or person-time, and unexposed noncases or person-time.

Thus, these two RR's correspond to a stratified analysis that controls for the second factor as present vs. absent. Their equality means that the RR for the outcome with respect to the first factor is the same in both strata of the second factor. Had we originally divided by RR_{01}, instead of RR_{10}, we would have found that the RR for the second factor is the same in both strata of the first factor.

To see the relationship with some familiar numbers, here is a portion of the Mann et al. data presented earlier:

Incidence of myocardial infarction (MI) in oral contraceptive (OC) users age 40-44 years, per 100,000 women-years

Cigarettes/day	OC*	$\overline{OC}$*	RR**	AR***
0-14	47 (R_{01})	12 (R_{00})	4	35
15 +	246 (R_{11})	61 (R_{10})	4	185

* Rate per 100,000 women-years

** RR=relative risk (rate ratio)

*** AR=attributable risk (rate difference, absolute difference)

We saw in the chapter on effect modification that the full table conformed quite closely to a multiplicative model. If we look back at the table we see that the RR's for the first two rows (3) were the same and those for the second two rows (4, shown above) were the same.

Suppose we let the four rates in the table be represented by R_{00}, R_{10}, R_{01},and R_{11}, with the first subscript denoting smoking and the second denoting OC. Then we can write:

$$R_{11} = \frac{R_{10} \times R_{01}}{R_{00}}$$

and

$$246 \approx \frac{61 \times 47}{12}$$

The above equality is only approximate, but then the rate ratios weren't exactly the same (3.92 versus 4.03). Therefore, the statement that the RR is the same in all strata is equivalent to saying that the data conform to a multiplicative model.

We could equally well have demonstrated this fact by using the OR (try it!). Had we instead used the rate or risk difference as the parameter of interest, we would find (by subtraction, rather than division) that equality of the stratum-specific difference measures is equivalent to having the data conform to an additive model (try this, too!).

$$R_{11} = R_{10} + R_{01} - R_{00}$$

$$R_{11} - R_{01} = R_{10} + R_{01} - R_{00} - R_{01} = R_{10} - R_{00}$$

This relationship between the multiplicative and additive models on the one hand and stratified analysis on the other is fundamentally trivial, but also fundamental, so it is worth a little more time.

Stratified analysis as "tables" or "columns"

A stratified analysis involving a dichotomous outcome, a dichotomous exposure, and a dichotomous stratification variable involves two 2 x 2 tables, each with two columns of cases and noncases (or person-time). If we look at the data as columns, rather than as tables, we can almost "see" the multiplicative or additive model structure in the stratification. For example, here are two 2 x 2 tables created with hypothetical numbers that produce rates similar to those in the Mann et al. data above and presented in the form of our earlier stratified analyses.

Hypothetical data on incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years, controlling for smoking (after Mann et al.)

Cigarettes /day	15+	15+	0-14	0-14
OC use	OC	$\overline{\text{OC}}$	OC	$\overline{\text{OC}}$
CHD	49	11	19	8
Women-years*	20	18	40	66
Rate**	245	61	48	12
	R_{11}	R_{10}	R_{01}	R_{00}

* (in thousands)

** per 100,000 (some differ slightly from Mann et al.'s)

The lefthand 2 × 2 table shows the relationship between OC and CHD among women who smoke 15+ cigarettes/day; the righthand table shows the relationship among women who smoke less than 15 cigarettes/day. **Equivalently**, the four columns show the number of cases, women-years of risk, and CHD rate in, from left to right:

15+ cigarette/day OC users	(R_{11}, = 49/20,000 = 245/100,000wy)
15+ cigarette/day OC nonusers	(R_{10}, = 11/18,000 = 61/100,000wy)
0-14 cigarette/day OC users	(R_{01}, = 19/40,000 = 48/100,000wy)
0-14 cigarette/day OC nonusers	(R_{00}, = 8/66,000 = 12/100,000wy)

Similarly, all of the relevant RR estimates can be obtained by forming ratios of the appropriate rates, e.g.:

Rate ratios

Both factors (versus neither)	$RR_{11} = R_{11} / R_{00} = 245/12 = 20$
Smoking (1st factor) acting alone	$RR_{10} = R_{10} / R_{00} = 61/12 = 5$
Smoking (1st factor) in presence of OC (2nd factor)	$RR_{S\|O} = R_{11} / R_{01} = 245/48 = 5$
OC (2nd factor) acting alone	$RR_{01} = R_{01} / R_{00} = 48/12 = 4$
OC (2nd factor) in presence of smoking (1st factor)	$RR_{O\|S} = R_{11} / R_{10} = 245/61 = 4$

So the multiplicative model for joint effects, introduced in the chapter on effect modification, is equivalent to stratified analyses in which the ratio measure is the same in all strata. The same can be shown for the additive model and the difference measure, though not with these data since they do not fit an additive model.

"Homogeneity" and "heterogeneity" vs. "synergy" or "antagonism"

In the terminology used when discussing summary measures of association, stratum-specific measures are said to be "homogeneous" when they are the same and "heterogeneous" when they are meaningfully different. Obviously, a summary measure works best in a situation where the measure being summarized is homogenous across strata. In the usual case, for a ratio measure of effect, homogeneity across strata is equivalent to rates, odds, or ratios that conform to a multiplicative model of joint effects. In the case of difference (absolute) measures, homogeneity is equivalent to an additive model of joint effects. "Effect modification" (or "effect measure modification", in Greenland and Rothman's new terminology) signifies heterogeneity for that measure.

Typically, epidemiologic analyses of risk factors employ ratio measures of effect. On the ratio scale, summary measures from stratified analysis (and as we will soon see, from mathematical models) are derived on the premise of homogeneity of effects across strata, equivalent to a multiplicative model of expected joint effects, and also generally inconsistent with an additive model. So the term "effect modification" is most commonly applied to situations where the ratio measure of effect is heterogeneous across strata – even if it should happen (admittedly as the exception) that the data do conform to an additive model! In contrast, "synergism" from a public health perspective is now generally regarded as an observed effect greater than expected from an **additive** model. So when there is "effect modification of the relative risk" there is generally "interaction from a public health perspective".

Such inconsistency is undoubtedly an indication that these concepts were designed by mortals, rather than by a higher power, and also underlines the point that "effect modification" is relative to the scale of measurement or expected model for joint effects. We can hope that as the discipline evolves, a new synthesis will develop that will avoid this "schizophrenic" approach. In the meantime, perhaps the following summary table will help.

Homogeneity, heterogeneity, and effect modification in relation to additive and multiplicative models

	Public health impact perspective	Summary measure perspective
1. Data conform to an additive model (**homogeneity of the difference measure** across strata)	**No interaction** (no synergism)	**No effect modification (of difference measure)**, summary difference measure is adequate **Effect modification (of ratio measure)**, summary ratio measure is not adequate
2. Joint effect exceeds expectation under an **additive** model ("supra-additive" – may or may not equal or exceed multiplicative model)	**Public health interaction (synergistic effect)**	**Effect modification** (of difference measure, perhaps also ratio measure), summary difference measure is not adequate (perhaps also summary ratio measure)
3. Data conform to expectation under a multiplicative model (**homogeneity of ratio measure** across strata)	**Public health interaction (synergistic effect)**	**No effect modification (of ratio measure)**, summary ratio measure is adequate
4. Joint effect exceeds expectation under a **multiplicative** model ("supra-multiplicative")	**Public health interaction (synergistic effect)**	**Effect modification (of difference and ratio measures)**, summary difference and ratio measures are not adequate

Types of overall summary measures

When the crude and stratum-specific measures are all similar, then the crude measure serves as a fully satisfactory summary measure. When there is meaningful heterogeneity, then we will need to present the stratum specific measures themselves. There remains the situation where the stratum-specific measures are sufficiently homogenous that a summary measure of some kind is of interest but, due to confounding, the crude measure cannot serve this roll. In such cases the crude measure is outside the range of the stratum-specific measures or so far from the middle of the range that it would be a

misleading summary. These circumstances call for an adjusted measure, generally some form of weighted average of the stratum-specific measures.

Suppose that all of the stratum-specific measures are close together (i.e., homogeneous), so that we are inclined to regard all of them as estimates of the same population parameter (the "true" measure of association) plus or minus some distortion from sampling variability (if we want to quantify the compatibility of the data with this supposition, we can employ a statistical test, such as the Breslow-Day homogeneity chi-square, to assess the expected range of chance variability). If there is a "true" underlying value, how can we best estimate it? Obviously some sort of weighted average is called for, but what kind?

If there is only one "true" measure of association and each of the strata provides an estimate of that true measure, then we will want to pay more attention to strata that provide "better" (i.e., more precise) estimates. So the averaging procedure we employ should give more weight to the estimates from such strata. We can meet this objective by using as weights the estimated precision of each stratum-specific estimate. Such a weighted average provides the best estimate of the "true" measure of association, under the assumptions on which we have been proceeding. (Rothman refers to summary estimates derived in this way as "directly pooled" estimates. However, the term "pooled" is sometimes used to refer to the crude total over a set of strata or studies.)

[Note: the calculation of summary measures of association as explained below is NOT a required part of EPID 168. The only things from this discussion of summary measures that EPID 168 students are expected to know concern: (1) summary measures are typically weighted averages; (2) if the crude measure of association falls comfortably within the range of the stratum-specific measures, then it is not confounded and may serve as a summary measure; (3) if the crude measure is outside the range of the stratum-specific measures, then confounding is present and the crude measure is <u>not</u> an adjusted measure of association must be used to summarize the relationship; (4) if stratum-specific measures are meaningfully different from each other, then any summary measure (crude or adjusted) provides an incomplete picture of the relationship, so the investigator should report the stratum-specific results and take that heterogeneity into account in interpreting a summary measure. The following discussion is provided for the more advanced or adventurous. Others may wish to come back to this section during or after their next course in epidemiologic methods.]

Precision-based weighted summary measure estimates – *optional topic*

The **im**precision of an estimate can be defined as the width of the confidence interval around it. Since we are used to estimating 95% confidence intervals by adding and subtracting 1.96 times the standard error of the estimate, the total width is 2 x 1.96 × standard error. Since all of these width's will include the 2 x 1.96, all of the variability in precision is contained in the standard errors. The **smaller** the standard error, the greater the degree of precision, so weights consisting of the receiprocals of the standard errors will accomplish precision-weighting. In fact, the weights used are the squares of these reciprocals and are called "inverse variance weights".

Difference measure – the CID

The variance of the CID is an easy one to derive, since the CID is simply a difference of two proportions. When there are at least 5 "successes", the variance of a proportion (p) can be estimated simply as p(1-p)/n, where n is the size of the sample. The variance of a sum or difference of two independent random variables is the sum of their variances. So the variance (square of the standard error) of the CID is:

$$\text{var(CID)} = \text{var}(CI_1) + \text{var}(CI_0)$$

$$[\text{s.e.(CID)}]^2 = \frac{CI_1\,(1\text{-}CI_1)}{n_1} + \frac{CI_0\,(1\text{-}CI_0)}{n_0}$$

Using the notation from our 2 × 2 tables, where "a" represents exposed cases and "b" represents unexposed cases, we can write this formula as:

$$[\text{s.e.(CID)}]^2 = \frac{a/n_1\,(c/n_1)}{n_1} + \frac{b/n_0\,(d/n_0)}{n_0}$$

$$[\text{s.e.(CID)}]^2 = \frac{ac}{n_1^3} + \frac{bd}{n_0^3} = \frac{n_0^3 ac + n_1^3 bd}{n_1^3\, n_0^3}$$

whose reciprocal (and the stratum-specific weight) is:

$$w = \frac{1}{[\text{s.e.(CID)}]^2} = \frac{n_1^3 + n_0^3}{n_0^3 ac + n_1^3 bd}$$

This value is computed for each stratum and used as the weight for the CID for that stratum. For two strata (indicated by subscripts 1 and 2):

$$\text{Summary CID} = \frac{w_1\,CID_1 + w_2\,CID_2}{w_1 + w_2}$$

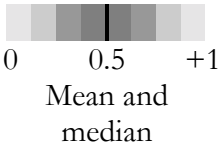
Since we have just derived the variances of the stratum-specific CID estimates and since the variance of the summary CID estimate is simply their sum, the variance of this summary CID estimate is simply $1/w_1 + 1/w_2$, and a 95% confidence interval for the summary CID estimate is:

$$\text{95\% CI for (summary) CID} = \text{CID} \pm 1.96 \sqrt{1/w_1 + 1/w_2}$$

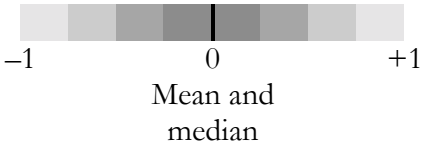
Ratio measures

A uniform random variable that is a proportion has a symmetric distribution, since its possible values lie between 0 and 1, and the mean of the distribution (0.5) is the same as its median. Similarly, the distribution of the CID, based on the difference in two uniform random proportions, is symmetric, since it lies between -1 and 1 and has its mean and median at its null value, 0.

Distribution of a proportion:



Distribution of a difference of two proportions:

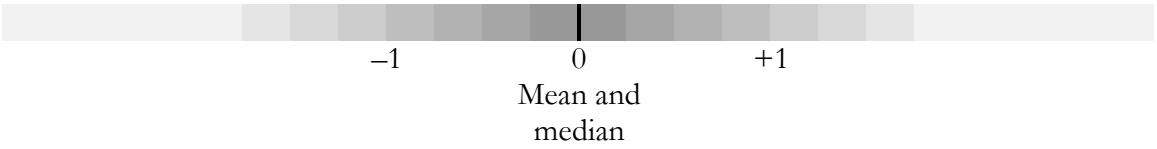


Because of this symmetry, variance estimates based on an approximate normal distribution could be used. Ratio measures, however, do not have symmetric distributions. The CIR (a ratio of two proportions) and the OR (a ratio of odds, which are in turn ratios of two non-independent proportions) both have a lower limit of 0, a median (and null value) at 1.0, and no upper limit.

Distribution of CIR, IDR, OR

This asymmetry makes the use of a normal approximation more problematic. However, the logarithm of a ratio measure **does** have a symmetric distribution, so that the normal approximation can be used.

Distribution of ln(CIR), ln(IDR), ln(OR):



Therefore, variances for the CIR, IDR, and OR are estimated using a logarithmic transformation.

Ratio measures – CIR:

The natural logarithm of the CIR is:

$$\ln(CIR) = \ln\left[\frac{CI_1}{CI_0}\right] = \ln(CI_1) - \ln(CI_0)$$

If each stratum-specific CI is an independent random proportion, then the variance of the logarithm of the estimate of the stratum-specific CIR is the sum of the variances of the logarithms of the estimates of the statum-specific CI's.

$$\mathrm{Var}(\ln(CIR)) = \mathrm{Var}(\ln(CI_1)) + \mathrm{Var}(\ln(CI_0))$$

The variance of these logarithms is obtained using a Taylor's series approximation as (Kleinbaum, Kupper, and Morgenstern; Rothman and Greenland):

$$\text{Var}(\ln(\text{CIR})) \approx \frac{c}{an_1} + \frac{d}{bn_0} = \frac{bcn_0 + adn_1}{abn_1n_0}$$

so that the stratum-specific weights are:

$$w = \frac{1}{\text{Var}(\ln(\text{CIR}))} = \frac{abn_1n_0}{adn_1 + bcn_0}$$

For two strata, then, the precision-weighted summary **ln(CIR)** is:

$$\text{Summary } (\ln(\text{CIR})) = \frac{w_1 \ln(\text{CIR}_1) + w_2 \ln(\text{CIR}_2)}{w_1 + w_2}$$

In order to obtain the summary estimate for the CIR, the summary ln(CIR) must now be converted to the natural scale by exponentiation:

$$\text{Summary CIR} = \exp(\text{summary } \ln(\text{CIR}))$$

Again, we can use the w_i to obtain the variance of the overall CIR estimate, though again a transformation of scale will be needed. The variance of the summary ln(CIR) estimate is simply $1/w_1 + 1/w_2$, so the 95% confidence interval is:

$$95\% \text{ confidence interval for } \ln(\text{CIR}) = \ln(\text{CIR}) \pm 1.96\sqrt{1/w_1 + 1/w_2}$$

$$95\% \text{ confidence interval for CIR} = \exp[\ln(\text{CIR}) \pm 1.96\sqrt{1/w_1 + 1/w_2}]$$

Ratio measures – OR:

An approximate variance estimate for the ln(OR) in the ith stratum is:

$$\text{Var}(\ln(\text{OR})) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

so that the weight for the ith stratum is:

$$w_i = \frac{1}{\left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)}$$

(Notice that a small number in any cell makes the variance large and, therefore, the weight small.) The overall ln(OR) is then estimated as:

$$\ln(OR) = \frac{w_1 OR_1 + w_2 OR_2}{w_1 + w_2}$$

and the overall OR as:

$$OR = \exp(\ln(OR))$$

The variance of the ln(OR) is $1/\Sigma w_i$ and can be used to obtain a 95% confidence interval for the ln(OR), which can then be exponentiated to obtain a confidence interval for the OR, as for the CIR.

Mantel-Haenszel summary measures:

Nathan Mantel and William Haenszel, in their classic 1959 paper, introduced a summary OR that is particularly easy to calculate:

$$OR_{MH} = \frac{\Sigma[a_i d_i / n_i]}{\Sigma[b_i c_i / n_i]}$$

Rothman shows that the OR_{MH} is a weighted average, with stratum-specific weights of $b_i c_i / n_i$. These weights are also precision-based, since they are inversely proportional to the variance of the logarithm of the stratum-specific OR's. The difference between these weights and the ones in the previous formula is that for the OR_{MH} the weights are based on variances that apply on the assumption that the OR's are 1.0, whereas the previous weights did not require that assumption. However, the two summary measures produce similar results and are essentially equivalent when the stratum-specific OR's are not far from 1.0. An advantage of the OR_{MH} is that it can be used with sparse data including an "occasional" zero cell (see Rothman).

Formulas for the these and other summary measures of association (IDD, IDR), confidence intervals, and overall tests of statistical significance can be found in the textbooks by

Kleinbaum, Kupper, and Morgenstern; Hennekins and Buring; Schlesselman; and Rothman. The Rothman text includes discussion of maximum likelihood methods of estimating summary measures.

Although the discussion here has emphasized the usefulness of summary measures in analyses where there is little heterogeneity across strata, at times an investigator may wish to present a summary measure even when substantial heterogeneity is present. Standardized (rather than adjusted) measures are used in these situations (see Rothman and/or Kleinbaum, Kupper, and Morgenstern).

[Note: Time to tune back in if you skipped through the section on weighting schemes for summary measures of association. On the other hand, if you are already familiar with mathematical models you may wish to skim or skip this section.]

Matched designs

As we saw in the chapter on confounding, when the study design uses matching, it may be necessary to control for the matching variables in the analysis. In a follow-up study, analyzing the data without taking account of matching may not yield the most precise estimates, but the estimates will not be biased. A case-control study with matched controls, however, can yield biased estimates if the matching is not allowed for in the analysis. Thus, the matching variables should always be controlled in analyzing matched case-control data. If the result is no different from that in the unmatched analysis, then the unmatched analysis can be used, for simplicity.

The most straightforward way to control for matching variables is through stratified analysis, as presented above. If matching was by category (i.e., frequency matching, e.g., by sex and age group) was employed, then the analysis procedure is a stratified analysis controlling for those variables. If individual matching (e.g., pair matching, matched triples, etc.) was employed, then each pair or "n-tuple" is treated as a strata.

Suppose that the data from a case-control study using pair matching are as shown in the following table.

Pair	Case	Control	Type
6	n	n	A
9	n	n	A
10	n	n	A
1	Y	n	B
2	Y	n	B

5	Y	n	B
3	n	Y	C
8	n	Y	C
4	Y	Y	D
7	Y	Y	D

If each pair is a stratum, then the stratified analysis of the above data consists 10 tables, each with one case and one control. There will be 3 tables like table A, 3 like table B, 2 like table C, and 2 like table D.

	Exp	Unexp	Exp	Unexp	Exp	Unexp	Exp	Unexp
Case	0	1	1	0	0	1	1	0
Control	0	1	0	10	1	0	1	0
Type	A		B		C		D	

Although we cannot compute any stratum-specific measures of association, we can compute a Mantel-Haenszel summary odds ratio using the formula:

$$OR_{MH} = \frac{\Sigma[a_i d_i / n_i]}{\Sigma[b_i c_i / n_i]}$$

where a_i, b_i, c_i, d_i are the cells in table i, and n_i is the number of participants in table i. This general formula becomes much simpler for pair-matched data, because all of the n_i are 2 and many of the terms disappear due to zero cells. When we remove these terms and multiply numerator and denominator by 2 (n_i), we are left with (a) a one ($a_i d_i$) in the numerator for each table where the control is exposed and the case is not (table type B); and (b) a one ($b_i c_i$) in the denominator for each table where the case is exposed and the control is not (table type C). For the above data:

$$OR_{MH} = \frac{1 + 1 + 1}{1 + 1} = \frac{3}{2} = 1.5$$

So the formula becomes simply OR=B/C, where B is the number of discordant pairs in which the case is exposed and C is the number of pairs in which the control is exposed. Note that the concordant pairs (types A and D) have no effect on the OR.

Mathematical models

Earlier in this chapter we showed that when the RR is the same in all strata of a stratified analysis, then data conform to a multiplicative model, and vice-versa. We also stated that for difference measures, equality of the stratum-specific difference measures is equivalent to having the data conform to an additive model. In fact, these simple models can serve as a jumping off point for understanding mathematical models used to control confounding.

Returning to the topic of breast cancer in relation to obesity and/or reserpine use, suppose that the following table shows data from a cohort study. (Note that this is hypothetical - reserpine was at one time suspected of being related to breast cancer risk, but that evidence has since been discounted.)

Ten-year risk of breast cancer, by obesity and use of reserpine (hypothetical data)

Risk factors	Numeric (illustrative)	Algebraic
None (background risk)	.01	R_{00}
Obesity only	.03	R_{10}
Reserpine only	.02	R_{01}
Both reserpine and obesity	.04	R_{11}

Thus:

R_{00} indicates background risk (no reserpine, non-obese)

R_{10} indicates risk for obesity (without reserpine)

R_{01} indicates risk for reserpine (without obesity)

R_{11} indicates risk both reserpine and obesity

In this example, the joint risk conforms to an additive model:

$RD_{11} = RD_{10} + RD_{01}$ (Risk differences are additive)

$$R_{11} - R_{00} = (R_{10} - R_{00}) + (R_{01} - R_{00})$$

$$(.04 - .01) = (.03 - .01) + (.02 - .01)$$

$$0.03 = 0.02 + 0.01$$

or, equivalently:

$$R_{11} = R_{10} + R_{01} - R_{00}$$

$$0.04 = 0.03 + 0.02 - 0.01$$

We can also express the various risks in terms of the baseline risk and the "effect" of the risk factors:

$R_{10} = R_{00} + RD_{10}$ (.03 = .01 + .02) (Obesity "effect")

$R_{01} = R_{00} + RD_{01}$ (.02 = .01 + .01) (Reserpine "effect")

$R_{11} = R_{00} + RD_{01} + RD_{10}$ (.04 = .01 + .02 + .01) (Both)

Note that the word "effect" is used here by convention and for convenience, rather than to suggest causality.

Another way we might think about these various risk equations is to try to put them all into a single equation with "switches" for which effects are "turned on". The baseline risk R_{00} is always present, so we require only two "switches", one for the obesity effect and one for the reserpine effect:

Risk	=	R_{00}	+	Obesity effect		Obesity "switch"	+	Reserpine effect		Reserpine "switch"
Risk	=	R_{00}	+	RD_{10}	×		+	0.01	×	
Risk	=	0.01	+	0.02	×		+	0.01	×	

When a "switch" is on (=1) then the 0.02 (obesity effect) or 0.01 (reserpine effect) comes into play, making the Risk from the model larger.

Risk	=	R_{00}	+	Obesity effect		Obesity "switch"		Reserpine effect		Reserpine "switch"	
Risk	=	0.01	+	0.02	×	0	+	0.01	×	0	= 0.1
Risk	=	0.01	+	0.02	×	1	+	0.01	×	0	= 0.03
Risk	=	0.01	+	0.02	×	0	+	0.01	×	1	= 0.02
Risk	=	0.01	+	0.02	×	1	+	0.01	×	1	= 0.04

We now have a "model" that we can use to compute the risk for any combination of the two risk factors. Although this example is trivial, as well as contrived, the model structure is the same as in multiple linear regression. To see our model in a more sophisticated form, we have merely to replace the "switches" by indicator variables that can take the value of 0 or 1.

Linear models:

If we let:

B = 1 if the woman is obese and 0 if she is not

E = 1 if the woman uses reserpine and 0 if she does not

then our model becomes:

$$R(B,E) = R_{00} + (RD_{10})B + (RD_{01})E$$

Substituting values from the table:

$$R(B,E) = .01 + (0.02)B + (0.01)E$$

Our two dichotomous variables (B=1 or 0, E=1 or 0) yield four possible combinations of reserpine use and obesity, just as did our switches model. We now have a professional-looking linear model for breast cancer risk in terms of baseline risk, presence or absence of each of two dichotomous risk factors, and the risk difference (or increase in risk) attributable to each factor. The risk differences (0.02, 0.01) are called "coefficients" and are often represented by the Greek letter β; the baseline risk is often represented by the Greek letter α.

You may well wonder what is the value of the above machinations, since we have no more information from our model than we had in our table of risks (i.e., in our stratified analysis). The accomplishment lies in the ability to estimate risk differences for each factor, controlling for the other(s), by estimating the coefficients in the model. The power of modeling is the ability to use the study data to estimate model coefficients by using a statistical technique known as regression analysis. The estimated coefficients yield epidemiologic measures that are adjusted for the effects of the other variables in the model.

We can make our model more complex and professional-looking by adding a third variable and introducing additional notation:

$$\text{Risk} = \Pr(D=1 \mid X_1, X_2, X_3) = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Here, we express risk as the probability that the disease variable equals 1 (as opposed to 0) based on the values of X_1, X_2, X_3. Each β represents the risk difference, or increase in risk, for the corresponding factor (X). The estimate of each β would be based on the observed risk difference across each stratum of the other variables.

This model, of course, is just like the one we developed, except that to make it more impressive, α's and β's are used instead of RD's, X's are used instead of more familiar letters, and a third term has been added. For example, if X_1 is obesity, X_2 reserpine, and X_3 parity (also coded as a dichotomous variable, e.g., nulliparous vs. parous) then the coefficient for X_1 will be a weighted average of the risk difference for obesity use among the four subgroups defined by the other two risk factors:

1. no reserpine-nulliparous women
2. no reserpine-parous women
3. reserpine-nulliparous women
4. reserpine-parous women.

Therefore, each coefficient (risk difference) will be adjusted for the effects of the other variables in the model, more or less as if we had computed an adjusted overall measure in a stratified analysis.

Just as in stratified analysis, the suitability of the coefficient as an adjusted risk difference depends on whether the risk difference for reserpine is essentially the same across the four groups. The model is designed to handle random variability in the risk differences, but not biological (or sociological, artefactual, etc.) reality. So as with any summary measure, the suitability of the linear regression coefficient (i.e., the estimate of the overall risk difference) can be compromised by meaningful heterogeneity of the risk difference across strata of the other variables (i.e., on the extent of statistical interaction or effect modification of the risk difference).

If necessary, the model can accomodate some heterogeneity with the help of an "interaction" term to represent the "difference in risk differences". Interaction terms are usually created as a product of the

two (or more) factors that "interact", since such a term is zero if either of the factors is absent and one only when both are present. For the price of one more Greek letter (γ , gamma) we can write the model:

$$\text{Risk} = \Pr(D=1 \mid X_1,X_2,X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 X_1 X_2$$

provides for the effect of X_1 to depend upon whether X_2 is present or absent (as well as for the effect of X_2 to depend upon whether X_1 is present or absent). But if we incorporate interaction terms for all possible pairs, triplets, . . ., of variables, we will find ourselves right back where we started from – a fully-stratified analysis and no summary measure to use.

The linear model we have just seen has many attractive features, not unimportantly its simplicity and the ease with which statistical estimation of its coefficients can be carried out. Moreover, although we have developed and illustrated the model using only dichotomous, or "binary" variables, the model can readily accomodate count and continuous variables, and with some caution, ordinal variables. (For a nondichotomous variable, the coefficient is the risk difference for a one-unit increase in the variable.)

But linear models also have several drawbacks. First, of course, the data may not conform to an additive model, perhaps to an extent beyond which a single interaction term will suffice to "fit" the data. Second, it is possible to obtain estimates of coefficients that will result in "risks" that are less than zero or greater than one. The linear model in the homework assignment will do that for certain combinations of risk factors, though this is more of a technical objection. Third, linear regression estimates risk differences, but epidemiologists are usually interested in estimating ratio measures of association.

Logistic models:

More widely used in epidemiologic analysis is the logistic model (also referred to as the multiple logistic model or the logit analysis model). In our linear model, above, we chose to model risk as a linear function of two risk factors. In the logistic model, we model the "logit" as a linear function of the risk factors:

$$\text{Logit}(D=1 \mid X_1,X_2,X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The logit is the natural logarithm of the odds, ln(odds) or ln[p/(l-p)]. It may seem a bit farfetched to work with the logit, rather than risk, but recall our explanation for the use of a logarithmic transformation in order to estimate the variance of a ratio measure.

Whereas risk ranges from 0 to 1, a confining situation for mathematicians, the logit has no bounds. Whereas the risk ratio and the OR have their null value (1.0) way to one side of the range of possible values (zero to infinity), the log(OR) has an unlimited range, with its null value (zero) right in the middle (i.e., it has a symmetrical distribution). We generally use Naperian or "natural" logarithms (base e), abbreviated as ln.

Moreover, the logistic model, we will see, corresponds to a multiplicative model, which we saw earlier is the model that is implied by stratified analysis based on the OR or the risk ratio. Furthermore, the coefficients that we estimate using logistic regression can be converted into OR's, so that we now have a ratio measure of association.

It is easy to discover what the logistic coefficients are. Since the logit is the logarithm of the odds, then the difference of two logits is the logarithm of an OR (because subtraction of logs corresponds to division of their arguments – see the appendix to the chapter on Measures of Frequency and Extent).

Suppose that X_3 is a dichotomous (0-1) variable indicating absence (0) or presence (1) of an exposure. First write the model with the exposure "present" (X_3=1), and underneath write the model with the exposure "absent" (X_3=0).

$$\begin{aligned} &\text{logit}(D=1\,|\,X_1,X_2,X_3=1) \;=\; \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \quad (X_3 = 1, \text{ present}) \\ -\;&\text{logit}(D=1\,|\,X_1,X_2,X_3=0) \;=\; \alpha + \beta_1 X_1 + \beta_2 X_2 + 0 \quad\;\; (X_3 = 0, \text{ absent}) \\ &\text{———————————————————————————} \end{aligned}$$

When we subtract the second model from the first, all the terms on the right are removed except the coefficient for X_3. On the left, we have the (rather messy) difference of the two logits, one for X_3 present and the other for X_3 absent:

$$\text{logit}(D=1\,|\,X_1,X_2,X_3=1) - \text{logit}(D=1\,|\,X_1,X_2,X_3=0) = \beta_3$$

Spelling out the logits:

$$\ln(\text{odds}(D=1\,|\,X_1,X_2,X_3=1)) - \ln(\text{odds}(D=1\,|\,X_1,X_2,X_3=0)) = \beta_3$$

and, since a difference of logarithms is the logarithm of a ratio:

$$\ln\left[\frac{\text{odds}(D=1\,|\,X_1,X_2,X_3=1)}{\text{odds}(D=1\,|\,X_1,X_2,X_3=0)}\right] = \beta_3$$

A ratio of odds is simply an OR, in this case, the OR for the disease with respect to the exposure represented by X_3:

$$\ln[\,\text{OR}\,] = \beta_3$$

$$\exp(\ln[\,\text{OR}\,]) = \exp(\beta_3)$$

$$OR = \exp(\beta_3)$$

β_3 is the difference of the logits, hence the log of the OR for the exposure represented by X_3. Therefore $\exp(\beta_3)$ is the OR for a one-unit change in X_3.

Note: $\exp(\beta_1)$ means the anti-logarithm: e, the base for Naperian logarithms, raised to the β_1 power. Since the coefficients are on the logarithmic scale, to see the result on the OR scale, we needed to take the anti-logarithm. For example, a logistic model coefficient of 0.7 corresponds to an OR of about 2.0 for a dichotomous variable or 2.0 for a one-unit increase in a measurement variable.

So the coefficient of a dichotomous explanatory variable is the log of the OR of the outcome with respect to that explanatory variable, controlling for the other terms included in the model. The constant term (α) in a model with only dichotomous risk factor variables is the baseline logit (log odds) for the outcome – the log of the disease odds for a person who has none of the risk factors ($\ln[\Pr(CI_0/(1-CI_0)]$).

For a nondichotomous risk factor, we can compare the odds at two different levels. For example, if age is expressed by a continuous variable X_1 for the number of years, then $\exp(\beta_1)$ gives the OR per year of age and $\exp(10\ \beta_1)$ gives the OR per decade of age.

The logistic model can also be written in terms of risk (i.e., probability) by taking anti-logs (exponents) and employing some algebra. The tranformation is left as an optional exercise for those of you who are interested. The result is:

$$\Pr(D=1 \mid X_1,X_2,X_3) = \frac{1}{1 + \exp(-\alpha - \beta_1X_1 - \beta_2X_2 - \beta_3X_3)}$$

or, if we let L = logit = $\alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$

$$\Pr(D=1 \mid X_1,X_2,X_3) = \frac{1}{1 + \exp(-L)}$$

From the risk formulation we can readily see that the logistic function must range between zero and one, a desirable property for modeling risk. When L (the logit) is "infinitely negative", then exp(-L) is "infinitely large" and the probability estimate is zero. When L is "infinitely large", then exp(-L) is also "infinitely small" and the probability estimate is one. When L is zero, then exp(-L) is 1, and the probability estimate is one-half.

Key epidemiologic assumptions in the logistic model

1. the log odds of disease are linearly related to each of the risk factors (X variables), or equivalently, the disease odds are exponentially related to each of the risk factors, or equivalently, the disease risk is related to each of the risk factors by the logistic (sigmoidal) curve;

2. the joint effects of the risk factors are multiplicative on disease odds (e.g., if a one-unit increase in X_1 alone multiplies incidence odds two-fold and a one-unit increase in X_2 alone multiplies incidence odds three-fold, then a simultaneous one-unit increase in both X_1 and X_2 multiplies incidence odds six-fold) (Greenland, *AJPH*, 1989; Rothman, Modern epidemiology).

In addition, to estimate the coefficients using regression procedures, it must be assumed that the subjects are a random sample of independent observations from the population about which inferences are to be drawn (Harrell, Lee, and Pollock, 1988).

Thus the logistic model corresponds to the multiplicative model for the stratified analysis we considered above. The true OR is assumed constant across all strata. As with the linear model, it is the assumption of homogeneity that permits us to estimate coefficients that are simple to interpret.

We can relax the assumption by including product terms, as illustrated above for the linear model. But then the coefficients are more difficult to interpret. In addition, carried too far that tactic will return us toward a fully-stratified situation and will exhaust our sample size, computer resources, and imagination.

Though we have illustrated both of these models with dichotomous (zero-one) variables, they can readily accomodate continuous variables. Again, the model structure is based on an assumption – that the relationship of the dependent variable (risk, for the linear model, or the logit, for the logistic model) with the independent variable is linear.

For some relationships, this assumption is readily tenable, e.g., CHD risk and number of cigarettes smoked. For others, e.g., mortality risk and body weight, the relationship is U-shaped, so that a simple linear or logistic model will not be suitable (more complex forms of the linear and logistic models are available for U-shaped variables through such techniques as the incorporation of squares of variable values).

Other limitations of the logistic model are that ORs are not the preferred epidemiologic measure of association, and where the outcome is not rare, the proximity of the OR to the risk ratio does not hold. Also, the model cannot provide what the study cannot. Although the logistic model in the above form can be used with case control data, estimates of risk require follow-up data. Mathematics can substitute for data only to a point.

Other regression models *[Optional for EPID 168]*

Two other mathematical model forms that epidemiologists commonly use to control for confounding and to obtain adjusted measures of effects are the proportional hazards and Poisson models.

For an outcome with an extended risk period, especially an outcome that is not rare, it is frequently desirable to use an analysis approach, such as incidence density or survivorship, that takes into account time to the occurrence of the event. The proportional hazards model, developed by David R. Cox, is a widely-used mathematical model for analyzing epidemiologic data where "time to occurrence" is important. The "hazard" (conventionally represented by the Greek letter lambda, λ) is essentially the same concept as instantaneous incidence density.

For three independent variables, the proportional hazards model can be written:

$$\log[ID(t|X_1,X_2,X_3)] = \log[ID_0(t)] + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$$

(i.e., the natural log of incidence density as a function of time is the sum of the log of a background or underlying incidence density plus an increment for each predictor variable).

The model can also be formulated in terms of survivorship:

$$S(t|X_1,X_2,X_3) = [S_0(t)]\exp(\beta_1X_1 + \beta_2X_2 + \beta_3X_3)$$

where S(t) is the probability that the event has not occurred by time t.

The coefficient of a dichotomous predictor is the logarithm of the incidence density ratio [ln(IDR)] for that predictor:

$$\log[ID(t|X_1,X_2,X_3=1)] = \log[ID_0(t)] + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 \quad (X_3 \text{ present})$$

$$-\ \log[ID(t|X_1,X_2,X_3=0)] = \log[ID_0(t)] + \beta_1X_1 + \beta_2X_2 + 0 \quad (X_3 \text{ absent})$$

$$\log[IDR(t)] = \beta_3$$

$$IDR(t) = \exp(\beta_3)$$

In addition to the assumptions required for the logistic model, the Cox proportional hazards model requires that the hazard ratio (the IDR) be constant over time, though more complex survivorship models employing "time-dependent covariates" relax this assumption.

The Poisson model is similar to the logistic model and the proportional hazards model in that the three involve a logarithmic transformation of the risk function (i.e., odds, hazard) being estimated and have a linear combination (i.e., an expression of the form:

$a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots)$ on the right-hand side. The Poisson model is of particular interest when outcomes are very rare.

Key points [EPID 168 students please tune back in here.]

Some guiding principles for multivariable analysis are:

1. Keep in mind that our principal objectives are to describe and interpret the data at hand, using informed judgment, insight, and substantive knowledge as well as technique.
2. Stratified analysis is a very powerful approach. Although it does not hold when we try to analyze many variables simultaneously, we can control for two or three at a time, using different subsets, and let judgment help to fill the gaps. It is always possible that an observed association that is not eliminated when we control for smoking, cholesterol, blood pressure, and Type A behavior pattern individually could still be due to some combined effect of all of these. But how likely is it, especially if we have controlled for each pair of these risk factors and still found the association?
3. In carrying out a stratified analysis for a variable or a combination of variables, we are asking the question "is that combination of variables responsible for the observed result?" The question must be a reasonable one for us to ask. If a few principal risk factors individually do not account for an observed finding, the probability that some combination of them would do so appears less likely. [But no one has demonstrated that proposition empirically.]
4. Mathematical modeling is a very powerful approach to data analysis. But in all cases, a key question is whether the form of the model is appropriate for the data, and the underlying relationships, at hand. Using an inappropriate model can produce biased results. There are statistical techniques for assessing the statistical appropriateness of the models employed ("ask your statistician").

(It is recommended (see Greenland, *AJPH*, 1989; 79(3):340-349 and Vanderbroucke JP: Should we abandon statistical modeling altogether? *Am J Epidemiol* 1987; 126:10-13) that before embarking on modeling exercises that cannot be directly validated against the results of stratified analyses, one should first perform parallel analyses with the same variables in order to validate model choices and results against the stratified data.)

Expectations for EPID 168

- Know the relationship between the multiplicative model and stratified analysis, and (only) basic concepts of linear regression models and logistic regression models. Expectations for your understanding of mathematical modeling are modest:
- Know advantages and disadvantages of modeling (compared to, for example, stratified analysis), as presented in the chapter on confounding.

- Know the epidemiologic meaning of the coefficient of an exposure term in a linear regression model and how the linear regression model relates to stratified analysis and the additive model discussed in the Effect Modification chapter.
- Know the epidemiologic meaning of the coefficient of an exposure term in a logistic model and how that model relates to stratified analysis and the multiplicative model.
- Know the epidemiologic meaning of the coefficient of an exposure term in a proportional hazards model and that that model is used for analyses in terms of incidence density [survivorship]
- For all three models, the coefficients in a model with several variables are all "adjusted" for the effects of the other variables in the model.

Bibliography

Rothman, Modern epidemiology, pp. 285-295; Schlesselman, Case-control studies, pp. 227-234.

Harrell, Frank E., Jr.; Kerry L. Lee, Barbara G. Pollock. Regression models in clinical studies. *JNCI* 1988; 80(15):1198-1202.

Godfrey, Katherine. Simple linear regression in medical research. N Engl J Med 1985;313:1629-36.

Silberberg, Jonathan A. Estimating the benefits of cholesterol lowering: are risk factors for coronary heart disease multiplicative. *J Clin Epidemiol* 1990; 43(9):875-879.

J. Paul Leigh. Assessing the importance of an independent variable in multiple regression: is stepwise unwise? *J Clin Epidemiol* 1988; 41:669-678.

Vandenbroucke, Jan P. Should we abandon statistical modeling altogether? *Am J Epidemiol* 1987; 126:10-13.

Greenland, Sander. Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 1989; 79:340-349 (Advanced)

Kleinbaum, Kupper, and Morgenstern. *Epidemiologic research: Principles and quantitative methods*. Chapters 16-17.

Breslow and Day. *Statistical methods in cancer research. I. The analysis of case-control studies.* Chapters 3-7 (Primarily Chapter 3).

Wilcosky, Timothy C. and Lloyd E. Chambless. A comparison of direct adjustment and regression adjustment of epidemiologic measures. *J Chron Dis* 1985; 38:849-356.

See also: Flanders, W. Dana; and Philip H. Rhodes. Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *J Chron Dis* 1987; 40(7):697-704. [includes SAS program]

Deubner, David C., William E. Wilkinson, Michael J. Helms, et al. Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. *Am J Epidemiol* 1980;112:135-143, 1980.

Weinstein, Milton C.; Pamela G. Coxson, Lawrence W. Williams, Theodore M. Pass, et al. Forecasting coronary heart disease incidence, mortality, and cost: the Coronary Heart Disease Policy Model. *Am J Pubic Health* 1987; 77:1417-1426.

McGee, Daniel; Dwayne Reed, Katsuhika Yano. The results of logistic analyses when the variables are highly correlated: an empirical example using diet and CHD incidence. *J Chron Dis* 1984; 37:713-719.

Breslow and Storer. General relative risk functions for case-control studies. *Am J Epidemiol* 1985;

Szklo, Moyses; F. Javier Nieto. Epidemiology: beyond the basics. Gaithersburg MD, Aspen, 2000. Chapter 7 has an excellent presentation of stratified analysis and mathematical modeling at a basic and understandable level.

14. Data analysis and interpretation

Concepts and techniques for managing, editing, analyzing and interpreting data from epidemiologic studies.

Key concepts/expectations

This chapter contains a great deal of material and goes beyond what you are expected to learn for this course (i.e., for examination questions). However, statistical issues pervade epidemiologic studies, and you may find some of the material that follows of use as you read the literature. So if you find that you are getting lost and begin to wonder what points you *are* expected to learn, please refer to the following list of concepts we expect you to know:

- Need to edit data before serious analysis and to catch errors as soon as possible.
- Options for data cleaning – range checks, consistency checks – and what these can (and can not) accomplish.
- What is meant by data coding and why is it carried out.
- Basic meaning of various terms used to characterize the mathematical attributes of different kinds of variables, i.e., nominal, dichotomous, categorical, ordinal, measurement, count, discrete, interval, ratio, continuous. Be able to recognize examples of different kinds of variables and advantages/disadvantages of treating them in different ways.
- What is meant by a "derived" variable and different types of derived variables.
- Objectives of statistical hypothesis tests ("significance" tests), the meaning of the outcomes from such tests, and how to interpret a p-value.
- What is a confidence interval and how it can be interpreted.
- Concepts of Type I error, Type II error, significance level, confidence level, statistical "power", statistical precision, and the relationship among these concepts and sample size.

Computation of p-values, confidence intervals, power, or sample size will not be asked for on exams. Fisher's exact test, asymptomatic tests, z-tables, 1-sided vs. 2-sided tests, intracluster correlation, Bayesian versus frequentist approaches, meta-analysis, and interpretation of multiple significance tests are all purely for your edification and enjoyment, as far as EPID 168 is concerned, not for examinations. In general, I encourage a nondogmatic approach to statistics (*caveat*: I am not a "licensed" statistician!).

Data analysis and interpretation

Epidemiologists often find data analysis the most enjoyable part of carrying out an epidemiologic study, since after all of the hard work and waiting they get the chance to find out the answers. If the data do not provide answers, that presents yet another opportunity for creativity! So analyzing the data and interpreting the results are the "reward" for the work of collecting the data.

Data do not, however, "speak or themselves". They reveal what the analyst can detect. So when the new investigator, attempting to collect this reward, finds him/herself alone with the dataset and no idea how to proceed, the feeling may be one more of anxiety than of eager anticipation. As with most other aspects of a study, analysis and interpretation of the study should relate to the study objectives and research questions. One often-helpful strategy is to begin by imagining or even outlining the manuscript(s) to be written from the data.

The usual analysis approach is to begin with descriptive analyses, to explore and gain a "feel" for the data. The analyst then turns to address specific questions from the study aims or hypotheses, from findings and questions from studies reported in the literature, and from patterns suggested by the descriptive analyses. Before analysis begins in earnest, though, a considerable amount of preparatory work must usually be carried out.

Analysis - major objectives

1. Evaluate and enhance data quality
2. Describe the study population and its relationship to some presumed source (account for all in-scope potential subjects; compare the available study population with the target population)
3. Assess potential for bias (e.g., nonresponse, refusal, and attrition, comparison groups)
4. Estimate measures of frequency and extent (prevalence, incidence, means, medians)
5. Estimate measures of strength of association or effect
6. Assess the degree of uncertainty from random noise ("chance")
7. Control and examine effects of other relevant factors
8. Seek further insight into the relationships observed or not observed
9. Evaluate impact or importance

Preparatory work – Data editing

In a well-executed study, the data collection plan, including procedures, instruments, and forms, are designed and pretested to maximize accuracy. All data collection activities are monitored to ensure adherence to the data collection procotol and to prompt actions to minimize and resolve missing

and questionable data. Monitoring procedures are instituted at the outset and maintained throughout the study, since the faster irregularities can be detected, the greater the likelihood that they can be resolved in a satisfactory manner and the sooner preventive measures can be instituted.

Nevertheless, there is often the need to "edit" data, both before and after they are computerized. The first step is "manual" or "visual editing". Before forms are keyed (unless the data are entered into the computer at the time of collection, e.g., through CATI - computer-assisted telephone interviewing) the forms are reviewed to spot irregularities and problems that escaped notice or correction during monitoring.

Open-ended questions, if there are any, usually need to be coded. Codes for keying may also be needed for closeed-end questions unless the response choices are "precoded" (i.e., have numbers or letters corresponding to each response choice). Even forms with only closed-end questions having precoded responses choices may require coding for such situations as unclear or ambiguous responses, multiple responses to a single item, written comments from the participant or data collector, and other situations that arise. (Coding will be discussed in greater detail below.) It is possible to detect data problems (e.g., inconsistent or out of range responses) at this stage, but these are often more systematically handled at or following the time of computerization. Visual editing also provides the opportunity to get a sense for how well the forms were filled out and how often certain types of problems have arisen.

Data forms will usually then be keyed, typically into a personal computer or computer terminal for which a programmer has designed data entry screens that match the layout of the questionnaire. For small questionnaires and data forms, however, data can be keyed directly into a spreadsheet or even a plain text file. A customized data entry program often checks each value as it is entered, in order to prevent illegal values from entering the dataset. This facility serves to reduce keying errors, but will also detect illegal responses on the form that slipped through the visual edits. Of course, there must be some procedure to handle these situations.

Since most epidemiologic studies collect large amounts of data, monitoring, visual editing, data entry, and subsequent data checks are typically carried out by multiple people, often with different levels of skill, experience, and authority, over an extended period and in multiple locations. The data processing procedures need to take these differences into account, so that when problems are detected or questions arise an efficient routing is available for their resolution and that analysis staff and/or investigators have ways of learning the information that is gained through the various steps of the editing process. Techniques such as "batching", where forms and other materials are divided into sets (e.g., 50 forms), counted, possibly summed over one or two numeric fields, and tracked as a group, may be helpful to avoid loss of data forms. Quality control and security are always critical issues. Their achievement becomes increasingly complex as staff size and diversity of experience increase.

Preparatory work – Data cleaning

Once the data are computerized and verified (key-verified by double-keying or sight-verified) they are subjected to a series of computer checks to "clean" them.

Range checks

Range checks compare each data item to the set of usual and permissible values for that variable. Range checks are used to:

1. Detect and correct invalid values
2. Note and investigate unusual values
3. Note outliers (even if correct their presence may have a bearing on which statistical methods to use)
4. Check reasonableness of distributions and also note their form, since that will also affect choice of statistical procedures

Consistency checks

Consistency checks examine each pair (occasionally more) of related data items in relation to the set of usual and permissible values for the variables as a pair. For example, males should not have had a hysterectomy. College students are generally at least 18 years of age (though exceptions can occur, so this consistency check is "soft", not "hard"). Consistency checks are used to:

1. Detect and correct impermissible combinations
2. Note and investigate unusual combinations
3. Check consistency of denominators and "missing" and "not applicable" values (i.e., verify that skip patterns have been followed)
4. Check reasonableness of joint distributions (e.g., in scatterplots)

Preparatory work – Data coding

Data coding means translating information into values suitable for computer entry and statistical analysis. All types of data (e.g., medical records, questionnaires, laboratory tests) must be coded, though in some cases the coding has been worked out in advance. The objective is to create variables from information, with an eye towards their analysis. The following questions underlie coding decisions:

1. What information exists?
2. What information is relevant?
3. How is it likely to be analyzed?

Examples of coding and editing decisions

- A typical criterion for HIV seropositivity is a repeatedly-positive ELISA (enzyme linked immunosorbent assay) for HIV antibody confirmed with a Western blot to identify the

presence of particular proteins (e.g., p24, gp41, gp120/160). Thus, the data from the laboratory may include all of the following:

a. An overall assessment of HIV status (positive/negative/indeterminant)

b. Pairs of ELISA results expressed as:

   i. + + / + – / – – / indeterminate

   ii. optical densities

c. Western Blot results (for persons with positive ELISA results) expressed as:

   i. (+ / – / indeterminant)

   ii. specific protein bands detected, e.g., p24, gp41, gp120/160

How much of this information should be coded and keyed?

- How to code open-ended questionnaire items (e.g., "In what ways have you changed your smoking behavior?", "What are your reasons for quitting smoking?", "What barriers to changing do you anticipate?", "What did you do in your job?")
- Closed-end questions may be "self-coding" (i.e., the code to be keyed is listed next to each response choice), but there can also be:

  a. Multiple responses where only a single response is wanted – may be

     1. Inconsistent responses (e.g., "Never" and "2 times or more")
     2. Adjacent responses indicating a range (e.g., "two or three times" and "four or five times", by a respondent who could not choose among 2-5 times).

  b. Skipped responses – should differentiate among

     1. Question was not applicable for this respondent (e.g., age at menarche for male respondents)
     2. Respondent declined to answer (which respondents sometimes may indicate as "N/A"!)
     3. Respondent did not know or could not remember
     4. Respondent skipped without apparent reason

It is necessary to achieve a balance between coding the minimum and coding "everything".

- Coding is much easier when done all at once.
- One can always subsequently ignore coded distinctions not judged as meaningful.
- Information not coded will be unavailable for analysis (e.g., date questionnaire received, which questionnaires were randomly selected for 10% verification survey).
- More detail means more recodes for analysis means more programming means more opportunities for error.

- Decisions deferred have to be made sometime, so why not decide up front (e.g., When a respondent circles adjacent response choices, such as "3. Once or twice" and "4. Two to five times", what should be coded – 3?, 4?, 3.5? a missing value code? a code to be replaced at a later date when a decision is made?)

It is important to document how coding was done and how issues were resolved, so that consistency can be achieved and the inevitable questions ("How did we deal with that situatiion?") answered.

Types of variables - levels or scales of measurement

Constructs or factors being studied are represented by "variables". Variables (also sometimes called "factors") have "values" or "levels". Variables summarize and reduce data, attempting to represent the "essential" information.

Analytic techniques depend upon variable types

Variables can be classified in various ways. A **continuous variable** takes on all values within its permissible range, so that for any two allowable values there are other allowable values in between. A continuous variable (sometimes called a "measurement variable") can be used in answer to the question "how much". Measurements such as weight, height, and blood pressure can, in principle, be represented by continuous variables and are frequently treated as such in statistical analysis. In practice, of course, the instruments used to measure these and other phenomena and the precision with which values are recorded allow only a finite number of values, but these can be regarded as points on a continuum. Mathematically, a **discrete variable** can take only certain values between its maximum and minimum values, even if there is no limit to the number of such values (e.g., the set of all rational numbers is countable though unlimited in number). Discrete variables that can take any of a large number of values are often treated as if they were continuous. If the values of a variable can be placed in order, then whether the analyst elects to treat it as discrete and/or continuous depends on the variable's distribution, the requirements of available analytic procedures, and the analyst's judgment about interpretability.

Types of discrete variables

1. **Identification** – a variable that simply names each observation (e.g., a study identifying number) and which is not used in statistical analysis;

2. **Nominal** – a categorization or classification, with no inherent ordering; the values or the variable are completely arbitrary and could be replaced by any others without affecting the results (e.g., ABO blood group, clinic number, ethnicity). Nominal variables can be **dichotomous** (two categories, e.g., gender) or **polytomous** (more than two categories).

3. **Ordinal** – a classification in which values can be ordered or ranked; since the coded values need only reflect the ranking they can be replaced by any others with the same relative ranking (e.g., 1,2,5; 6,22,69; 3.5,4.2, 6.9 could all be used in place of 1,2,3). Examples are injury severity and socioeconomic status.

4. **Count** – the number of entities, events, or some other countable phenomenon, for which the question "how many" is relevant (e.g., parity, number of siblings); to substitute other numbers for the variable's value would change its meaning. In epidemiologic data analysis, count variables are often treated as continuous, especially if the range is large.

Types of continuous variables

1. **Interval** – differences (intervals) between values are meaningful, but ratios of values are not. That is, if the variable takes on the values 11-88, with a mean of 40, it is meaningful to state that subject A's score of 60 is "twice as far from the mean" as subject B's score of 50. But it is not meaningful to say that subject A's score is "1.5 times the mean". The reason is that the zero point for the scale is arbitrary, so values of the scores have meaning only in relation to each other. Without loss of information, the scale can be shifted: 11-88 could be translated into 0-77 by subtracting 11. Scale scores can also be multipled by a constant. After either transformation, subject A's score is still twice as far from the mean as is subject B's, but subject A's score is no longer 1.5 times the mean score. Psychological scales (e.g., anxiety, depression) often have this level of measurement. An example from physics is temperature measured on the Fahrenheit or Celsius scale.

2. **Ratio** – both differences and ratios are meaningful. There is a non-arbitrary zero point, so it is meaningful to characterize a value as "x" times the mean value. Any transformation other than multiplying by a constant (e.g., a change of units) will distort the relationships of the values of a variable measured on the ratio scale. Physiological parameters such as blood pressure or cholesterol are ratio measures. Kelvin or absolute temperature is a ratio scale measure.

Many variables of importance in epidemiology are dichotomous (i.e., nominal with two levels) – case vs. noncase, exposed vs. unexposed. For an apparently ordinal or continuous variable, the phenomenon itself may not warrant treatment as such. It is necessary to ask such question as: "Is "more" really more?" and "Are thresholds or discontinuities involved?" Again, the underlying reality (or, rather, our conceptual model of it) determines the approach to quantification. Variable values are often collapsed into a small number of categories for some analyses and used in their original form for others.

Preparatory work – Data reduction

Data reduction seeks to reduce the number of variables for analysis by combining single variables into compound variables that better quantify the construct. Variables created during coding attempt to faithfully reflect the original data (e.g., height, weight). Often these variables can be used directly for analysis, but it is also often necessary to create additional variables to represent constructs of interest. For example, the construct overweight is often represented by a variable derived from the values for height and weight. Data reduction includes simplifying individual variables (e.g., collapsing six possible values to a smaller number) and deriving compound variables (e.g. "socioeconomic status" derived from education and occupation).

In general:

- Simpler is better
- Avoid extraneous detail
- Create additional variables, rather than destroy the original ones (never overwrite the raw data!).
- Inspect detail before relying on summaries
- Verify accuracy of derived variables and recodes by examining crosstabulations between the original and derived variables.
- Take into account threshold effects, saturation phenomena, and other nonlinearities
- Categorize based on the nature of the phenomenon (e.g., a study of Down's syndrome can collapse all age categories below 30 years; a study of pregnancy rates will require a finer breakdown below 30 years and even below 20 years).

Types of derived variables

Scales - In a pure scale (e.g., e.g., depression, self-esteem) all of the items are intended as individual measures of the same construct. The scale score is usually the sum of the response values for the items, though items with negative valence (e.g., "I feel happy" in a depression scale) must be inverted. The purpose of deriving a scale score by having multiple items is to obtain a more reliable measure of the construct than is possible from a single item. Scale reliability (internal consistency) is typically assessed by using Cronbach's **coefficient alpha**, which can be thought of as the average of all of the inter-item correlations. If the items did indeed measure the same construct in the same way and were indeed answered in an identical manner, then the only differences in their values should be due to random errors of measurement. Cronbach's alpha gives the proportion of the total variation of the scale scores that is not attributable to random error. Values of 0.80 or greater are considered adequate for a scale that will be used to analyze associations (if the scale is used as a clinical instrument for individual patients, its alpha should be at least 0.90 – see Nunally's textbook, *Psychometrics*). When the scale consists of separate subscales, internal consistency may be more relevant for the individual subscales than for the scale as a whole. Analyses of relationships between individual items (inter-item correlation or agreement), between each item and the remaining items (item-remainder correlation), between each item and the total scale (item-scale correlation), and among groups of items (factor analysis) are standard methods of analyzing item performance.

Indexes - An index consists of a group of items that are combined (usually summed) to give a measure of a multidimensional construct. Here, each of the items measures a different aspect or dimension, so that internal consistency measures like Cronbach's alpha are either not relevant or require a different interpretation. Examples of indexes derived from several variables include socioeconomic status (e.g., occupation, income, education, neighborhood), social support (e.g., marital status, number of close family members, number of close friends), sexual risk behavior (number of partners, types of partners, use of condoms, anal intercourse). Items may have different weights, depending upon their relative importance and the scale on which they were measured.

Algorithms - A procedure that uses a set of criteria according to specific rules or considerations, e.g., major depressive disorder, "effective" contraception (I have not seen this term used to designate a type of variable before, but I am not aware of any other term for this concept).

Preparatory work – Exploring the data

Try to get a "feel" for the data – inspect the distribution of each variable. Examine bivariate scatterplots and cross classifications. Do the patterns make sense? Are they believable?

- Observe shape – symmetry vs. skewness, discontinuities
- Select summary statistics appropriate to the distribution and variable type (nominal, ordinal, measurement)

  Location - mean, median, percentage above a cut-point

  Dispersion - standard deviation, quantiles
- Look for relationships in data
- Look within important subgroups
- Note proportion of missing values

Preparatory work – Missing values

Missing data are a nuisance and can be a problem. For one, missing responses mean that the denominators for many analyses differ, which can be confusing and tiresome to explain. Also, analyses that involve multiple variables (e.g., coefficient alpha, crosstabulations, regression models) generally exclude an entire observation if it is missing a value for any variable in the analysis (this method is called **listwise deletion**). Thus, an analysis involving 10 variables, even if each has only 5% missing values, could result in excluding as much as 50% of the dataset (if there is no overlap among the missing responses)! Moreover, unless data are **missing completely at random** (**MCAR** – equivalent to a pattern of missing data that would result from deleting data values throughout the dataset without any pattern or predilection whatever), then an analysis that makes no adjustment for the missing data will be biased, because certain subgroups will be underrepresented in the available data (a form of selection bias).

Imputation for missing values - optional topic

As theory, methods, and computing power have developed over the years, analytic methods for handling missing data to minimize their detrimental effects have improved. These methods seek to **impute** values for the missing item responses in ways that attempt to increase statistical efficiency (by avoiding the loss of observations which have one or a few missing values) and reduce bias. Earlier methods of imputation, now out of favor, include replacing each missing value by the mean or median for that variable. Even though such practices enable all observations to be used in regression analyses, these methods do not

reduce bias and tend to introduce additional distortion. More sophisticated methods reduce bias from missing data while minimizing distortion from imputation. These methods derive imputations that make use of the values of variables for which data are present and which are related to the variable being imputed.

Typically, **complete data cases** (observations that have no missing values for the variables of interest) serve as the raw material for the imputations. Factors that are theoretically related to the variables to be imputed and with which they are associated in the complete data cases are used to develop "predictive" models for the imputed variables. These models are then applied to the remaining observations, providing predicted ("imputed") values for their missing responses. The resulting imputations are said to be **conditioned on** the variables in the model.

For example, suppose the available data show a positive correlation between blood pressure and age. By conditioning imputations on age, we impute (on average) higher blood pressures to older subjects with missing blood pressure data and lower blood pressures to younger subjects missing blood pressure data. This technique preserves the relationship between age and blood pressure that exists in the complete data cases. Moreover, if older subjects are more likely to be missing a blood pressure reading, then the conditioning reduces the bias from analyzing only the complete data cases.

If the process that led to the missing data is uniformly random except for being positively related to identifiable factors (e.g., subject's age), then the missing data process is called **missing at random** (**MAR**), rather than MCAR. In such a situation, the overall mean blood pressure for the complete data cases is biased downwards (due to underrepresentation of older subjects), but the overall mean based on imputations conditioned on age is not.

If predicted values are simply substituted for missing values, however, then although bias will be reduced so will standard errors. The reason is that the imputation models were created based on (imperfect) associations between the conditioning variables and the variables being imputed. In contrast, the predicted values are directly computed from the model as if, in our example, blood pressure were completely determined by age. In effect, the model functions as a "self-fulfilling prophecy". To avoid this problem a source of random variability is introduced into the imputation process. For example, rather than substituting the predicted values themselves for the missing data, the imputed values may be sampled from distributions whose means are the predicted values (e.g., if the estimated mean for a yes-no response were 0.30 [where 1="yes" and 0="no"], then the imputed value would be generated randomly from a binomial distribution with a proportion of "successes" of 0.30).

In addition, by using multiple imputations (typically five), the analyst can adjust the standard errors to reflect the uncertainty introduced by the imputation process. Carrying out multiple imputations means repeating the imputation process to create multiple versions of the dataset (one for each imputation), analyzing each dataset separately, and combining the results according to certain procedures.

Imputation causes the least distortion when the proportion of missing data is small, and data are available for variables that are strongly associated with the variable being imputed. Perversely, however, imputation is most needed when the proportion of missing data is large. Also, unfortunately, the available data may provide little guidance about whether the missing process is MCAR, MAR, or "nonignorable". Attention to causes of missing responses during data collection can be helpful (Heitjan, 1997).

[I would like to thank Michael Berbaum and Ralph Folsom for their patient explanations of imputation and for reading over earlier versions of this section.]

Descriptive analyses

Exploration of the data at some point becomes descriptive analysis, to examine and then to report measures of frequency (incidence, prevalence) and extent (means, survival time), association (differences and ratios), and impact (attributable fraction, preventive fraction). These measures will be computed for important subgroups and probably for the entire study population. Standardization or other adjustment procedures may be needed to take account of differences in age and other risk factor distributions, follow-up time, etc.

Evaluation of hypotheses

After the descriptive analyses comes evaluation of the study hypotheses, if the study has identified any. Here there will be a more formal evaluation of potential confounding, other forms of bias, potential alternative explanations for what has been observed. One aspect of both descriptive analysis and hypothesis testing, especially of the latter, is the assessment of the likely influence of random variability ("chance") on the data. Much of the field of statistics has grown up to deal with this aspect, to which we will now turn.

Evaluating the role of chance - inference

Whether or not we believe, in Albert Einstein's words, that "the Lord God doesn't play dice with the universe", there are many events in the world that we ascribe to "chance". When we roll a die, the resulting number is generally unpredictable and does not (or at least, should not) follow any evident pattern. Similarly, when we draw five cards from a freshly-shuffled, unmarked deck, we know that some outcomes are more or less likely than others (e.g., a pair is more likely than three of a kind), but we cannot predict what cards we will draw. The theories of probability and statistics were born in the gaming parlors of Monte Carlo and came of age in the fields of the British countryside. The computer revolution put their power, for good or for whatever, into the hands of any of us who can click a mouse.

The basis for the incorporation of the fruits of the theory of probability and statistics into medical and epidemiologic research has been recounted by Austin Bradford Hill as follows:

> "Between the two world wars there was a strong case for emphasizing to the clinician and other research workers the importance of not overlooking the effects of the play of chance upon their

> data. Perhaps too often generalities were based upon two men and a laboratory dog while the treatment of choice was deduced from a difference between two bedfuls of patients and might easily have no true meaning. It was therefore a useful corrective for statisticians to stress, and to teach the need for, tests of significance merely to serve as guides to caution before drawing a conclusion, before inflating the particular to the general." (pg 299 in The environment and disease: association or causation. Proceedings of the Royal Society of Medicine, 1965: 295-300)

From this innocent and commonsensical beginning, statistical procedures have (like kudzu?? – just kidding!) virtually engulfed the thinking of researchers in many fields. Hill continues:

> "I wonder whether the pendulum has not swung too far – not only with the attentive pupils but even with the statisticians themselves. To decline to draw conclusions without standard errors can surely be just as silly? Fortunately I believe we have not yet gone so far as our friends in the USA where, I am told, some editors of journals will return an article because tests of significance have not been applied. Yet there are innumerable situations in which they are totally unnecessary - because the difference is grotesquely obvious, because it is negligible, or because, whether it be formally significant or not, it is too small to be of any practical importance. What is worse the glitter of the t table diverts attention from the inadequacies of the fare. . . . "

He admits that he exaggerates, but he suspects that the over-reliance on statistical tests weakens "our capacity to interpret data and to take reasonable decisions whatever the value of P." Hill is referring to tests of significance, which are probably the most common procedures for assessing the role of chance, or perhaps more precisely, the amount of numerical evidence that observed differences would not readily arise by chance alone.

Illustration of a statistical test

Consider the following data, from the first study to report an association between adenocarcinoma of the vagina and maternal use of diethylstilbestrol (DES). During the 1960's, a handful of cases of adenocarcinoma of the vagina were observed in young women, a highly unusual occurrence. Investigation into the histories of the affected women revealed that in most cases the girl's mother had taken diethylstilbestrol (DES) while she was carrying the girl in her uterus. At that time DES had been prescribed in the belief that it might prevent premature delivery in women who had lost pregnancies. In how many patients would this history have to emerge for it before the investigators could be confident that it was not a chance observation? This question is usually answered by means of a statistical test.

Prenatal exposure to diethylstilbestrol (DES)
among young women with adenocarcinoma of the vagina

	Exposed to diethylstilbesterol?		
	Yes	No	Total
Cases	7	1	8
Controls	0	32	32
Total	8	33	40

Source: Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *New Engl J Med* 1971; 284:878-881. [From Schlesselman JJ. *Case-Control Studies*. New York, Oxford, l982: 54]

All but one of the cases had a positive history for intrauterine exposure to DES. In contrast, not one of 32 controls did. The relative risk from this table cannot be calculated directly, because of the zero cell, but adding 0.5 to all four cells yields a relative risk (OR) of 325, a stronger association than most of us will ever encounter in our data. However, this study has only eight cases. Could these results be due to chance?

A statistical test of significance is a device for evaluating the amount of numerical data on which an observed pattern is based, to answer a question like, "How often could such a strong association arise completely by chance in an infinite number of analogous experiments with the same number of subjects and the same proportion of cases (or of exposed)?" This question is not quite the same as "How likely is it that chance produced the association in the table?" nor as "How much of the association is due to chance?". But if such a strong association would arise only very infrequently by chance alone, then it is reasonable to suppose that at least some potentially identifiable factor has contributed to the observed association. That factor could be bias, of course, rather than the exposure, but at least it would be something other than chance. Conversely, it is also possible that much stronger associations could readily arise by chance and yet the one we observed might reflect a causal process. The significance test simply evaluates the strength of numerical evidence for discounting chance as a likely sufficient explanation.

In order to conduct a test of significance, we need to operationalize the concept of "analogous experiment". There's the rub. What kind of experiment is analogous to an epidemiologic study, all the more so an observational one? For the above table, the significance test that would be used is Fisher's Exact Test. The analogous experiment (**probability model**) here is equivalent to the following:

> Suppose that you have 40 pairs of socks – 7 pairs of red socks, and 33 pairs of blue socks. You want to pack 8 pairs of socks in your travel bag, so without looking you take 8 pairs at random and put them in your bag. How many red pairs have you packed for your trip?

When this "analogous experiment" is repeated a sufficient number of times, the proportion of trials in which the bag has 7 red pairs will provide the probability that chance alone would produce a situation in which you had packed 7 pairs of red socks. This probability is the "p-value" for the significance test of the relationship between adenocarcinoma of the vagina and maternal DES in the above table.

Fortunately, the distribution of the number of red pairs in the bag has already been worked out theoretically, so that the exact probability can be computed without having to carry out what in this case would be a VERY large number of trials. The formula for the (hypergeometric) distribution is:

$$\Pr(A=j) = \frac{C^{n_1}_{j} \; C^{n_0}_{(m_1 - j)}}{C^{n}_{m_1}} = \frac{n_1! n_0! m_1! m_0!}{n!\, j!\, (n_1 - j)!\, (m_1 - j)!\, (n_0 - m_1 - j)!}$$

where Pr(A=j) is the probability of obtaining j red pairs of socks in the travel bag and m_0, m_1, n_0, n_1, and n are the row and column totals in the table:

	Color		
	Red	Blue	Total
Travel bag	j	$m_1 - j$	m_1
In drawer	$n_1 - j$	$n_0 - m_1 - j$	m_0
Total	n_1	n_0	n

Here is how the formula is applied:

	Red (DES)	Blue	Total
Packed (cases)	7	1	8
In drawer (controls)	0	32	32
Total	8	33	40

Possible outcomes (Colors of pairs of socks in travel case)		Probability of each outcome	
Red	Blue		
0	8	.181	
1	7	.389	
2	6	.302	
3	5	.108	
4	4	.019	
5	3	.0015	$\frac{7!\ 33!\ 8!\ 32!}{40!\ 5!\ 2!\ 3!\ 30!}$
6	2	.00005	
7	1	4.3 x 10-7	p-value
8	0	0	
		1.0000	

Comments on the "red socks" model:

1. A model is a system or structure intended to represent the essential features of the structure or system that is the object of study. The above model is a very simplified representation!

2. The model is derived on the basis of certain constraints or assumptions (e.g., in this case, 8 cases, 7 DES-exposed mothers, and 40 subjects in all – "fixed marginals" – as well as "all permutations are equally likely').

3. The model underlying hypothesis testing assumes a repeatable experiment and an *a priori* specification of the "hypothesis" being tested – a "null" hypothesis [this is embodied in the model of "equally likely" permutations] and an "alternative hypothesis" [this deals with what results we would regard as inconsistent with the null hypothesis].

4. The above model is tedious to compute for large tables, though computers have solved that problem.

Concept of hypothesis testing (tests of significance)

What we really want to know is: "Is the observed association due to chance?", or "How likely is it that the observed association is due to chance?". This probability is sometimes referred to as the "posterior [*a posteriori*] probability", the probability that the hypothesis is true given the observed results. (The "prior [*a priori*] probability" that the hypothesis is true is our belief in the hypothesis before we have the results in question). The frequentist school of statistics, from which significance testing derives, cannot answer this question directly. Instead, significance tests and p-values attempt to provide an indirect answer, by reformulating the question as: "How often would an association as strong as that observed occur by chance alone?". The role of chance is played by a suitable probability model, chosen to represent the probability structure of the data and the study design. But most epidemiologic studies deviate rather markedly from the probability models on which statistical

tests are based (e.g., see Sander Greenland, Randomization, statistics, and causal inference), so although statistical theory is extremely precise, it must be thoughtfully applied and thoughtfully interpreted.

A compromise version of the question that underlies a significance test is "How consistent are the numerical data with what would be expected 'by chance' - as played by a suitable probability model". The probability model is most often one that assumes no systematic difference between groups, partly because such models are easier to derive and also because it is often convenient for the hypothesis-testing framework. The result of the significance test is a probability (the **p-value**), which provides a quantitative answer to this compromise question. (Note: The statistical "null hypothesis" is rarely of interest from a substantive perspective. A study hypothesis should be stated in terms of no association only if that is indeed what the investigator hopes to demonstrate. In fact, it is quite difficult to demonstrate the absence of association, since the evidence for no association is related to the Type II error probability (1 – statistical power) for the study, which is generally considerably greater than the significance level – see below).

The p-value itself can be regarded as a descriptive statistic, a piece of evidence that bears on the amount of numerical evidence for the association under study. When a decision is called for, though, then some method of assigning an action to the result of the significance test is needed. Decision-making entails the risk of making errors. Ideally the loss function (the costs of errors of various types) is known explicitly. Under broadly applicable assumptions, though, the theory of decision-making provides a technique for decision-making based on the results of a statistical test. That technique is statistical hypothesis testing.

As noted, the hypothesis to be tested is generally a "null hypothesis" (usually designated H_0). H_0 is the probability model that will play the role of chance (for example, the red socks model). In the present context, that model will be based on the premise that there is no association. If there is sufficient numerical evidence to lead us to reject H_0, then we will decide that the converse is true, that there is an association. The converse is designated as the "alternate hypothesis" (H_A). The decision-making rule is to reject H_0, in favor of H_A, if the p-value is sufficiently small, and to otherwise accept H_0.

Since we have a decision between two alternatives (H_0 and H_A) we can make two kinds of errors:

Type I error: Erroneously reject H_0 (i.e., conclude, incorrectly, that data are not consistent with the model)

Type II error: Erroneously fail to reject H_0 (i.e., conclude, incorrectly, that data are consistent with the model)

(The originator of these terms must have been more prosaic than the originators of the terms "significance", "power", "precision", and "efficiency") Traditionally, the Type I error probability has received more attention and is referred to as the "**significance level**" of the test.

In a strict decision-making mode, the result of the significance test is "Reject null hypothesis" or "Fail to reject null hypothesis". (Note that "fail to reject the null hypothesis" is not equivalent to declaring that the null hypothesis is true.) However, rarely must a decision be made based on a single study, so it is preferable to report the calculated p-value (probability that the assumed probability model would produce data as extreme or more so). The p-value gives more information than the statement "results were significant at the 5% level", since it quantifies the degree to which the data are incompatible with "chance" (as played by the probability model), allowing the reader to apply his/her tolerance for a Type I error. Note that the p-value is not a direct index of the strength of an association in an epidemiologic sense nor of its biologic, clinical, or epidemiologic "significance". The p-value simply assesses the compatibility of the observed data with the assumed probability model that serves to represent H_0.

There are many methods for obtaining a p-value or conducting a test of statistical significance. The choice depends upon the level of measurement of the variables (dichotomous, nominal polytomous, ordinal, continuous), the sampling design from which the data came, and other factors. The statistical test illustrated above is an "exact" test (Fisher's exact test), since it is based on a model that considers all possible outcomes and in how many ways each can occur. In an exact test, the probability model is readily apparent.

Illustration of an asymptotic test

More commonly-used, because they are much simpler to compute, are *asymptotic tests* (e.g., chi-square, t-test). Asymptotic tests are approximations whose accuracy improves as the sample size increases, and the underlying probability model on which they are based tends to be more abstract. Typically, asymptotic tests are based on the "normal" (Gaussian) distribution. Why the Gaussian distribution? Because it offers a number of analytic advantages and, most especially, because of the Central Limit Theorem ("one of the most remarkable theorems in the whole of mathematics", Mood and Graybill, 1963:149). The Central Limit Theorem holds that if we take large enough random samples from any distribution with a finite variance, the means of those samples will have an approximately Gaussian distribution.

The general form for such a test is (see Rothman, *Modern epidemiology,* p. 139 or Kleinbaum, Kupper, and Morgenstern, *Epidemiologic research*):

$$Z = \frac{a - E(a)}{\sqrt{var(a)}}$$

where "a" is the observed value (e.g., the number of exposed cases), E(a) is the expected value for "a" under the null hypothesis (a.k.a. analogous experiment), and var(a) is the variance of "a" under the null hypothesis. Thus, Z is the number of standard deviations by which "a" differs from what would be expected if there were no association and has an approximate unit normal distribution. (Z is occasionally written as χ. (called "chi", a unit normal distribution is the same as the square root of a one-degree-of-freedom chi-square distribution).

The probability associated with being "Z" standard deviations away from the mean of a normal distribution can be computed and is readily available in statistical tables (see table excerpt below). The value of a normally-distributed random variable is usually (i.e., probability 95%) less than two standard deviations from its mean, so if Z exceeds 1.96 we say "p<.05", or more precisely, we take the value we have calculated for Z, look it up in a table of the normal distribution and read off the corresponding p-value.

The table excerpt below shows various probabilities derived from the unit normal distribution. For example, the probability associated with a distance of 1.645 standard deviations above the mean is shown in column B (0.05) and is identical to the probability associated with a distance of 1.645 standard deviations below the mean (since the normal distribution is symmetrical). The probability associated with obtaining a value of z that is either above or below a distance of 1.645 standard deviations from the mean is shown in column D (0.10). So if using the formula above (or one of those below) we obtain a value of Z equal to 1.645, then the p-value is either 0.05 or 0.10, depending upon the alternative hypothesis.

Excerpt from a table of the Normal Distribution

z	h	A	B	C	D	E
0.00	0.3989	0.0000	0.5000	0.0000	1.0000	0.5000
0.01	0.3989	0.0040	0.4960	0.0080	0.9920	0.5040
0.02	0.3989	0.0080	0.4920	0.0160	0.9840	0.5080
. . .	. . .	. . .	. . .	. . .	. . .	. . .
0.8416	0.2800	0.30	0.20	0.60	0.40	0.80
. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.282	0.1755	0.40	0.10	0.80	0.20	0.90
. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.645	0.1031	0.45	0.05	0.90	0.10	0.95
. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.960	0.0585	0.475	0.025	0.95	0.05	0.975
. . .	. . .	. . .	. . .	. . .	. . .	. . .
2.576	0.0145	0.495	0.005	0.99	0.01	0.995
. . .	. . .	. . .	. . .	. . .	. . .	. . .
3.090	0.0034	0.499	0.001	0.998	0.002	0.999
. . .	. . .	. . .	. . .	. . .	. . .	. . .

Legend:

z = number of standard deviations to the right of the mean

h = height of the normal curve for that number of standard deviations from the mean

A = area between the mean and z

B = area to the right of z (or to the left of -z)

C = area between -z and +z

D = area beyond |z| (i.e., to the left of -z and the right of +z)

E = area to the left of z

(Source: National Bureau of Standards – Applied Mathematics Series–23, U.S. Government Printing Office, Washington, D.C., 1953, as abstracted in Table A-4 in Richard D. Remington and M. Anthony Schork, *Statistics with applications to the biological and health sciences*. Englewood Cliffs, NY, 1970.]

One-sided versus two-sided p-values

Recall that the p-value is the probability of obtaining an association as strong as (or stronger than) the association that was observed. It turns out, though, that the phrase "as strong as (or stronger than)" is ambiguous, because it doesn't specify whether or not it is intended to include inverse associations, i.e., associations in the opposite direction from the putative association that motivated the study. For example, if we observe a 2.5 relative risk, does "as strong" mean only relative risks of 2.5 or larger, or does it also mean relative risks of 0.4 or smaller? If the former (only 2.5 and larger), then the corresponding p-value is "one-sided" (or "one-tailed"). In contrast, if H_A is "either greater than or equal to 2.5 or [inclusive] less than or equal to 0.4", then a two-sided p-value is indicated. [Only one-sided p-values can be interpreted as the "probability of observing an association as strong or stronger under the chance model" (Rothman and Greenland,185).]

The issue of one-sided versus two-sided p-values can arouse strong emotions. For a given calculated value of Z, the one-sided p-value is exactly half of the two-sided p-value. Proponents of two-sided p-values argue that a one-sided p-value provides an inflated measure of the statistical significance (low probability of obtaining results by chance) of an association. Appropriate situations for using one-sided p-values are sometimes characterized as ones where the investigator has no interest in finding an association in the opposite direction and would ignore it even it occurred. However, a posting on the EPIDEMIOL-L listserv asking for situations of this sort produced very few persuasive examples.

Here is a dramatical presentation of some of the issues in choosing between 1-sided and 2-sided p-values:

> The wife of a good friend of yours has tragically died from lung cancer. Although she was a life-long nonsmoker, your friend used to smoke quite heavily. Before her death she had become an anti-smoking activist, and her last wishes were that your friend bring suit against R.J. Morris, Inc., the manufacturer of the cigarette brand your friend used to smoke. Knowing that he cannot afford expert consultation, your friend turns to you and prevails upon you to assist him with the lawsuit.
>
> In preparation for the trial, the judge reviews with both sides the standard of evidence for this civil proceeding. She explains that for the court to find for the plaintiff (your side) it must

conclude that the association is supported by "a preponderance of the evidence", which she characterizes as "equivalent to a 90% probability that R.J. Morris' cigarettes caused the disease". The R.J. Morris attorney objects, declaring that, first of all, only the probability that cigarettes can cause disease can be estimated, not the probability that cigarettes did cause the disease. As the judge is about to say that the judicial interpretation of probability permits such a conclusion, the R.J. Morris attorney raises her second objection: since the plaintiff is basing his case on scientific evidence, the plaintiff's case should be held to the conventional standard of evidence in science, which requires a significance level of 5%. [Recall that the significance level is the probability of a Type I error, which in this case would mean finding the company responsible even though your friend's lung cancer was really due to chance. If the court were to fail to find the tobacco company responsible, even though the company's cigarettes did cause the cancer, that would be or a Type II error.]

Seeing an opportunity, you pass a note to your friend, who passes it on to his attorney. Upon reading it, his attorney says to the judge "Your Honor, my client is prepared to accept the R.J. Morris' insistence on a 5% significance level, provided that it is based on a one-sided alternative hypothesis." Beginning to regret that she introduced the probability metaphor, the judge turns to the R.J. Morris attorney, who is now hastily conferring with her biostatistician. After a quick consultation, the R.J. Morris attorney charges indignantly that plaintiff's attorney is trying, through deception, to obtain a lower standard of evidence. A 5% one-tailed significance level, she charges, is actually a 10% significance level, since everyone knows that two-tailed tests are more appropriate. Your friend's attorney senses that this charge will be a telling point with the judge and anxiously looks back to you for advice on how to respond.

With your coaching, your friend's attorney replies that a two-tailed test is warranted only when the appropriate alternative hypothesis (H_A) is two-sided. The question in this case is whether R.J. Morris is or is not liable, i.e., whether their cigarettes did or did not cause the cancer. This question corresponds to a one-sided H_A, i.e., the court can (1) reject H_0 (no causation) in favor of the alternative that R.J. Morris is liable or (2) fail to reject H_0, if the court finds the evidence insufficient. "May it please the court," she continues, "there is no issue here that the cigarette smoke could have acted to <u>prevent</u> the cancer from occurring, so requiring a two-tailed alternative hypothesis is tantamount to imposing a significance level of 2.5%, which is closer to the standard for a criminal, rather than a civil, trial".

With the benefit of further consultation, the R.J. Morris attorney "strenuously objects". "Plaintiff may see this case as involving a one-sided H_A, but notwithstanding the proposed tobacco settlement, as far as the R.J. Morris Company is concerned the relationship between smoking and cancer has not been proved. Therefore a finding that cigarette smoking can in fact prevent cancer is just as relevant as plaintiff's contention that the cigarettes were responsible."

You are naturally outraged by the R.J. Morris lawyer's assertion that the relationship between smoking and cancer is not proved, but you have to put that aside as your friend's lawyer asks you is it not correct that the significance level is simply a mechanism for deciding how many standard deviations away from the mean are required to exclude chance as an explanation. Usually, people exclude chance when the statistical test comes out two standard deviations from the center of a normal distribution (actually, 1.96 standard deviations, which corresponds to a

two-tailed 5% significance level). If the judge does accept the one-tailed 5% significance level, even with a good argument that because the appropriate H_A is one-sided so that the Type I error probability is really only 5%, a decision that meets the test of being only 1.65 standard deviations from the mean (corresponding to a one-tailed 5% significance level) may be vulnerable on appeal. Since the scientific evidence is firm, would it be better to agree to the two-tailed test?

The judge looks at her watch, and you see beads of perspiration breaking out on your friend's attorney's forehead. Meanwhile you're trying to sort through the issues. You've only just received your epidemiology degree, and you aren't yet sure that it works. It's true that an appeals court might reject the idea of a one-tailed test, since appellate judges tend to be conservative, and R.J. Morris will certainly appeal an adverse judgment. But then a dark thought jumps into your mind. What if R.J. Morris has concocted evidence that will somehow make it appear that your friend <u>is</u> responsible for his wife's death from lung cancer? You know that that is outlandish, but what if they could? With a two-sided H_A, the court could reject H_0 <u>and</u> find your friend responsible, thereby destroying him financially and emotionally. "One-sided!", you cry out …and then suddenly you wake with a start. The professor and your fellow students are looking at you with puzzlement, wondering what question you thought that you were responding to. As you emerge from your daydream you hope that you have not slept through too much of the lesson and vow to go to bed earlier in the future.

Significance testing in a two-by-two table

For a two-by-two table, the formula can be more easily expressed for computational purposes by defining "a" as the contents of a single cell in the table, conventionally the "a" (upper left corner) cell, so that E(a) is the value expected for "a" under the null hypothesis (n_1m_1/n), and Var(a) is the variance of "a" under the null hypothesis $\{(n_1n_0m_1m_0)/[n^2(n-1)]$, based on the hypergeometric distribution. The test statistic is then simply:

$$Z = \frac{a - n_1m_1/n}{\sqrt{\{ (n_1\, n_0m_1\, m_0) \,/[\, n^2\, (n-1)]\}}}$$

An equivalent, but more easily remembered computation formula, is:

$$Z = \sqrt{X^2} = \sqrt{\frac{(ad - bc)^2\,(n-1)}{n_1\, n_0m_1\, m_0}}$$

[Note: you may also see the above formula with n, instead of (n-1) [e.g., Hennekins and Buring, p. 251 uses T instead of (n-1)]. The reason is that the above formula gives a Mantel-Haenszel chi-square statistic (based on the hypergeometric distribution) instead of the Pearson chi-square statistic (based on the normal distribution). For large samples the two are essentially equivalent. There are parallel formulas for person-time data.]

	Exposed to diethylstilbesterol?		
	Yes	No	Total
Cases	a	b	m_1
Controls	c	d	m_0
Total	n_1	n_0	n

Whatever misgivings we may have about the statistical model and its application, results with as small a p-value as that obtained in this study will be very satisfying to practically any investigator who obtains them. But to appreciate the dynamics of this procedure, and the problems of interpretation that arise in more equivocal circumstances, let us analyze what underlies a small p-value.

A small p-value (i.e., low probability that results similar to those observed would be produced by "chance" [as played by a given statistical model]) reflects:

- a strong observed association (or a large observed difference)

or

- a large sample size (roughly speaking).

Therefore, when the p-value is not small, there are two possibilities (ignoring the possibilities of systematic error, inappropriate statistical model, etc.):

1. the observed association or difference is not strong.
2. the observed association is of a respectable size but the study size was too small to judge it "significant."

How we interpret a failure to obtain a low p-value depends upon our judgment of the magnitude of the observed association and of the statistical power of the study to detect an important real difference.

If the p-value is small (e.g., less than five percent [typical], ten percent [less common], or one percent [for the demanding or rich in data]), the observed results are somewhat inconsistent with an explanation based on chance alone, so we are inclined to view them as having some origin worth inquiring about (e.g., systematic influences from the way the study was designed or conducted, biological or psychosocial processes related to the factors under study, etc.). If the observed difference or association is too small to be scientifically or clinically significant (as opposed to statistically significant), we will not care to pursue the matter regardless of the p-value.

If the p-value is not small (i.e., the results are "not significant"), was an association observed? If no association was observed, then the appropriate characterization of the finding is "no association was observed" (but see below). If an association was observed, then we can say "an association was observed but the data were insufficient to discount chance as an explanation" [not "there was no association"!].

If no association was observed, then we also need to ask what were our chances of detecting a meaningful association if one exists? If statistical power was low, then we cannot say much. If statistical power was high, then we can say the data provide evidence (assuming, always, that bias is not present) against the existence of a strong association.

If the observed association is strong enough to be important if it is not due to chance, then the only conclusion we can draw is that the data do not provide sufficient evidence to discount an explanation of chance alone – this is not equivalent to a conclusion that "an association was not observed" [since one was] or that "the observed association is due to chance" [which no one knows]. Other characterizations often stated are also unfortunate:

"The observed association is not significant" [which tends to impugn it]

"The association did not reach statistical significance" [which implies that the association should have been stronger – it may be as strong as it should be but be based on too few subjects.]

Better to say "an association of ____ was observed, but the data were too few to discount an explanation based on chance" or some similar expression. [Note: Any result can become "nonsignificant" if we stratify enough.]

An alternate possibility is that the observed association was too weak to be meaningful even if it had been associated with a small p-value. Here our conclusion depends upon the size of the study, i.e., its statistical power to detect an association of some particular magnitude. If the power was low, if the study's ability to detect a difference we would regard as important is low, then there really is not much we can say or conclude, except that our failure to find an association could well be due to chance (i.e., we may well have made a "Type II error"). This inability is one of the reasons for discouraging researchers from conducting small studies except as a pilot study to develop procedures and instruments. If the power was high, then we are in a better position to interpret our results as evidence against the existence of a real association.

Statistical power and sample size

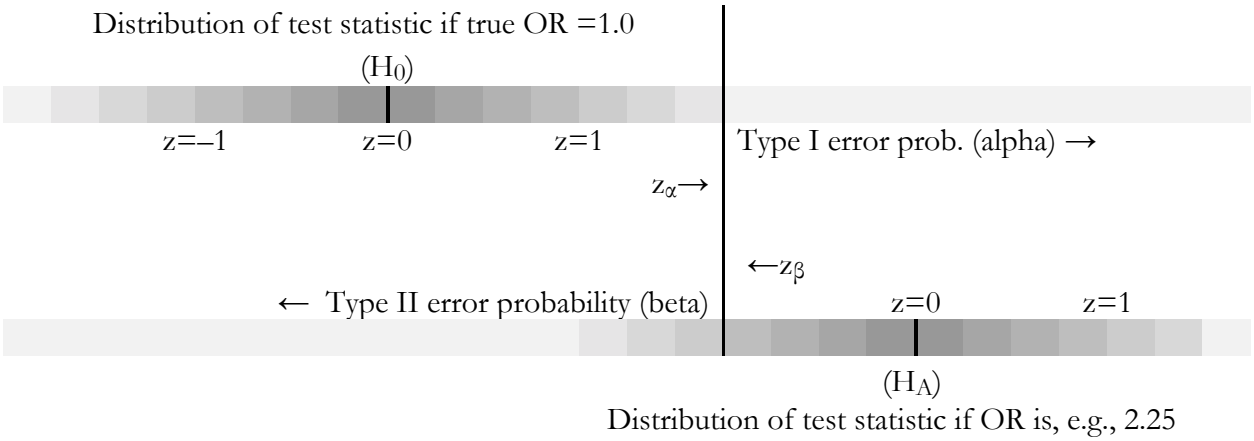
Statistical power refers to the ability to detect an association of interest in the face of sampling error. Suppose that there is a true association of a certain type and degree, but that through the workings of chance our studies will observe the association to be weaker or stronger. In order to be reasonably certain that our study will detect the association, the study has to be large enough so that sampling error can be contained.

For example, suppose that we are comparing a group of cases of Alzheimer's disease cases to a group of controls to see whether the cases are different in respect to presence of a specific gene. Suppose that this gene is actually present in 20% of cases and in 10% of the population from which the cases arose (i.e., the OR in a large, unbiased case-control study would be 2.25). If we study 20 cases and 20 controls, we may well find 4 cases with the gene and 2 controls with the gene, so that we correctly estimate the prevalence of the gene in cases and in the population and the OR.

With such few subjects, however, we could very easily get only 3 cases with the gene and 3 controls with the gene, completely failing to detect the difference in prevalence (OR = 1.0). In fact, we might even get 4 controls with the gene and only 2 cases with the gene, so that the gene appear to be protective (OR = 0.44). Of course, we would not want to react to a difference or an OR that may readily be due to chance, so we will test whatever result we observe to make sure that it is greater than is expected to occur by chance alone (i.e.,"significant"). That means that we will discount any association we observe if it is less what we regard as within chance expectation. (Or recalling our courtroom fantasy, a "preponderance of the evidence", not merely suspicion.)

Therefore, in order to detect an association, we must both (1) observe it in our study and (2) decide that chance would not likely have created it. Each of these requirements places a demand on the size of the study. We need at least some minimum number of subjects so that (1) we have a reasonable expectation of observing an association if one exists (i.e., that we will not make a type II error), and (2) we will think it unlikely that chance could produce an association of that size.

Statistical power to detect an OR ≠ 1.0 with a one-tailed significance test



This diagram illustrates the overlap between the central portions of the distributions of a test statistic (e.g., Z) expected under the null hypothesis (e.g.., true OR is 1.0) and alternate hypothesis (e.g., true OR is 2.25). When we obtain the results from the study we will compute the test statistic (e.g., Z) and compare it to its distribution under the H_0 (the upper of the two distributions in the diagram). If the calculated value of Z is smaller than z_α, i.e., it falls to the left of the cutpoint we have set (defined by the Type I error probability, alpha), then we will conclude that the data we observed came from the upper distribution (the one for no association, true OR=1.0). Even if the

OR we observed was greater than 1.0 (which implies that Z was greater than 0), because Z was not greater than our cutpoint we regard the observed OR as a chance deviation from 1.0. If the unseen truth is that there really is no association, then our conclusion is correct. If instead the true OR is really 2.25, so that the data we observed really came from the lower distribution, then our conclusion represents a Type II error. The area to the left of the cutpoint on the lower distribution represents the probability of making a Type II error, "beta". Statistical power – the probability of detecting a true difference – is equal to one minus beta (i.e., 1 - beta).

Conversely, if we observe a value of Z to the right of the cutpoint, we will conclude that the data we observed did not come from the upper distribution and that therefore the true OR is greater than 1.0. If we are incorrect – if the association we observed was in fact simply a chance finding – then our conclusion represents a Type I error. The area to the right of the cutpoint on the upper distribution represents the probability of making a Type I error, "alpha".

If we abhor making a Type I error, we can move the cutpoint to the right, which reduces alpha – but increases beta. If we prefer to reduce beta, we can move the cutpoint to the left – but that increases alpha. What we would really like to do is to reduce both alpha and beta, by making the distributions narrower (so that more of the shading is located at the center of the each distribution, symbolizing greater precision of estimation). The width of the distribution is controlled by the sample size. With a powerful light we can easily distinguish between, for example, a snake and a stick. But with a weak light, we cannot be certain what we are seeing. We can elect to err on one side or the other, but the only way to reduce our chance of error is to get a more powerful light.

Commonly used values for alpha and beta are, respectively, 0.05 and 0.20 (power=0.80), for a total probability of error of 0.25. If the study size is limited due to the low incidence of the disease, the low prevalence of the exposure, or the low amount of the budget, then our study estimates will be imprecise – the distributions in the above diagram will be wide. The total error probability will be below 0.25 only when the lower distribution is farther to the right, i.e., corresponds to a stronger association.

In essence, intolerance for error (i.e., small alpha and beta) and desire to detect weak associations must be paid for with sample size. In our courtroom daydream, the better the chance we want of winning the case against R.J. Morris (our power) and/or the more R.J. Morris can persuade the judge to raise the standard of evidence (the significance level), the higher the price we will have to pay for our legal representation (more study subjects).] The Appendix contains a section that translates these concepts into estimated sample sizes.

Small studies bias

In crude terms, big studies are powerful; small studies are weak. The concept of "small studies bias" illustrates the importance of having an understanding of statistical power when interpretating epidemiologic studies.

The idea behind small studies bias (Richard Peto, Malcolm Pike, et al., *Br J Cancer* 34:585-612, 1976) is that since small studies are easier to carry out than large studies, many more are carried out. Small studies that do not find a "significant" result are often not published. The journals tend not to be interested, since as explained above, there is not much information in a negative study that had low power. In fact, the investigators may not even write up the results – why not just conduct another study. In contrast, large studies are expensive and involve many investigators. Whatever the results from a large study, there is more interest on everyone's part to publish it.

To the extent that this scenario describes reality, the body of published studies contains primarily small studies with "significant" results and large studies with "significant" and "nonsignificant" results. However, if there are many small (i.e., easy, inexpensive) studies going on, then a 5% probability of making a Type I error translates into a large number of positive findings and resultant publications. Thus, many of the small studies that appear in the literature are reporting Type I errors rather than real associations.

The following example, based on randomized trials of new treatments, comes from the article by Peto, Pike, et al. Assume that there are 100 large and 1,000 small trials of treatments that are not really different, and 20 large and 200 small trials of treatments which are really different. The large trials have statistical power of 95%; the small trials have statistical power of 25%. The significance level is 5%, and only trials reporting significant results are published. These somewhat pessimistic, but perhaps very realistic, assumptions lead to the following hypothetical scenario for the number of treatment trials in progress that will be "statistically significant" ($p<0.05$):

Planned trial size	True death rate in		Postulated # of trials	Expected number to find	
	Control	Treatment		$p>0.05$	$p<0.05$
250	50%	50%	100	95 (TN)*	5 (FP)*
250	50%	33%	20	1 (FN)	19 (TP)
25	50%	50%	1,000	950 (TN)	50 (FP)
25	50%	33%	1,000	150 (FN)	50 (TP)

* TN, FP, FN, TP are for analogy with sensitivity and specificity (see below).

In this scenario, 100 small trials with "significant" results will be published, but only half of them will reflect a real difference between treatments. Peto, Pike et al.'s conclusion is to pay attention only to large trials, particularly ones that are large enough to be published even if they do not find a significant difference in treatments.

These results can be thought of in terms of the concepts of sensitivity, specificity, and predictive value. In this conceptualization, sensitivity corresponds to the statistical power to detect a true difference (95% for large trials, 25% for small trials), specificity corresponds to one minus the significance level – the probability of correctly identifying a chance result (95% specificity for a 5% significance level), and positive predictive value is the probability that a "significant" result in fact reflects a true difference in treatment effectiveness.

Large trials (e.g., 250 deaths)

	True death rate in treatment group (assuming 50% death rate in control group)		
P < 0.05	33%	50%	Total
Yes	19	5	24
No	1	95	96
Total	20	100	120

Thus, the predictive value of a p < 0.05 = 19/24 = 79%

Small trials (e.g., 25 deaths)

	True death rate in treatment group (assuming 50% death rate in control group)		
P < 0.05	33%	50%	Total
Yes	50	50	100
No	150	950	1,100
Total	200	1,000	1,200

Predictive value of a P<.05 = 50/100 = 50%

Evaluating the role of chance - interval estimation

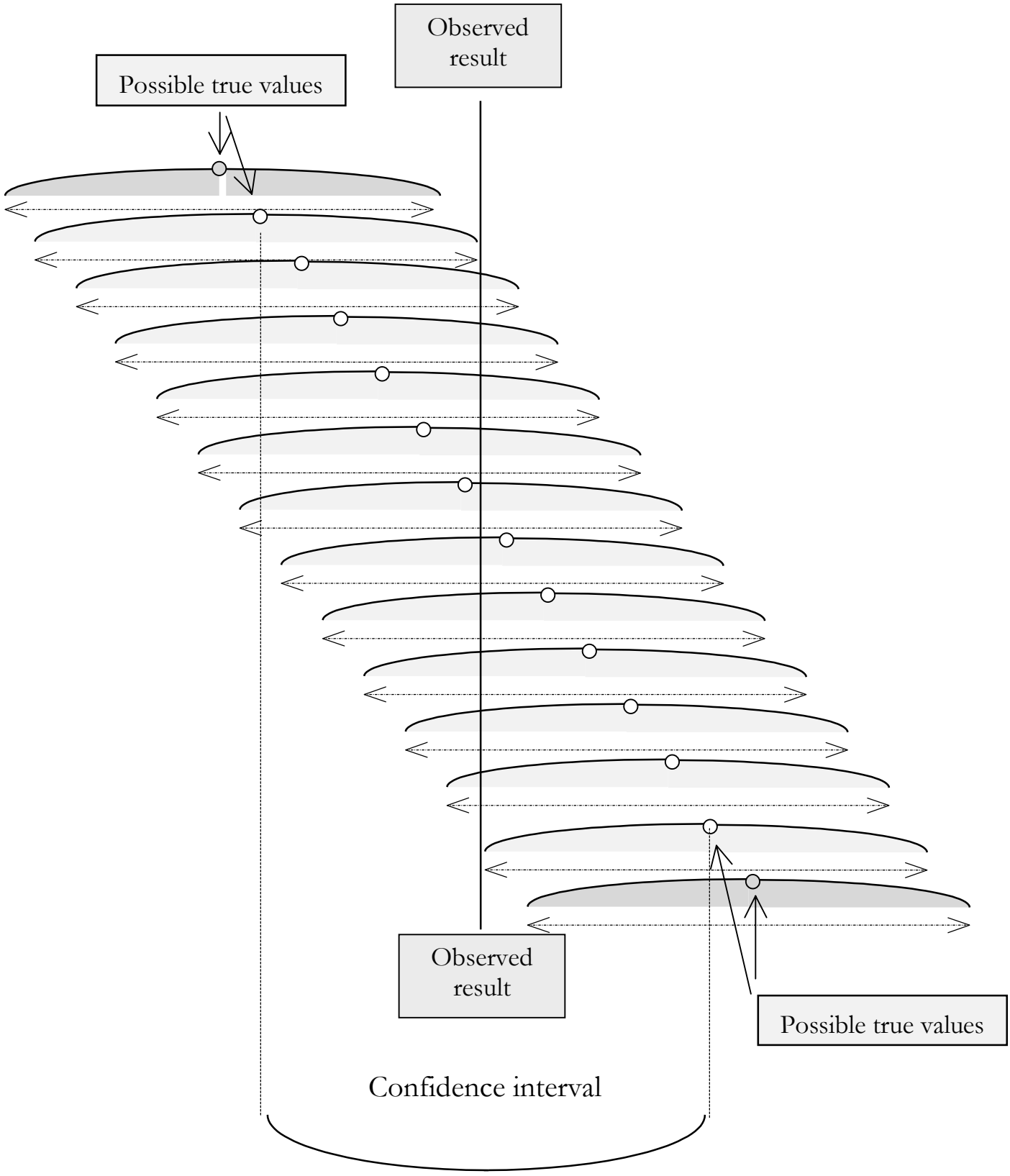
[EPID 168 students are responsible for these concepts, but not for the computations]

Statistical significance testing, with its decision-making orientation, has fallen somewhat out of favor for reporting data from epidemiologic investigations. On the premise that an epidemiologic study is essentially a measurement process (see Rothman), it is argued that the more appropriate statistical approach is one of estimation (e.g., of a measure of effect) rather than significance testing. Of course, there is still a need to quantify the role of chance, but in an estimation framework chance is quanitified by a confidence interval or confidence limits about the point estimate. Confidence limits quantify the amount of uncertainty in an estimate by defining an interval which should cover the population parameter being estimated (e.g., measure of effect) a known percentage of the time. Various authors have argued that confidence intervals are superior to p-values as a means of quantifying the degree of random error underlying an association.

Confidence intervals address the question, "what possible values for a population parameter (e.g., incidence density ratio) are consistent with the observed results?" Stated another way, "what is the range of true values which, when distorted by haphazard influences, could well have produced the observed results?" Confidence intervals provide a statement about the precision of an estimate or estimates based on the amount of data available for the estimate. If a "significant" association was not observed, then the confidence interval can give some idea of how large an association might nevertheless exist but, due to the luck of the draw, not be observed.

The nature of a confidence interval and what it does and does not provide, however, is a little tricky (judging from a discussion of confidence intervals on the STAT-L internet listserv that continued for weeks and drew a host of responses and counter-responses). The frequentist view is that a "95% confidence interval" is an interval obtained from a procedure that 95% of the time yields an interval containing the true parameter. Ideally, a 95% confidence interval would be one that "contains the parameter with 95% probability". But frequentists argue that the interval is set by the data, and the population parameter already exists in nature. The parameter is either in the interval of it is not. There is no probability about it. All that can be said is that 95% of the time the procedure will yield an interval that embraces the value of the parameter (and therefore 5% of the time the procedure will yield an interval that does not). In this view, a 95% confidence interval is like a student who typically scores 95% – the probability that he/she will give the correct answer to a question is 95%, but the answer he/she gave to any particular question was either correct or incorrect.

The concept behind the confidence interval



Computing a confidence interval for a ratio measure of effect

Introductory biostatistics courses cover the method for obtaining a 95% confidence interval for the estimate of a population proportion p. If the sample is large enough so that $np > 5$ and $n(1-p) > 5$, then the confidence limits are:

$$p \pm 1.96 \sqrt{[var(p)]}$$

$$p \pm 1.96 \sqrt{[p(1-p)/n]}$$

where p is the observed proportion, var(p) is the variance of the estimate of p (so $\sqrt{var(p)}$ is the standard error), and n is the number of observations. For a proportion, var(p) equals $p(1-p)/n$.

This method can be used to estimate confidence intervals for prevalence, cumulative incidence, and other simple proportions. Many epidemiologic measures, however, are ratios (e.g., CIR, IDR, and OR). Since ratio measures of effect have a highly skewed distribution (most of the possible values lie to the right of the null value of 1.0), the usual approach is to first estimate the confidence interval for the natural logarithm [ln(CIR), ln(IDR), or ln(OR)] and then take the anti-log (exponent) of the confidence limits:

$$95\%\ CI\ for\ ln(OR) = ln(OR) \pm 1.96 \sqrt{\{var[ln(OR)]\}}$$

$$95\%\ CI\ for\ OR = exp\{ln(OR) \pm 1.96 \sqrt{([var[ln(OR)])\}}$$

$$= OR\ exp\{\pm 1.96 \sqrt{([var([ln(OR)])\}}$$

To obtain the variance of the ln(OR), we use a simple formula (that has been derived by means of a Taylor series approximation to the ln[OR]):

$$var\{[ln(OR)] = 1/a + 1/b + 1/c + 1/d\}$$

which works well if a, b, c, d are all at least 5.

The 95% confidence interval for the ln(OR) is therefore:

$$ln(OR) \pm 1.96 \sqrt{[(1/a + 1/b + 1/c + 1/d)]}$$

and the 95% confidence interval for the OR is:

$$OR\ exp\{\pm 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}\}$$

or

$$OR\ e^{+ 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}}$$

Formulas for confidence intervals for the CIR and IDR can be found in Kleinbaum, Kupper and Morgenstern and Rothman and Greenland. Of course, if the study population is highly-selected (i.e., unrepresentative of any other population of interest), how useful is the value of the estimate?

IMPORTANT CAVEAT: Everything in this section, of course, has been based on the assumption of perfect (unbiased, independent) sampling and measurement. Anything other than an unbiased simple random sample and any error in measurement will invalidate the above at least to some extent.

Meta-analysis

Meta-analysis is a quantitative approach to summarizing and synthesizing the findings from different studies of a particular relationship of interest. Meta-analysis proceeds from the recognition that the failure to find "significant results" can be due as much to the limited statistical power of individual studies as to the absence of a relationship. Combining the information from multiple studies can yield a more precise and definitive assessment of the existence and strength of a relationship than is available from any one study or, it is argued, from a nonquantitative distillation of the literature.

There are four steps in carrying out a meta-analysis: 1) formulating the problem, 2) identifying the studies (published and unpublished), 3) coding and evaluating the studies, and 4) statistical analysis. Steps 2) and 3) are critical for the validity of the meta-analysis, since the judgments from the meta-analysis will depend upon the adequacy with which the evidence about the relationship is represented by the studies that are finally analyzed (the possibility of publication bias against "negative" studies implies that some effort should be made to locate unpublished studies). The strategy for statistical analysis can be similar to that for stratified analysis, regarding each study as a separate "stratum". More refined approaches recognize that the studies themselves can be regarded as a sample from some universe of possible studies, so that the weighting scheme needs to take into account inter-study variability as well as intra-study variability (as in the random-effects model of analysis of variance).

Interpretation of results

Key questions

1. How good are the data?
2. Could chance or bias explain the results?
3. How do the results compare with those from other studies?
4. What theories or mechanisms might account for findings?
5. What new hypotheses are suggested?
6. What are the next research steps?
7. What are the clinical and policy implications?

Bibliography

General

Ahlbom, Anders. *Biostatistics for epidemiologists*. Boca Raton, Florida, Lesis Publishers, 1993, 214 pp., $45.00 (reviewed in *Am J Epidemiol*, April 15, 1994).

Bailar, John C., III; Thomas A. Louis, Philip W. Lavori, Marcia Polansky. Studies without internal controls. *N Engl J Med* 1984; 311:156-62.

Bulpitt, C.J. Confidence intervals. *The Lancet* 28 February 1987: 494-497.

Feinstein, Alvan R. The fragility of an altered proportion: a simple method for explaining standard errors. *J Chron Dis* 1987; 40:189-192.

Feinstein, Alvan R. X and iprr: An improved summary for scientific communication. *J Chron Dis* 1987; 40:283-288.

Frank, John W. Causation revisited. *J Clin Epidemiol* 1988; 41:425-426.

Gerbarg, Zachary B.; Ralph I. Horwitz. Resolving conflicting clinical trials: guidlines for meta-analysis. *J Clin Epidemiol* 1988; 41:503-509.

Glantz, Stanton A. *Primer of biostatistics*. NY, McGraw-Hill, 1981.

Godfrey, Katherine. Comparing means of several groups. *N Engl J Med* 1985;313:1450-6.

Northridge, Mary E.; Bruce Levin, Manning Feinleib, Mervyn W. Susser. Statistics in the journal–significance, confidence, and all that. Editorial. *Am J Public Hlth* 1997;87(7):1092-1095.

Powell-Tuck J, MacRae KD, Healy MJR, Lennard-Jones JE, Parkins RA. A defence of the small clinical trial: evaluation of three gastroenterological studies. *Br Med J* 1986; 292:599-602.

Ragland, David R. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* 1992;3:434-440

Rothman - *Modern Epidemiology*, Chapters 9, 10, 14.

Schlesselman - *Case-control studies*, Chapters 7-8. (Especially the first few pages of each of these chapters).

Woolf SH, Battista RN, Anderson GM, Logan AG, et al. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol* 1990; 43:891-905.

Zeger, Scott L. Statistical reasoning in epidemiology. *Am J Epidemiol* 1991; 134(10):1062-1066.

The role of statistical hypothesis tests, confidence intervals, and other summary measures of statistical significance and precision of estimates

Allan H. Smith and Michael N. Bates. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology* 1992;3:449-452

Browner, Warren S.; Thomas B. Newman. Are all significant P values created equal? *JAMA* 1987; 257:2459-2463.

Fleiss, Joseph L. Significance tests have a role in epidemiologic research: reactions to A.M. Walker (*Am J Public Health* 1986; 76:559-560). See also correspondence (587-588 and 1033).

George A. Diamond and James S. Forrester. Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine* 1983; 93:385-394

Greenland, Sander. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421-429.

Maclure, Malcome; Greenland, Sander. Tests for trend and dose response: misinterpretations and alternatives. *Am J Epidemiol* 1992;135:96-104.

Mood, Alexander M. and Franklin A. Graybill. *Introduction to the theory of statistics*. 2ed. NY, McGraw-Hill, 1963.

Oakes, Michael. *Statistical inference*. Chestnut Hill, Mass., Epidemiology Resources, 1986.

Peace, Karl E. The alternative hypothesis: one-sided or two-sided? *J Clin Epidemiol* 1989; 42(5):473-477.

Poole, Charles. Beyond the confidence interval *Am J Public Health* 1987; 77:195-199.

Poole, C. Confidence intervals exclude nothing *Am J Public Health* 1987; 77:492-493. (Additional correspondence (1987; 77:237)).

Savitz DA, Tolo KA, Poole C. Statistical significance testing in the *American Journal of Epidemiology*, 1970-1990. *Am J Epidemiol* 1994;139:1047-.

Thompson, W. Douglas. Statistical criteria in the interpretation of epidemiologic data *Am J Public Health* 1987; 77:191-194.

Thompson, W.D. On the comparison of effects *Am J Public Health* 1987; 77:491-492.

Walker, Alexander M. Reporting the results of epidemiologic studies *Am J Public Health* 1986; 76:556-558.

Woolson, Robert F., and Joel C. Kleinman. Perspectives on statistical significance testing. *Annual Review of Public Health* 1989(10).

Sample size estimation

Donner A, Birkett N, and Burk C. Randomization by Cluster: sample size requirements and analysis. *Am J Epidemiol* 1981; 114:706

Snedecor GW, Cochran WG. *Statistical Methods*, 1980 (7th ed) see pages 102-105, 129-130 (Table A is from page 104).

Imputation

Heitjan, Daniel F. Annotation: what can be done about missing data? Approaches to imputation. *Am J Public Hlth* 1987;87(4):548-550.

Little RJA, Rubin DB. *Statistical analysis with missing data*. NY, Wiley, 1987.

Interpretation of multiple tests of statistical significance

Bulpitt, Christopher J. Subgroup analysis. *Lancet* 1988 (July 2);31-34.

Cupples, L. Adrienne; Timothy Heeren, Arthur Schatzkin, Theodore Coulton. Multiple testing of hypotheses in comparing two groups. *Annals of Internal Medicine* 1984; 100:122-129.

Holford, Theodore R.; Stephen D. Walter, Charles W. Dunnett. Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *J Clin Epidemiol* 1989; 42(5):427-434.

Jones, David R. and Lesley Rushton. Simultaneous inference in epidemiological studies. *Int J Epidemiol* 1982;11:276-282.

Lee, Kerry L., Frederick McNeer, Frank Starmer, et al. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 61:508-515, 1980.

Stallones, Reuel A. The use and abuse of subgroup analysis in epidemiological research. *Preventive Medicine* 1987; 16:183-194 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

See also Rothman, *Modern Epidemiology*.

Interpretation of "negative" studies

Freiman, Jennie A., Thomas C. Chalmers, Harry Smith, Jr., and Roy R. Kuebler. The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690-694.

Hulka, Barbara S. When is the evidence for `no association' sufficient? Editorial. *JAMA* 1984; 252:81-82.

Meta-analysis

Light, R.J.; D.B. Pillemer. *Summing up: the science of reviewing research*. Cambridge MA, Harvard University Press, 1984. (very readable)

Longnecker M.P.; J.A. Berlin, M.J. Orza, T.C. Chalmers. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 260(5):652-656. (example)

Wolf, F.M. *Meta-Analysis: quantitative methods for research synthesis*. Beverly Hills, CA, Sage, 1986.

Bias

Greenland, Sander. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; 112:564-569.

Walter, Stephen D. Effects of interaction, confounding and observational error on attributable risk estimation. *Am J Epidemiol* 1983;117:598-604.

Appendix

Estimating sample size to compare two proportions or means

(Adapted from a summary provided by of Dana Quade, UNC Department of Biostatistics, June 1984)

Let N be the number of subjects (observational units) required in **each** of two groups to be compared. Then

$$N = I \times D \times C$$

Where:

I = Intolerance for error, which depends on:

a. Alpha = Desired significance level that we want to use for our hypothesis test (e.g., 5%, two-sided)

b. Beta = Type II error (e.g.,.10 – same as 1 - power)

Formula: $I = (Z_{alpha} + Z_{beta})^2$

Z_{alpha} and Z_{beta} are, respectively, the critical values corresponding to alpha and beta from the normal distribution (see Table A on next page)

D = Difference to detect, which depends on the narrowness of the difference between the true proportions or means, in relation to the standard deviation of that difference. D can be regarded as the inverse of the "signal-to-noise ratio" – the softer the signal or the louder the noise, the more subjects needed

$$D = \frac{\text{noise}}{\text{signal}} \quad \text{OR} \quad \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \quad \text{OR} \quad \frac{2(\sigma^2)}{(\mu_1 - \mu_2)^2}$$

(for differences in proportions, where (p_1 and p_2 are the two proportions – see table on next page)

(for differences in means, where μ_2 and μ_2 are the two means, and σ^2 is the variance of the difference)

C - Clustered observations, which depends on whether observations are selected independently or in clusters.

- If all observations are sampled independently, C = 1.
- If observations are sampled in clusters (e.g., by households, schools, worksites, census tracts, etc.), then sample size must be increased to offset the fact that observations within a cluster are more similar to each other than to observations in other clusters. If rho is the intracluster correlation among observations within clusters, then:

  $$C = 1 + (m-1)\,\text{rho}$$

  where m is the average cluster size (i.e., n = km, where k is the number of clusters). C is often referred to as the "design effect". If the clusters are large or if the people in them tend to be very similar, then individual subjects contribute little information and you therefore need to study a very large number of them. If you choose "independent thinkers", you will learn more from each one.

Table A: Intolerance for error

Desired power	Two-Tailed Tests			One-Tailed Tests		
	Significance Level			Significance Level		
	0.01	0.05	0.10	0.01	0.05	0.10
0.80	11.7	7.9	6.2	10.0	6.2	4.5
.90	14.9	10.5	8.6	13.0	8.6	6.6
0.95	17.8	13.0	10.8	15.8	10.8	8.6

Table B: Difference to be detected

		p_2					
		.10	.20	.30	.40	.50	.60
p_1	.05	55	9.2	4.1	2.4	1.5	1.0
	.10	–	25	7.5	3.7	2.1	1.3
	.15	87	115	15	5.9	3.1	1.8
	.20	25	–	37	10	4.6	2.5
	.25	12.3	139	159	19	7.0	3.5

Complications

1) Unequal sample sizes

Let n be the **average** sample size = $(n_1+n_2)/2$

Let $\text{lambda}_1 = n_1/2n$, $\text{lambda}_2 = n_2/2n$ $(\text{lambda}_1 + \text{lambda}_2 = 1)$

$$D = \frac{\dfrac{p_1(1-p_1)}{2\,\text{lambda}_1} + \dfrac{p_2(1-p_2)}{2\,\text{lambda}_2}}{(p_1-p_2)^2} \quad \text{OR} \quad \frac{\dfrac{\sigma^2_1}{2\,\text{lambda}_1} + \dfrac{\sigma^2_2}{2\,\text{lambda}_2}}{(\mu_1-\mu_2)^2}$$

2) Covariables

If statistical tests are to be conducted separately within each stratum then n as determined above is required for each stratum.

If results for different strata are to be tested only for an overall average association, it is probably best not to try to allow for them in the sample size formulas explicitly, but make a modest overall increase in n.

Note: "more precise" formulas can be found in the literature, but the parameters needed for factors D and C are never really known.

Sample size for interval estimation

The tolerable width for a confidence interval can be used as the target for estimating the required sample size for a study population. Suppose, for example, that an investigator wishes to estimate the proportion (p) of condom use in a clinic population. If the investigator can obtain a simple random sample of that population, then her estimate of the proportion of condom users would be $p = u/n$, where u is the number of users in the sample and n is the size of her sample. As noted above, if $np > 5$ and $n(1-p) > 5$, then a 95% confidence interval for p is :

$$p \pm 1.96\ (1/\sqrt{[p(1-p)\ n]})$$

For example, if p is 0.50, then the confidence interval is:

$$0.5 \pm 1.96\ (1/\sqrt{[\ (0.5)(0.5)/n]}) \quad = \quad 0.5 \pm 1.96\ \frac{(0.5)}{\sqrt{n}}$$

[The square root of (0.5)(0.5) is, of course, 0.5]

Since 1.96×0.5 is approximately 1, for practical purposes this expression is equivalent to:

$$0.5 \pm 1/\sqrt{n}\ ,\ \text{so that the confidence limits are} \quad (0.5 - 1/\sqrt{n}\ ,\ 0.5 + 1/\sqrt{n}\ )$$

For example, suppose that n, the sample size, is 100. Then the 95% confidence interval around the point estimate of 0.5 is:

$$(0.5 - 1/\sqrt{100},\ 0.5 + 1/\sqrt{100})$$
$$= (0.5 - 1/10,\ 0.5 + 1/10)$$
$$= (0.5 - 0.1, 0.5 + 0.1)$$
$$= (0.4, 0.6)$$

Imprecision is often quantified in terms of the half-width of the interval, i.e., the distance between the point estimate and the interval's upper (or lower) limit, which we will refer to here as the "margin of error". The half-width of the above interval is 0.1 (i.e., the square root of n) in absolute terms or 20% (0.1/0.5) in relative terms. A 0.1 absolute or 20% relative margin of error is adequate for a "ballpark" estimate of a proportion, but not much more.

Since the above expressions involve the square root of the sample size, progressive narrowing of the interval width involves substantially greater increases in sample size. For example, to obtain a 0.05 absolute or 10% relative margin of error, sample size must be quadrupled, to 400:

$$(0.5 - 1/\sqrt{400},\ 0.5 + 1/\sqrt{400})$$
$$= (0.5 - 1/20,\ 0.5 + 1/20)$$
$$= (0.5 - 0.05,\ 0.5 + 0.05)$$
$$= (0.45,\ 0.55)$$

Similarly, a sample size of 900 yields confidence limits one-third as wide as from a sample of 100, a sample of 2,500 yields limits one-fourth as wide as for n=100, etc.

These numbers apply to a point estimate of 0.5, which produces the widest error margin in absolute terms. A smaller or greater point estimate will have a narrower (in absolute terms) interval, because the square root of $p(1 - p)$ cannot exceed 0.5 (try it! – or use calculus). The relative margin of error, on the other hand, is inversely related to the size of the point estimate. Examine the following table:

Point estimate	Sample size	Margin of error (rounded) Absolute *	Relative ** (%)
0.1	100	0.06***	60***
0.2	100	0.08	40
0.3	100	0.09	30
0.4	100	0.096	24
0.5	100	0.10	20
0.6	100	0.096	16
0.7	100	0.09	12
0.8	100	0.08	9.8
0.9	100	0.06	6.5
0.1	400	0.03	30
0.2	400	0.04	20
0.3	400	0.045	15
0.4	400	0.048	12
0.5	400	0.05	10
0.6	400	0.048	8.0
0.7	400	0.045	6.4
0.8	400	0.04	4.9
0.9	400	0.03	3.2

* Approximate half-width of 95% confidence interval in absolute terms

** Approximate half-width of 95% confidence interval in absolute terms, relative to the size of the point estimate

*** Calculation: $1.96\ (1/\sqrt{[(0.01)(1 - 0.01) / 100]}) = 1.96\ (0.03) = 0.0588 \approx 0.06$ absolute error margin

This table illustrates that:

1. quadrupling sample size halves the margin of error.
2. absolute error margin decreases as the point estimate moves away from 0.5
3. relative error margin is inversely – and very strongly – related to the size of the point estimate

For very small point estimates, as illustrated in the following table, very large samples are required to obtain a small relative margin of error. Even a sample size of 2,500 still produces a relative error margin of 17% for a proportion of 0.05.

Point estimate	Sample size	Margin of error (rounded) Absolute *	Relative * (%)
0.5	100	0.10	20
0.5	400	0.05	10
0.5	900	0.033	6.6
0.5	1,600	0.025	5.0
0.5	2,500	0.020	4.0
0.05	100	0.043	85
0.05	400	0.021 **	43 **
0.05	900	0.014	28
0.05	1,600	0.011	21
0.05	2,500	0.009	17

* See previous table

** Calculation: $1.96 \times (1/\sqrt{[(0.05)(0.95)/400]}) = 1.96 \times 0.0109$

$= 0.0214 \approx 0.021$ absolute error margin

Relative $= 0.0214 / 0.05 = 0.427 = 42.7\%$ (approximately 43%)

Recall that this formula requires that $nP \geq 5$, which is just met for P=0.05 and n=100.

How large a sample is large enough? If the objective is to set an upper or lower bound on a proportion, then a small absolute margin of error may suffice. For example, if one is testing for hepatitis C antibody and wants to be reassured that the seroprevalence is below 5%, then a sample size of 900 will produce an interval with an absolute error margin no wider than 0.033 (for a point estimate of 0.5 – see above table) and more likely 0.011 (for a point estimate of 0.05) or smaller. Since we expect the seroprevalence to be very small, then the 0.011 is much more relevant than the 0.033. If when we carry out the study we obtain a point estimate of exactly 0.05, then the 95% confidence interval will be (0.039,0.061) which will tell us that the true value is at least not likely to be greater than 6%. If the point estimate is below 0.04, then the upper confidence limit will be below 5% and we are reassured that the seroprevalence is no greater than that value.

Note that the above is all based on the assumption of perfect (unbiased) simple random sampling and measurement. Anything other than an unbiased simple random sample and any error in measurement will invalidate the above at least to some extent.

Meditations on hypothesis testing and statistical significance

The statistical theory of hypothesis testing and assessment of statistical "significance" proceeds from an analysis of decision-making with respect to two competing hypothesis: a "null" hypothesis and an alternative hypothesis. Two types of errors are possible:

Type I: Erroneously reject the "null hypothesis" (H_0), in favor of the alternate hypothesis (H_A), i.e., erroneously reject chance as a sufficient explanation for the observed results.

Type II: Erroneously fail to reject H_0, i.e., erroneously accept chance as an explanation. [A parallel dichotomy will be seen later in the course when we discuss sensitivity and specificity.]

Traditionally, the Type I error probability has received more attention and is referred to as the "significance level" of the test. The Type I error presumably owes its prominence to the scientific community's desire to avoid false alarms, i.e., to avoid reacting to results that may readily have been chance fluctuations. Also the Type I error probability is easier to estimate, since the Type II error probability depends on stating the size of true difference that one seeks to detect.

During recent decades, the calculation and presentation of "p-values" (which give information about the Type I error probability) have become *de rigeur* in the empirical scientific literature. Indeed, a significant (!) number of people refuse to pay any attention to results that have p-values greater than .05 (5% probability of a Type I error).

Such a stance is a considerable labor-giving device, but is perhaps a bit brutal. After all, a result with a p-value of .10 would result from a chance process in only one in ten trials. Should such a finding be dismissed? Moreover, since the p-value reflects the number of subjects as well as the size of the observed difference, a small study will have very small p-values only for large (and perhaps unrealistic?) observed differences. If the size of the observed difference is unreasonably large, then we may be suspicious of the finding despite a small p-value. If the observed difference is plausible, but because the study is small the p-value is "not significant", we may nevertheless want to pay some attention.

Another reason for a reflective, rather than a reflexive, approach to p-values (and statistical inference generally) is that the probability estimates themselves are accurate only with respect to the models that underlie them. Not only may the mathematical models not adequately capture the real situation, but the context in which they are used clouds the issue. One critical assumption is that of random sampling or randomization (as in a randomized controlled trial). Although this assumption is the basis for virtually all of the statistical theory of hypothesis testing and confidence intervals, it is rarely met in observational studies and the limitations that it imposes on the interpretation of statistical tests are often underappreciated (Greenland S. Randomization, statistics, and causal inference *Epidemiology* 1990;1:421-249).

Even in randomized trials problems of interpretation exist. For example, the p-value for a single result in a single study may be 5 percent. But that means that 20 independent studies of two identical phenomena would observe, on the average, one difference that was "significant" at the five percent level. A prolific investigator who conducts 200 studies in his/her professional life can expect to have ten that are "significant" by chance alone. Moreover, a study will often examine multiple outcomes, including multiple ways of defining the variables involved.

Such "multiple comparisons" increase the likelihood of chance differences being called "significant". But the statistical procedures for dealing with this "significance inflation" tend, like measures to suppress price inflation or grade inflation, to produce recession or even depression [of study findings]. Should an investigator be required to take an oath that he/she had (1) fully specified an a priori hypothesis, including the procedures for defining and manipulating all variables, decisions about all relationships to examine, what factors to control, etc; (2) proceeded directly to the pre-specified statistical test without looking at any other data; and (3) will not perform any further statistical tests with the same data? (See Modern Epidemiology for more on these points.)

What about so called "fishing expeditions" in which an investigator (or her computer) pores over a dataset to find "significant" relationships. Is such a procedure better characterized as "seek and ye shall find" or as "search and destroy"? Some analysts recommend adjusting the significance level to take account of such "multiple comparisons", but an energetic investigator can easily perform enough tests so that the adjusted significance level is impossible to satisfy. Other writers (e.g., Rothman, Poole) assert that no adjustment is required – that once the data are in, the number of tests is irrelevant. Others (e.g., Greenland) have proposed more sophisticated approaches to adjustment. Perhaps the best course at this time is twofold:

(1) If you are conducting a study, for example, a randomized trial, in which you have a good chance of satisfying the assumptions for a statistical hypothesis test and are hoping to test a specific hypothesis, especially one that may lead to some decision, then it is probably better to adhere to the Neyman-Pearson hypothesis testing format as much as possible. This approach ensures maximum impact for your results;

(2) If you are conducting an inquiry with few of the above characteristics, or have already completed the a priori hypothesis test, analyze all that you like but be candid in describing how you proceeded. Then readers can interpret as they judge most appropriate.

Apparent (calculated) power is rarely achieved because it often assumes no errors in classification of subjects. A study with advertised power of 90% could well have much less probability of detecting a given true difference because of dilution by information bias. Similarly we can in principle improve the effective power of a study if we can increase the precision with which important variables are measured.

Louis Guttman has written that estimation and approximation, never forgetting replication, may be more fruitful than significance testing in developing science. [Louis Guttman. What is not what in statistics. *The Statistician* 25(2):81-107.]

Independent replication is the cornerstone of scientific knowledge.

Bayesian approach to p-value interpretation

The application of the concepts of sensitivity, specificity, and predictive value to interpreting statistical hypothesis tests suggests an analogy between statistical tests and diagnostic tests (see Browner and Newman, 1987; Diamond and Forrester, 1983; and Feinstein, *Clinical Biostatistics*). Just as the interpretation of a diagnostic test depends upon disease prevalence (the "*a priori* likelihood that the patient has the disease"), the interpretation of statistical tests can be regarded as dependent upon "truth prevalence", i.e., on the reasonableness of the hypothesis.

As noted earlier, we would like statistical inference to provide an estimate of the probability that a hypothesis of interest (H) is true given the observed results. The p-value provides instead the probability of observing an extreme result under a null hypothesis (typically the inverse of the hypothesis of interest). The Bayesian approach to interpreting p-values tries to provide an answer that comes closer to the original objective. In the Bayesian approach, we begin with a prior probability for the truth of the hypothesis and then adjust that probability based on the results of a study, to obtain a posterior probability. The effect that the study results have on our assessment of the credibility of the hypothesis depends on our original assessment of its credibility.

Let T mean that a statistical test is "significant". According to Bayes Theorem, if Pr(H) is the "*a priori*" probability of H, i.e., the probability that H is true based only on previous information, then the *a posteriori* probability of H (the probability that H is true based on previous information and the current test result) is:

$$\Pr(H|T) = \frac{\Pr(H)\ \Pr(T|H)}{\Pr(H)\ \Pr(T|H) + \Pr(h)\ \Pr(T|h)}$$

[where Pr(T|h) means the probability of a positive test given that the hypothesis is not true] which can be written:

$$\Pr(H|T) = \frac{\Pr(H)\ \Pr(T|H)}{\Pr(H)\ \Pr(T|H) + [1 - \Pr(H)]\ \Pr(T|h)}$$

Since Pr(T|H) is the statistical power (the probability of a positive test given that the hypothesis is true) and Pr(T|h) is the p-value (the probability of a positive test given that the hypothesis is not true), the posterior probability can be written:

$$\Pr(H|T) = \frac{\Pr(H)\ \text{(power)}}{\Pr(H)\ \text{(power)} + [1 - \Pr(H)]\ \text{(p-value)}}$$

Pr(H|T) is therefore a function of the "*a priori*" probability of the hypothesis, the statistical power, and the p-value. Therefore the p-value has more impact on Pr(H|T) when Pr(H) is small (i.e., when a hypothesis is not supported by prior research or laboratory data) (see Diamond and Forrester).

To get an idea of how these formulas work with typical values for the various elements, take a look at the following table:

Evaluation of posterior probability based on prior probability, statistical power, and p-value

	Prior probability (Before the study) Pr(H)	Statistical power of the study Pr(T|H)	P-value (Findings of the study) Pr(T|h)	Posterior probability (After the study) Pr(H|T)	
Credible hypotheses	0.60	0.8	0.100	0.92	
	0.60	0.8	0.050	0.96	High power
	0.60	0.8	0.001	1.00	
	0.60	0.5	0.100	0.88	
	0.60	0.5	0.050	0.94	Low power
	0.60	0.5	0.001	1.00	
Long shot hypotheses	0.05	0.8	0.100	0.30	
	0.05	0.8	0.050	0.46	High power
	0.05	0.8	0.001	0.98	
	0.05	0.5	0.100	0.21	
	0.05	0.5	0.050	0.34	Low power
	0.05	0.5	0.001	0.96	

In this table, for example, a very strong p-value (e.g., 0.001) gives high credibility (posterior probability) even for a long shot hypothesis examined in a study of low statistical power. A p-value that is "just significant", however, does not make a hypothesis highly credible unless it was judged more likely than not before the study. Even a "nonsignificant" p-value (e.g., 0.10) provides some increase in credibility of the hypothesis, so in the Bayesian framework a p-value of 0.10 would not be regarded as a "negative" result casting doubt on the existence of an association. Meta-analysis, in which results are combined across studies to obtain a quantitative assessment of an association from the full body of evidence, also takes into account evidence for the association from studies that observed an association but had a p-value greater than 0.05. Formal use of Bayesian methods in everyday work, however, is somewhat constrained by the absence of an obvious method for obtaining a prior probability.

More meditations on interpreting statistical significance tests

Some of the concepts in the interpretation of statistical tests of significance can perhaps be illustrated through an example based on one glorious origin of probability theory – games of chance. Suppose that our friend tells you that he has an intuition about roulette wheels. By watching the operator spin the wheel, your friend can, he claims, predict where the ball will land within a very small margin. If, for simplicity, the wheel has numbers 1-100 on it, your friend says he can predict the numbers where the ball will land. He wants you to put up some money to send him to Monte Carlo to make our fortunes.

Naturally you're excited by the prospect of instant riches but also a bit skeptical. To verify your friend's claim, you undertake a statistical test. You give your friend $5 to prove his prowess at the local gambling casino, and you wait to see how he does.

The null hypothesis for your statistical test is that your friend has no special ability, so that his chances of predicting the resting place of the ball on any one try are simply 1 out of 100 (.01). The 1-sided alternate hypothesis is that your friend does have this ability and can predict the correct number more often than 1 out of 100 times. [The 2-sided alternate hypothesis is that your friend will predict the resting place of the ball either more than would be expected by chance or less than would be expected.]

Your friend returns with $400. Knowing that the probability of his being correct on a given try by chance alone was only 1%, your are impressed. His performance was "significant at the .01 level"! Do you underwrite his trip to Monte Carlo? How do you interpret his correct prediction?

Is it correct to say that there is only a 1% chance that his accurate prediction was due to "luck"? Not quite. According to the frequentist interpretation, the prediction was made and the roulette wheel has already been spun. The accuracy was due either to "chance" ("luck") or your friend's ability, but only one of these was actually responsible that time. So the probability that his correct prediction was due to chance is either zero (i.e., your friend can predict) or one (your friend cannot predict). The only trouble is, you don't know which!

You can say (before the wheel was spun and assuming it was a balanced wheel) that if your friend had no special ability there was only a one percent probability of his making a correct prediction and that therefore his winning is evidence against the null hypothesis (of no ability) and in favor of the alternate hypothesis (ability to predict). If you have to decide that day, you might figure that it would be worth underwriting his trip to Monte Carlo, but you would be aware that his correct prediction could have been due to chance because there was a one percent probability that in the absence of any clairvoyance his prediction would have been correct (not quite the same as a one percent probability that his correct prediction was due to chance). So you give your friend $2,000. He thanks you profusely, and in parting, tells you that it actually took him 30 tries to make a correct prediction – he borrowed the money for the other 29 tries.

That information gives you pause. Certainly you would not have been so impressed if he had told you he could make a correct prediction in 30 tries. If the probability of a correct prediction (i.e., a correct guess) in the absence of any special ability is 0.01, then the probability of one or more correct guesses in 30 tries is 0.26 (1.0 minus the quantity 0.99 raised to the 30th power). Twenty-six percent is still less than 50 percent, i.e., the probability of winning a coin flip, but not so impressively. The evidence against the null hypothesis is now not nearly so strong. This change in your interpretation illustrates the issue that arises in connection with multiple significance tests and small studies bias.

It is possible, using statistical theory, to adjust significance levels and p-values to take into account the fact that multiple independent significance tests have been done. But there are various practical problems in applying such procedures, one of which is the lack of independence among multiple tests in a particular set of data. For example, if your friend explained that he so rarely makes an incorrect prediction that when he did he became so upset that it took him a whole hour (and 29 more predictions) to regain his predictive ability, then even if you remained skeptical you would be hard-put to calculate an adjusted p-value for your test if you thought he was telling the truth. Similarly, in a given dataset, does the fact that an investigator tested the same difference in various ways (e.g., obesity as indexed by weight/height2 [Quetelet's index], weight/height3 [ponderal index], percent above ideal weight, skinfold thickness, and body density) weaken the findings for each test? If she also looked at blood pressure differences, would that weaken the credibility of statistical significance of differences in obesity?

"You pays your money, and you takes your choice."

Data analysis and interpretation - Assignment

Part I

The following questions are based on Rosenberg et al., "Oral contraceptive use in relation to nonfatal myocardial infarction". *Am J Epidemiol* 1980; 111:59-66.

1. Control for menopausal status appears to have been accomplished through: (Choose one)

   A. Restriction

   B. Matching plus stratified analysis

   C. Stratified analysis without matching

   D. Mathematical modeling (logistic regression)

2. Using the data in Table 1, label and complete a 2x2 table for the crude (i.e., not stratified by age of hospitalization) odds ratio for MI and current (versus "never") OC use.

3. Compute the odds ratio for the above table (show all work).

4. Is age a confounder of the relationship between MI risk and current OC use, based on the data in Table 1? Justify your answer (1-3 sentences), referring to specific measures or estimates in the data.

5. Is age an effect modifier of the relationship between MI risk and current OC use, based on the data in Table 1? Justify your answer (1-3 sentences).

6. Based on the data in Table 4:

   a. Is MI associated with cigarette smoking? Give a sentence to support your answer (yes, no, cannot determine from the data in Table 4).

   b. Is the association between MI and the combination of smoking and hypertension greater than or less than the association between MI and hypertension alone? Give a sentence to support your answer (greater than, less than, cannot determine from the data in Table 4).

7. Using the data in table 4, create and label a 2x2 table relating current OC use and hospitalization for MI among nurses who have no history of hypertension, regardless of their smoking status.

8. What is the relative risk estimate for women who have all three characteristics (OC, CIG, HYP) compared to women who have none?

9. Assuming that the relative risk estimates in table 4 are precise (i.e., ignoring the variability from small cell sizes), what would be the relative risk estimate for women who have all three characteristics (OC, CIG, HYP) compared to women with none under a multiplicative model? Show your work.

Part II

1. In table 2 from the Rosenberg et al. study (p. 62), which variables are not found to be risk factors for MI?

2. Show the relationship between the logistic regression coefficient for Current OC use and the relative risk estimate for that factor.

3. Has age been controlled in this logistic model?

4. Comparing the information in Table 2 with that in Table 1, do you see evidence of confounding in Table 1 with respect to the relationship between MI risk and OC use? Briefly discuss, citing the most relevant measures or statistics from the two tables.

5. What is peculiar about this logistic regression model, in terms of the form of the variables in it?

6. What provision has been made in this model for possible statistical interaction, i.e., deviation from a multiplicative model?

7. Based on the logistic model shown, what would be the odds ratio for the combination of both cigarette smoking and Current OC use (versus neither)? Compare this result to the corresponding figure(s) in the stratified analysis in Table 4 and suggest possible explanations for the difference, if any.

Data analysis and interpretation - Assignment solutions

Part I

1. (a) Restriction

   The methods for this paper were included in the Sources of Bias assignment, so we need to look at that paper to find the answer. The 3rd paragraph of the Methods section (page 118) says: "One hundred and fifty-six respondents reported having been hospitalized for MI before their menopause …. For each of these case, we selected 20 control subjects from respondents … and who were premenopausal at the time of hospitalization of the case." Thus, it appears that both cases and controls were premenopausal at least to the time that the MI occurred or the comparable date for the matched controls. The controls were matched on several factors, including being premenopausal. But because ONLY premenopausal women were studied, the method is Restriction (to one level of the variable) rather than matching (enforcing the same distribution of the matching variable) and stratified analysis (which involves dividing the dataset into strata, not collecting data from only one stratum).

2.

	OC use	
	Current	Never
MI	(16 + 7) = 23	(42 + 53) = 95
No MI	(190 + 114) = 304	(991 + 1045) = 2036

3. $$OR = \frac{(23)(2036)}{(95)(304)} = 1.62$$

4. Age (at hospitalization) is not a confounder:

   Crude OR = 1.6

   Stratum-specific OR's are 2.0 and 1.2, so that the crude lies well within their range.

5. The concept of effect modification can be approached from different perspectives. One perspective is to regard effect modification as a departure from a multiplicative model, since the multiplicative model is most often employed in investigations of etiology and, from a practical standpoint, departure from multiplicativity means that a weighted average of stratum-specific ratio measures of effect (e.g., OR's) may be misleading. This example is complicated by the fact that controls were matched to cases on age, so that the effect of age cannot be evaluated from the data presented in the paper. However, homogeneity across strata of age can be examined. We already know, of course, that MI rates increase sharply with older age.

If the OR for Current OC use were the same in the two age strata, then we could conclude that the observed odds ratios fit a multiplicative model, so that there would be no effect modification based on the above perspective. However, the OR_{OC} for the older women is smaller than the OR_{OC} for the younger women. That suggests that the combined effect of current OC use and greater age is less than would be expected based on a multiplicative model, which could be interpreted as evidence of effect modification. The evidence is weak, however, since although confidence intervals are not presented, the OR estimates are based on rather small numbers of exposed cases and are therefore imprecise. Unless a statistical test for heterogenity of the OR across strata indicated that the observed difference in the OR's (1.2 versus 2.0) is beyond that expected from chance alone, one would say that there is, at most, slight evidence for effect modification.

The other perspective on effect modification relates to impact, i.e., that if the combined effect is greater than expected from an additive model, then interventions may be worth targeting to those dually exposed. This perspective cannot be fully investigated in the data we have here, because of the matching. But since the combined effect is less than expected based on a multiplicative model, the combined effect is presumably not much greater than expected based on an additive model.

6. (a) Yes: OR for CIG only is 5.0 (2.7-9.0).

   (b) Greater than: OR for CIG + HYP = 8.9, compared to OR for HYP only = 7.6. But the confidence intervals are broad and have substantial overlap, so "cannot determine" is also a reasonable conclusion.

7. Hospitalization for MI and OC use in nurses with no history of hypertension.

	OC use	
	Current	Never
MI	(5 + 7) = 12	(12 + 39) = 51
No MI	(150 + 107) = 257	(1022 + 669) = 1691

$$OR = \frac{(12)(1691)}{(51)(257)} = 1.55$$

8. From last line of the table: 170 (31 - 1100)

9. $OR_{CIG,HYP,OC}$ [multiplicative model]

$$= OR_{CIG|HYP,OC} \times OR_{HYP|CIG,OC} \times OR_{OC|CIG,HYP}$$

$$= 5.0 \times 7.6 \times 2.8$$

$$= 106$$

(This is less than the observed OR.)

Part II

1. Past OC use (Relative risk estimate 0.9) and Overweight (Relative risk estimate 1.2) are both not importantly related to MI risk. Though an OR of 1.2 indicates some elevation of risk, the confidence interval extends so far below 1.0 that the elevation is consistent with an interpretation in terms of chance.

2. The relationship between the coefficient for Current OC use and the relative risk estimate is:

   Relative risk estimate = OR = exp(0.59) = $e^{0.59}$ = 1.8

3. Age has not been controlled in this logistic model. The cases and controls were, however, matched by year of birth. It is not clear that this matching eliminates possible confounding by age. Nevertheless, from Table 1, there does not appear to be confounding by age, and while it is theoretically possible to have confounding in the multivariable analysis even though none was observed in the stratified analysis of Table 1, that likelihood is probably small.

4. The crude OR from Table 1 is 1.6; the summary OR (controlling for age) is also 1.6. there may be a small amount of confounding caused by the other risk factors, therefore, since the OR from the multiple logistic model is 1.8 for Current OC use. But the difference between 1.6 and 1.8 is not important.

5. This logistic model consists entirely of indicator (dichotomous) variables. In part, this fact was necessitated by the study questionnaire, which asked for history of various conditions, rather than their actual values (e.g., blood pressure). Overweight could presumably have been entered as a continuous variable. Using a single indicator variable to express the value of a continuous variable loses information. There are some offsetting advantages, however.

6. There are no interaction (product) terms in this model, so no provision has been made for deviation from the underlying model that the odds ratio for a combination of factors equals the product of their respective odds ratios (or equivalently, that the logarithm of the odds of MI equals the sum of the logarithms of the odds ratios for the factors, plus a constant), and that this relationship is not altered by the value of other variables in the model.

7. The odds ratio for the combination of cigarette smoking and Current OC use (compared to neither factor) is, since there are no product terms to consider, simply the product of the odds ratios (relative risks) for each of these two factors:

$$OR_{cig,OC} = OR_{cig} \times OR_{OC} = 2.8 \times 1.8 = 5.04 \approx 5.0$$

This OR is close to the value in Table 4 for "OC and CIG only" (5.6) for normotensive individuals. For hypertensives, the estimate for the joint effect of OCs and cigarettes is obtained by dividing 170 (the OR for "OC, CIG, and HYP") by 7.6 (the OR for "HYP only"); among hypertensives, therefore, the OR for OCs combined with cigarette smoking is 170 / 7.6 = 22.4, a value well above the 5.0 estimate from the model. Since there are no product terms involving hypertension, the logistic model assumes that the OR for each factor or combination of factors is unaffected by hypertension. In other words, the OR from this logistic model represents a type of average of the OR in normotensives and that in hypertensives.

The difference between the value from the logistic model (5.0) and the values from the stratified analysis (5.6 and 22.4) can be attributed to the "smoothing effect" of the logistic regression, which forces all odds ratios to fit the form of the assumed model (of multiplicative odds ratios with no heterogeneity). From the figures in Table 4, it is clear that most of the cases and controls were not hypertensive, so the logistic model odds ratio estimates will primarily reflect the odds ratios in normotensives. Hence the logistic value is closer to the 5.6 than to the 22.4.

Another possible reason for the difference between the stratified and logistic regression odds ratios is that the latter control for a variety of other risk factors that are not included in the stratified analysis. If these other factors confound the OC and cigarette smoking relationship with MI risk, then the stratified analysis results in Table 4 may be confounded.

15. Practical aspects of epidemiologic research

Epidemiology in the "real world": the practice of epidemiology and its institutional environment -- funding, logistics, collaborations, peers, publication, publicity, politics and policy, study conduct, data management.

Natural history of a study

1. Develop an idea for a study, see an RFA, be invited to work on a proposal, . . .
2. Explore the literature, talk with others, brainstorm
3. Develop the rationale and specific aims/questions
4. Design the study - architecture, setting, study population, eligibility criteria, measures, analysis
5. Arrange collaborations, secure access to needed resources
6. Prepare proposal and submit for review (human subjects and scientific) / approval / funding
7. Obtain resources - funds, release time, space, personnel, equipment, subcontracts, consultation, advice, and assistance
8. Create a management structure, timetable, workplan, communication infrastructure, quick reference resources, documentation procedures, filing systems
9. Identify measurement instruments, analytic procedures, etc.
10. Develop data collection protocol, forms, confidentiality, training
11. Obtain human subjects (IRB) approval for data collection instruments and procedures
12. Pretest questionnaires and data collection forms, and revise
13. Create tracking system(s) for subjects and forms
14. Develop data filing systems - manual and electronic
15. Design a system for data linkage (ID and numbers)
16. Arrange contemporaneous monitoring of process and output, with feedback to data collectors, including quality control measures
17. Pilot test questionnaires, data collection forms, and procedures
18. Modify instruments and procedures
19. Obtain IRB approval for revised data collection instruments and procedures
20. Schedule and train data collection personnel
21. Collect data, monitor the activity with frequent written reports
22. Review completed forms for completeness, consistency, and accuracy
23. Make modifications and provide feedback to get back on track

24. Submit manuscript from previous study
25. Develop formal edit specifications and coding rules, pilot test them, and implement
26. Computerize data
27. Compile Data Management Manual
28. Explore the data
29. Create data files for each data stream, create analysis files
30. Prepare an accounting for all data, check the N's very carefully!
31. Carry out preliminary analyses to inform planning and to look for big surprises
32. Write next grant proposal
33. Clean and summarize data
34. Create analysis variables and scales
35. Write a descriptive report
36. Address the research questions
37. Control for potential confounders, effect modifiers, other extraneous factors
38. Bring documentation up-to-date
39. Write up results of analyses
40. Write conclusion and introduction
41. Submit request for no cost extension
42. Fill in missing analyses
43. Complete manuscript and/or report
44. Arrange for storage for data, analyses, and documentation and/or make data and documentation available for use by others
45. Write and submit final report to funding agency

Funding an epidemiologic study

Most epidemiologic studies of any size (.e.g., costing $20,000 or more) conducted by independent agencies (e.g. universities, research institutes) are funded through research grants or contracts. The large majority of these are awarded by federal agencies, particularly institutes within the National Institutes of Health (NIH).

Major NIH agencies that fund epidemiologic research include the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), the National Institute of Allergy and Infectious Disease (NIAID), the National Institute of Environmental Health Sciences (NIEHS, located in Research Triangle Park), the National Institute of Child and Human Development (NICHD), the National Institute on Alcohol Abuse and Alcoholism (NIAAA), the National Institute on Drug Abuse (NIDA), and the National Institute of Mental Health (NIMH)). Other agencies of particular interest to epidemiologists seeking funding are the Centers for Disease Control

(CDC, which includes the National Institute of Occupational Safety and Health [NIOSH] and the Agency for Toxic Substances and Disease Registry [ATSDR]), the Environmental Protection Agency (EPA), and the Agency for Health Care Policy and Research (AHCPR).

Types of federal funding mechanisms

Grant applications to NIH are submitted through several funding mechanisms. The first mechanism is unsolicited investigator-initiated proposals (often called "R01's", since the application and grant, if awarded, will be assigned a number that begins R01-). Here the investigators develop a proposal on their own initiative and submit it on the hope (preferably with some informed judgment and informal advice) that an institute will have some interest in the proposed research.

Program announcement

Agencies often issue program announcements (PA's) describing particular interest areas and/or types of application the agency would like to receive. These announcements may or may not designate an amount of funding available. They usually do not involve a special review process, but they may request a Letter of Intent prior to the submission of the application.

The recipient of a grant award made in response to a program announcement has a considerable degree of lattitude in carrying out the research, subject to overall responsibility for the general scientific conduct of the study and the accurate accounting for all monies expended within the budgeted categories. Funds can generally be shifted between budget categories and other adjustments made, and research objectives can be modified (in consultation with the granting agency project officer) if necessary.

Request for applications

A second funding mechanism is proposals submitted in response to a Request for Applications (RFA). RFA's describe a specific research area in which the agency wants to generate research. The difference between RFA's and program announcements is that an RFA usually identifies specific funds for the successful applications. In addition, applications in response to the RFA may be reviewed by a special review group, chosen to be knowledgeable or sympathetic to the kinds of research solicited.

The RFA may include special requirements for applying or for eligibility for funding. For example, a Letter of Intent may be requested or required; a certain number or types of components may be required to be present in the proposed study. The RFA may specify a special deadline other than the usual review dates.

Within these guidelines, however, the investigator has complete freedom in the type of study and study population he/she proposes, the specific hypotheses to be tested, the manner of carrying out the investigation, and so forth. Should a grant be awarded, the investigator has the same lattitude as for other investigator-initiated research proposals. There is somewhat less freedom than in the first mechanism in that the agency has defined explicit goals and guidelines for the research proposals. But some of the guessing-game of what research would the agency like to fund has been eliminated

Cooperative agreement

A cooperative agreement is a grant where it is anticipated that there will be considerable interaction between the grantee and the funding agency, and often with other grantees who are part of the same cooperative agreement. This mechanism can provide considerable flexibility for both the grantee and the granting agency. The level of involvement of the granting agency can vary considerably, but often includes measures designed to achieve greater uniformity across multiple studies (e.g., common questionnaire items, uniform procedures for data collection, joint analyses).

An important feature of a cooperative agreement is that the funding agency may be able to redefine the goals and objectives, and other major aspects of the activity, in midstream, even terminating the agreement if it decides that other needs are of higher priority. Often a steering or executive committee is created from among principal investigators and NIH representatives to make decisions concerning the direction of a funding program.

Request for proposals

A request for proposals (RFP) differs substantially from the above mechanisms. An RFP is a solicitation for contract research rather than simply for the purpose of encouraging research in a particular area.

In issuing an RFP, the agency has decided that a particular study or studies are needed and has already decided the general framework of the study design. The investigators still have latitude in proposing how they would do the requested study or project, but much of this latitude will be directed toward interpreting and designing an approach to meet the specific objectives and criteria that have already been set out in the RFP.

A specific amount of money or other indication of resources available is usually given. If the agency likes your proposal but thinks it can persuade you to do it for less money, the agency may negotiate with you to reduce your budget.

If you receive an award under this mechanism, you will need to sign a contract specifying in some detail what you will provide to the sponsoring agency and when (a schedule of "deliverables"), and certain other conditions from the RFP. If you do not meet those conditions, or if the reports and "deliverables" you provide are not regarded as acceptable, your contract may be terminated.

RFP's are for the procurement of research. The data may belong to the funding agency.

Obviously, investigators generally prefer grants to cooperative agreements and contracts. There is much sentiment in favor of increasing the amount of government funding going to research project grants, which provide the greatest opportunity for investigators to pursue research of their own choosing.

Contracts need not be undesirable, however, particularly if the investigators are already interested in or would like to gain experience in an area of activity. Moreover, much of the major research in the

cardiovascular area -- notably the large collaborative randomized trials (HDFP, MRFIT, CPPT) and the ARIC study -- has been conducted through contract mechanisms.

Contract officers are generally understanding of difficulties that may arise in carrying out a project, and can be strong allies of the researchers. The contract officer's responsibility is to represent the interest of the agency in seeing the results of the project completed and of high quality. If you are working toward that end, you can generally expect a cordial relationship.

Access to data collected with federal funding

In October 1998, the U.S. Congress passed a law providing for access under the Freedom of Information Act (FOIA) to data collected with federal funds. The legislation stimulated considerable unease in the scientific community, and the Office of Management and the Budget (OMB) received numerous comments in response to the draft regulations for implementing the law. OMB published a reviewed draft in August 1999 and in October issued final regulations regulations, so that from this point forward (it is not clear whether or not the requirement applies retroactively) any investigator who receives federal funding to collect data (even if the federal funds represent only a portion of the total cost) can be required to provide that data under FOIA (FOIA contains certain provisions for the protection of privacy and proprietary material, though whether investigators will be regarded as having a proprietary interest research data they have been collected is an open question).

Research supplements for underrepresented minorities and for disabled scientists

Another advantage of a grant or cooperative agreement is the opportunity to apply for supplements. Competitive supplements, where additional funds are requested through a proposal that goes through peer review, are generally difficult to obtain. Administrative supplements (generally below a certain dollar limit) can be awarded by the agency.

Most supplements are made in response to individual, ad hoc requests. However, there are also two supplement programs. In order to increase the number of underrepresented minorities in biomedical research and also the number of scientists with disabilities, the National Institutes of Health has for a number of years funded a program of supplements whereby an investigator who has a grant or cooperative agreement with at least two years of funding remaining can apply for an administrative supplement (i.e., rapid review within the funding agency) to obtain additional funds to support a minority or disabled graduate student, postdoctoral fellow, or junior faculty member to work on the study or a closely related investigation. There are also programs for high school and undergraduate students.

Information about grants

One good source for learning about opportunities to apply for grants and contracts is the NIH Guide for Grants and Contracts, published by the NIH. The Guide appears approximately weekly and is available on the World Wide Web at http://www.nih.gov from where you can view institute program descriptions and access funding and other information.

Assurances by the applicant and her/his institution

In the British colonies in what is now the United States, the "power of the purse" was a primary strategy by which the colonial legislatures could influence the conduct of government. In more recent times (I would guess since about the 1960's), for the U.S. federal government has used the awarding of funds as a mechanism to achieve various governmental objectives. In some cases this strategy has enabled the federal government to enforce behaviors in areas of authority outside those assigned to it by the U.S. constitution (and therefore reserved by the Tenth Amendment to the states or the people). But even where the federal authority is clear, tying requirements for desired behaviors to funding gives a more detailed and powerful means of achieving compliance than that available through overburdened law enforcement and judicial systems. Thus, the applicant institution for a grant from the Public Health Service (of which NIH is a part) must provide assurances related to: Age Discrimination, Civil Rights, Debarment and Suspension, Delinquent Federal Debt, Financial Conflict of Interest, Handicapped Individuals, Human Subjects, Lobbying, Research Misconduct, Sex Discrimination, Vertebrate Animals. These assurances primarily relate to the presence of institutional practices and policies (with which individual researchers must, of course, comply). One area, though, risk to human subjects, often demands specific attention from researchers, such as epidemiologists, who study people (scientists who study non-human animals have a corresponding concern regarding vertebrate animals).

Before an investigator can collect data from or on human subjects, s/he must obtain permission from a human subjects protection review committee. In the United States, these are designated by NIH as Institutional Review Boards (IRB), though they may have other names as well (e.g., "Committee for the Protection of Human Subjects in Research"). The IRB mechanism was created in 1973, following a public outcry and Congressional hearings in response to media publicity about the now-infamous Tuskegee Syphilis Study (the history is presented in James Jones, ***Bad Blood: The Tuskegee Syphilis Experiment – a tragedy of race and medicine***, NY: Free Press, 1981; the following synopsis comes from Stephen B. Thomas and Sandra Crouse Quinn, The Tuskegee Syphilis Study, 1932 to 1972: implications for HIV education and AIDS risk education programs in the Black community, ***AJPH***, 1991; 81:1498-1504). In that study, carried out by the U.S. Public Health Service (PHS) in collaboration with the Tuskegee Institute, Alabama State Board of Health, the Macon County Medical Society and Board of Health, and the Milbank Memorial Fund, poor, African American men in rural Alabama who had been found to be infected with syphilis were studied over four decades to resolve questions about the longterm sequelae of syphilis.

The study had its origins in a series of programs to demonstrate that Black persons in the rural South could be tested and treated for syphilis, whose prevalence had been found to be as high as 40%. A combination of financial shortfall (the loss of the treatment funding that was being provided by the Julius Rosenwald Fund until the stock market crash of 1929) and a 1929 Norwegian study whose findings conflicted with prevailing theories concerning racial differences in the natural history of syphilis led the PHS to launch the study as an "unparalleled opportunity for the study of the effect of untreated syphilis" (p94 in Jones J, quoted on p1500 of Thomas and Quinn). The study was by no means a secret enterprise. While it was continuing, study investigators presented the study at medical/scientific conferences and published their findings in major medical/scientific journals. In order not to lose this "never-again-to-be-repeated opportunity" (p179 in Jones, quoted at p1501 in Thomas and Quinn), the PHS arranged with county, state, and other federal agencies to exclude study participants from treatment, even after penicillin became the standard treatment for syphilis in 1951. Indeed, as late as February 1969 a blue-ribbon panel convened by the Centers for

Disease Control (CDC) decided against treating the men in the study in order to harvest all of the scientific information it could provide.

Before collecting any data from human subjects, an investigator must provide a description of the study objectives, potential risks to participants, and procedures for data collection, participant enrollment, and informed consent to an IRB and obtain their approval. Changes to the study procedures must receive IRB approval, and the IRB must be informed of any harm that occurs to participants. Over the years the requirements related to human subjects protection, including training about use of human subjects and other ethical issues in research, have continued to expand. Most recently the Department of Health and Human Services instituted a requirement that all key personnel on DHHS-funded research must describe education they have received concerning human subjects protection (see http://grants.nih.gov/grants/guide/notice-files/NOT-OD-00-039.html). In their mandate to protect human subjects, IRB's have also begun to consider investigators' freedom from financial conflicts of interests in the research (see http://grants.nih.gov/grants/guide/notice-files/NOT-OD-00-040.html), another topic that since 1995 has been linked to funding of health research. The NIH web site contains links to a great deal of information on research ethics and human subjects protection. The information on human subjects research can be found by starting at the page for Bioethics Resources (www.nih.gov/sigs/bioethics/).

In response to Congressional action, NIH has also instituted and recently expanded requirements that minorities, women, and children be included in studies unless that would conflict with the scientific objectives of the research. The expanded guidelines may be found at: grants.nih.gov/grants/guide/notice-files/NOT-OD-00-048.html.

Study management

Definition of study management: Accomplishing the aims of the research.

Elaboration

1. Working out the details of the study design and implementation, often incompletely specified in the proposal - developing the protocol.
2. Garnering and managing the resources to carry out the study:

    Funding

    Space

    Equipment and infrastructure

    Supplies

    Staff

    Subjects

    Time (from one's other responsibilities)

The new element is the grant award. But grant funding isn't everything. The grant may not bring enough money. And money can't buy everything - space, telephones, secretarial assistance, time.

3. Meeting legal, ethical, institutional, and professional requirements.
4. Obtaining quality data (see Chapter 8 in Szklo and Nieto, 2000)
5. Publishing the study results and possibly other forms of dissemination.
6. Solo versus team

Even a study carried out by a single person research team involves study management in terms of the above. Typically, though, epidemiologic studies involve multiple investigators and staff. Larger and more complex (e.g., multidisciplinary) research teams bring in issues of collaboration, communication, personnel management, and greater needs for formal project management.

Study management tends not to be taught

Study management is somewhat neglected in teaching, for several important reasons, including:

1. Faculty teach what they have been taught. Since study management isn't taught, it is harder to teach.
2. Study management may be better learned "on the job", for example, in a research apprenticeship. That is undoubtedly true for various aspects, but it is also likely that much could be taught systematically.
3. Designing the study (more "intellectual") and securing the funding (money talks) have higher status than getting the work done. Doing the work takes longer, entails more tedium, and involves more mundane tasks. In industry, there is more glory for sales than for production. Consider the fascination of sex versus the reality of raising children.
4. Teaching study management is akin to teaching other aspects of methods - most often the teaching does not flow out of nor feed into the faculty member's research. Therefore the time demands of teaching are felt more strongly.

Natural history of study conduct

(on analogy to William Haddon's pre-crash, crash, post-crash for motor vehicle injury prevention)

Preaward

The principal investigator's role in study management begins with the proposal development activity. Production of the proposal is a foretaste of what production of the study will involve, only on a smaller scale. Developing and elaborating a vision, maintaining interest of a team, delegating responsibilities, communicating, supervising staff, negotiating, distributing rewards, organizing work, projecting costs and completion times, etc. Key considerations: are the right people being involved, is the concept sound, is there sufficient time set aside for quality thinking and review, is the budget realistic (in both directions)?

After submission but before the award, the PI may be the only study staff person as well as the PI. That means (if funding appears likely):

Writing job descriptions

Requesting space

Interviewing and hiring

Negotiating subcontracts

Developing instruments

Working out communications infrastructure

Clarifying uncertainties in the proposal

Publicity and media relations (news release)

Planning preaward expenditures? (equipment, personnel)

Problem: How to motivate before the award is official?

Award

Once the award arrives, there is often the painful necessity to reduce the budget: the study may have been underbudgeted (because of faulty projections or in an attempt to fit under a ceiling or to avoid or respond to criticism) or the award may be for less than the amount requested. Should the scope of the project be narrowed?

Startup tasks come next - many of the things that ideally would have been done before the award began likely remain to do - for example, staffing, space, equipment, subcontracts. All of these issues take time, and the PI will want the best staff, the most suitable space, the best type of equipment affordable, and the most desirable subcontract. She/he will probably feel lacking in expertise in some or all of these areas. And each has the potential to take much more time than appears to be available. In the meantime, various people (reporters, university administrators, community people, politicians, students) may want information about the study, whereas all that is available are the grant proposal and the PI!

With core staff and infrastructure in place, the PI is ready to confront the reality of the work that has been proposed and to find out how well have the production aspects of the research been anticipated? At this time, more conventional project management needs come into play:

1. Managing time
2. Managing people
3. Managing money

The basics of these three dimensions can be stated as follows:

A. Create a workplan

1. What needs to be done?
2. How long will it take?
3. When will it happen?
4. Who is responsible?
5. How much does it cost?

B. Revise the plan (so keep it easy to change)

(Adapted from Alan Gump, The basics of project management, *Symantec Newsletter*, Summer 1991, 30-31:)

Of course, managing means more than just making and revising a plan. Staff need to be found, oriented, trained, motivated, guided, supervised, and sometimes disciplined. Involving staff with a democratic management style can help staff to become invested in the study goals and its accomplishments. But there are costs to democracy (time, persuasion energy) and not everyone may want to participate in that way.

Two problems:

1. Since research almost by definition involves unfamiliar terrain, one typically doesn't know all that one needs to know to make the plan and carry it out.
2. Information (e.g., about what is going on in data collection, about spending) is often less available than would be helpful.

Some other issues that arise are:

- Data management - like study management, this is a major but largely unrewarded activity, especially documentation. Generating adequate, timely documentation is a constant challenge, and a balance must also be struck between documenting and doing.
- Monitoring - in principle, all activities should be monitored, so that adjustments can be made in time to meet study goals.
- E.g., a doctoral student found that after 4 months of his study, only 25% of WIC mothers recruited in the first 2 months had filled out their 2nd visit survey (because they missed the scheduled appointment and there was no other mechanism in place). So he changed the protocol to have the resurvey completed at any time the subject came for a WIC visit). He also revised his entry criteria when the number of eligible women was smaller than expected.
- All data should be reviewed as close to the time of its collection as possible, so that corrective action can be taken. For example, are all forms present and completed consistently? However, who will do this? PI? Project coordinator? Research assistant? Graduate assistant? Crucial but unpopular and always a "second" priority.
- Data analyses, especially those that will be published, require extra scrutiny. Ideally, analyses for publication should be replicated independently from the original work. It does not feel good to have to submit a letter like "We regretfully report that we have discovered an error in computer programming and that our previous results are incorrect." (New England

Journal of Medicine, Jan 14, 1999, p148). It is much, much easier to make programming errors than to find them.

- Professional meetings (formal presentations and works in progress) are both opportunities and challenges - opportunities to gain the insight and perspectives of outsiders and a stimulus to thinking through the study data and their interpretation; challenges because they typically take precedence over other important tasks and draw upon data that have not been fully cleaned.
- Getting time to write - the daily needs of the study tend to absorb all the time available for it, and so often it seems that the data are never ready until the study is over.
- How to manage authorship so that opportunities are appropriately shared, enough writing gets done, and the architects of the study get a reasonable share of articles. Increasingly, written policies are being developed to guide investigators and would-be authors.

Some other predictable and unpredictable challenges:

- Staff turnover - possible detours
- Continuation applications and progress reports
- Human subjects permission renewal
- Site visits
- Managing time and money so one comes out ahead but not too far
- Figuring out - beforehand - what data will need to be collected, entered (e.g., dates), analyzed
- Low response rates (NCM SHS & QFL)
- Low exposure rates (Irva)
- Missing data
- Fraudulent data (NCM, Russ Harris)
- Competition from other investigators
- Controversy (animal rights activism, safety hazards (lab, fieldwork), ethics attacks)
- Changing priorities of the funding agency
- Initiating new proposals - new research, funding for self and staff
- Staying in touch with the science

Industrial quality control technique (per King Holmes) - Identify the 5 ways this project can fail. Then try to reduce their likelihood.

Postaward

- Final report
- What to do with all of the data, files, equipment, etc.
- Finishing the papers
- Secondary analyses

Data management

Develop protocols and forms for data collection

- What data to collect
- How will data be transmitted
- Data processing, editing, coding
- Quality control - accuracy, completeness, consistency

Computerize data

- Design screens
- Field separators
- Range checks during data entry
- Double entry with verification
- Quality control - error rate

Compile Data Management Manual

- Study summary
- Brief description of all data streams
- Instrumentation, original sources, and modifications made
- Data collection protocols and (dated) forms
- Account of the data collection process, with dates and numbers
- Correction procedures (audit trail)
- Directory structure / Paths
- File identification - names, labels, computer runs, directory
- Variable identification - names, labels, formats, techniques for long labels, cross-references to forms and instruments

Explore the data

- Look at data (raw)
- Perform "quick and dirty" analyses

Create data files for each data stream, create analysis files

- Check for missing, bad, or duplicate IDs
- Check order of records
- Check for missing forms
- Define special missing values for skip patterns, special situations
- Make sure that all the numbers add up

Prepare an accounting for all data

- Check the N's - tabulations, tracking, records, and resolve
- Eligibility, consents, dispositions for all

Carry out preliminary analyses to inform planning

Clean and summarize data

- Variable distributions (SAS PROC UNIVARIATE FREQ), range checks, outliers, consistency

Create analysis variables and scales

- Derived variables, scales, coefficient alpha, indexes

(For further information, see chapter on Data management and data analysis.)

Typical problems that need to be caught

It is very important to review forms quickly in order to catch irregularites promptly -- when there may be a chance of correcting them. For example, towards the end of a study of partner notification for HIV exposure, it was discovered that partners who were located and notified, but refused to be interviewed, were not being systematically recorded anywhere. The data collection system had been designed to capture partner information on a partner interview questionnaire, but because of the refusal the interview questionnaire was not completed. But the fact of location / notification and who initiated it were important to log.

It is essential to:

1. log all forms in a computer file to be able to account for all data and quickly search for who is in the database.

2. maintain close supervision of the coding operation, to ensure that coding standards are being maintained but at least as important, to apply expert judgment in resolving irregularities detected (the coder's judgment is often not what the investigator would want) and finding out about the reality of the data.
3. pursue systematic data cleaning - subject ID's, linkage of forms, key variables, remaining variables.

Designing a questionnaire

The Inter-university Consortium for Political and Social Research (ICPSR) located within the Institute for Social Research at the University of Michigan provides Internet access to the world's largest archive of computerized social sciences data (http://www.icpsr.umich.edu). In most cases, an abstract and codebook are freely available over the Internet. Persons at institutions that belong to ICPSR can also obtain access to the data themselves.

Writing a paper (courtesy of Barbara Hulka, M.D., M.P.H.)

INTRODUCTION	From literature review
METHODS	Cull from more thorough version written for internal use
RESULTS	Make tables as go along, then cull, highlight and display
INTRODUCTION	Rewrite after doing Results Section

DISCUSSION

1. Repeat - simply - the key results.
2. Review possible biases and other limitations, and indicate how handled or how not. Don't be superficial. Consider counterweighting factors?
3. Refer to recent major studies addressing the same issue and try to identify reasons for differences
4. Depth and breadth:

   implications for theory (from other disciplines)

   implications for practice (for other disciplines)

   the "breadth aspect of EPID" — the ability to get enough grasp of another field's literature to use it.

TITLE - simple, informative, and accurate

ABSTRACT -- that's all many people will read:

1. Why - what's the issue, problem, research hypothesis (1 sentence)
2. What did you do (methods) - study design, N, length of follow-up (2-3 sentences)
3. What did you find (e.g. from first paragraph of discussion)
4. What do you conclude? -- so what?

Recently, standardization has come to abstract writing, in the form of "structured abstracts". Their use is now required by a number of journals.

The text by Szklo and Nieto (2000) includes a helpful chapter on "Communicating the results of epidemiologic studies", which includes a detailed outline of what to report taken from Kahn and Sempos (1989). Szklo and Nieto also include suggestions for writing style (e.g., avoiding scientific arrogance, and jargon), drawing appropriate inferences (including differentiating between association and causality and between statistical significance and strength of association), and preparation of tables and figures.

Bibliography

Funding of epidemiologic research

Baldwin, Wendy. An introduction to extramural NIH. http://grants.nih.gov/grants/intro2oer.htm

Inglehart JK. The NIH appropriation. *N Engl J Med* 1994; 311:1132-1136.

NIH Center for Scientific Review, http://www.csr.nih.gov/refrev.htm

Sultz, Harry. *Grant writing for health professionals*. Little, Brown and Company, Boston, Massachusetts, 1981.

Stallones, Reuel A. Research grants: advice to applicants. *The Yale Journal of Biology and Medicine* 48:451-458, 1975. (humorous)

Study conduct

Cartwright A, Seale C. The natural history of a survey: an account of the methodological issues encountered in a study of life before death. King Edward's Hospital Fund, London, 1990, 135pp ISBN 1-85551-056-1 (reviewed by Mary Monk in *Am J Epidemiol* 1991;134:711).

"Do's and Don'ts" in dealing with the press. *Epidemiology Monitor*. October 1984; 5:, 2-4.

Hulley, Stephen B.; Steven R. Cummings. *Designing clinical research*. Baltimore, Williams & Wilkins, 1988.

Petrie, Hugh G. Do you see what I see? The epistemology of interdisciplinary inquiry. *Educational Researcher* 1976; 5:9-15.

Schoenbach VJ, Arrighi HM. Data management and data analysis. In: Schoenbach VJ. *Fundamentals of epidemiology: an evolving text.* www.sph.unc.edu/courses/epid168/.

Publication

Epidemiology Working Group of the Interagency Regulatory Liaison Group. Guidelines for documentation of epiemiologic studies. *Amer J Epidemiol* 114:609-613, 1981.

Feinleib, Manning. Data bases, data banks and data dredging: the agony and the ecstasy. *J Chron Dis* 1984; 37:783-790.

Findley, Larry J.; Frederick J. Antczak. How to prepare and present a lecture. Commentary. JAMA 1985; 253: 246.

Gopen GD, Swan JA. The science of scientific writing. *American Scientist* 1990; 78:550-558.

"How good is peer review?" Editorial (827-829) and "The Journal's peer-review process" (837-839). *New Engl J Med* 1989; 321(12).

International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *New Engl J Med* 1991 (February 7);324(6):424-428.

International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *JAMA* 1993;269:2282-2286

Instructions for preparing structured abstracts. *JAMA* 1991; 266:42-43.

Kahn HA; CT Sempos, *Statistical methods in epidemiology* (Oxford University Press, 1989)

Lichtenstein, Michael J.; Cynthia D. Mulrow, Peter C. Elwood. Guidelines for reading case-control studies. *J Chron Dis* 1987; 40:893-903.

Northridge ME, Susser M. The paper route for submissions to the Journal. Editorial. *Am J Public Health* (May) 1994;84(5):717-718.

Northridge ME, Susser M. Seven fatal flaws in submitted manuscripts. *Am J Public Health* (May) 1994;84(5):717-718.

Raymond Richard Neutra. Counterpoint from a cluster buster. *Am J Epidemiol* 1990; 132:1-8.

Rennie, Drummond; Richard M. Glass. Structuring abstracts to make them more informative. Editorial. *JAMA* 1991; 266:116-117.

Rennie D, Yank V, Emmanuel L. When authorship fails. A proposal to make contributors accountable. *JAMA* 1997; 278:579-585

Riesenberg, Don; Goerge D. Lundberg. The order of authorship: who's on first? Editorial. *JAMA* 1990; 264(14):1857.

Rothman, Kenneth J. A sobering start for the cluster busters' conference. *Am J Epidemiol* 1990; 132(suppl):S6-S13.

Samet, Jonathan M. Dear Author – advice from a retiring editor. *Am J Epidemiol* 1999;150:433-436.

Szklo, Moyses; F. Javier Nieto. Epidemiology: beyond the basics. Gaithersburg MD, Aspen, 2000. Chapters 8, "Quality assurance and control", and 9, "Communicating the results of epidemiologic studies, 407-430.

Vaisrub, Naomi. Manuscript review from a statistician's perspective. Editorial. *JAMA* 1985; 253:3145-3147.

Yankauer, Alfred. Editor's ReportÄÄOn decisions and authorships. *Am J Public Health* 1987; 77:271-273.

Research ethics

UNC-CH Faculty Committee on Research. *Responsible conduct of research*. Chapel Hill, NC, UNC-CH Office of Research Services, 1994..

Sigma Xi. Honor in Science, 1986 (42pp, $2.50/copy from Sigma Xi).

Shapiro, Samuel. The Decision to publish: ethical dilemmas. J Chron Dis 38; 1985:366?-372.

Schulte, Paul A. The epidemiologic basis for the notification of subjects of cohort studies. Am J Epidemiol 1985; 121:351-361.

Soskolne, Colin L. Epidemiology: questions of science, ethics, morality, and law. Am J Epidemiol 1989; 129:1-18.

Policy

Goodman, Richard A.; James W. Buehler, Jeffrey P. Koplan. The epidemiologic field investigation: science and judgment in public health policy. Am J Epidemiol 1990; 132:9-16.

Morrison AB. Public policy on health and scientific evidence -- Is there a link? J Chron Dis 1984; 37:647-652.

Stallones, Reuel A. Epidemiology and public policy: pro- and anti-biotic. Amer J Epidemiol 115:484-491, 1982.

Weed, Douglas L. Between science and technology: the case of antihistamines and cancer. J Natl Cancer Inst 1994;86:740-741.

16. Data management and data analysis*

Data management: Strategies and issues in collecting, processing, documenting, and summarizing data for an epidemiologic study.

1. Data Management

1.1 Introduction to Data Management

Data management falls under the rubric of project management. Most researchers are unprepared for project management, since it tends to be underemphasized in training programs. An epidemiologic project is not unlike running a business project with one crucial difference, the project has a fixed life span. This difference will affect many aspects of its management. Some areas of management that are affected are hiring, firing, evaluation, organization, productivity, morale, communication, ethics, budget, and project termination. Although the production of a study proposal raises many management challenges, if the proposal is approved and funds allocated, the accomplishments of the project are dependent more upon its management than any other factor.

A particular problem for investigators and staff, if they lack specific training or experience, is to fail to appreciate and prepare for the implications and exigencies of mass production.

1.2 The Data Management System

The data management system is the set of procedures and people through which information is processed. It involves the collection, manipulation, storage, and retrieval of information. Perhaps its most visible tool is the computer; however, this is merely one of many. Other "tools" are the instruments and data collection forms, the data management protocol, quality control mechanisms, documentation, storage facilities for both paper and electronic media, and mechanisms of retrieval. The purpose of the data management system is to ensure: a) high quality data, i.e., to ensure that the variability in the data derives from the phenomena under study and not from the data collection process, and b) accurate, appropriate, and defensible analysis and interpretation of the data.

---

* The original version of this chapter was written by H. Michael Arrighi, Ph.D.

1.3 Specific Objectives of Data Management

The specific objectives of data management are:

1.3.1 Acquire data and prepare them for analysis

The data management system includes the overview of the flow of data from research subjects to data analysts. Before it can be analyzed, data must be collected, reviewed, coded, computerized, verified, checked, and converted to forms suited for the analyses to be conducted. The process must be adequately documented to provide the foundation for analyses and interpretation.

1.3.2 Maintain quality control and data security

Threats to data quality arise at every point where data are obtained and/or modified. The value of the research will be greatly affected by quality control, but achieving and maintaining quality requires activities that are often mundane and difficult to motivate. Quality control includes:

- Preventing and detecting errors in data through written procedures, training, verification procedures, and avoidance of undue complexity
- Avoiding or eliminating inconsistencies, errors, and missing data through review of data collection forms (ideally while access to the data source is still possible to enable uncertainties to be resolved) and datasets
- Assessing the quality of the data through notes kept by interviewers, coders, and data editors, through debriefing of subjects, and through reviews or repetition of data collection for subsamples
- Avoiding major misinterpretations and oversights through "getting a feel" for the data.

Security concerns include: (1) legal, (2) safety of the information, (3) protection from external sources, (4) protection from internal sources. While abuse is more salient, accidental problems are more common. Typical preventive measures are removal or isolation of information that identifies research subjects (to protect confidentiality), redundancy, and backups (to protect against human and machine malfunction). The loss of important data due to failure to have a secure backup copy could be construed as negligence. Unfortunately, there can be an inverse relationship between security and accessibility/usefulness of the data.

1.3.3 Support inquiries, review, reconstruction, and archiving

Inquiries and requests for instruments and/or data may arise at any time during the project and after its completion. The funding agency will require a final report. Other investigators or interested parties (e.g., corporations whose products are implicated as health threats) may request a copy of the data set to pursue their own analyses. Rarely, an investigation may be conducted because of the salience of the findings, the involvement of parties with a large stake in their

implications, or suspicions or charges concerning the study. For example, Herbert Needleman, a pioneering investigator into the effects of childhood lead exposure on cognitive function, had his data and results audited by a scientific committee (which included a UNC faculty member). Proctor and Gamble, Inc., brought suit against the CDC to require the provision of data from their case control studies of toxic shock and tampons.

Concern about scientific misconduct and fraud continues to increase, and investigators have the responsibility to maintain documentation to allay any such charges should they arise. Increasingly, journals require that data (and supporting documentation) be retained for several years following publication. On a more mundane level, innumerable questions will arise during the course of the data analysis, and the project's data management system needs to be able to provide accurate and timely answers.

An important principle in data management, at all levels and stages, is the full accounting for data. Thus when a data collection activity takes place, there should be a detailed record of the number of subjects (if known) in the universe from which subject recruitment takes place and a complete tabulation within a set of mutually exclusive categories (dispositions). Typical dispositions are -- ineligibles according to the reason for their ineligibility (e.g., out of age range, medical conditions), nonparticipants according to the reasons for their nonparticipation (e.g., no telephone number, disconnected telephone, out of town, refused), participants whose data are excluded (e.g., too many missing data items, interviewer skeptical of participant's truthfulness), etc.

An audit trail is an essential mechanism to identify changes to the data at every step. The audit trail should document what changes were made, who made them, and where, when, and how the changes were made. Audit trails are important for responding to or recovering from: (1) legal challenges, (2) procedural issues, (3) minor problems, and (4) disaster.

Note that the above objectives apply to both manual and computerized systems.

1.3.4 Special issues in blinded studies

The HIV epidemic has led to a major activity in conducting blinded serosurveys to determine the prevalence of HIV infection in different settings, subgroups, and geographical areas. In order to avoid bias from nonresponse, a particular concern in HIV studies given the low prevalence of the virus in most populations and the fear and stigma associated with HIV infection and risk factors, methods have been developed for conducting blinded (unlinked) studies. Such studies use leftover blood collected for other purposes (e.g., medical tests) and are analyzed in such a way that identification of the individuals involved in the study is impossible. Under certain circumstances, such studies do not require informed consent, so that they can be free from nonresponse bias.

Special care is needed to design a data management system that can prevent the possibility of linking data to individual subjects. For example, standard data collection procedures such as the use of sequential ID numbers, inclusion of exact dates on all forms, and recording of supplemental information to clarify atypical data items can compromise anonymity. Indeed, unlinked studies

engender a basic conflict between the need to prevent linkage and major data management objectives, such as monitoring and quality control, which require the ability to be able to trace back and verify information.

1.4 The Components of Data Management

1.4.1 Management

The general concepts of management are as applicable to data management as they are to project management. Management issues are critical components of the data management system. The data are merely the objects being manipulated by the data management system. Unless there is adequate attention to the process, the data will not be worthy of much attention.

The investigative team is ultimately responsible for the outcome of the project. Even in those large projects where a project manager and a data manager are present, the investigative team are the project's board of directors. Management skills are required to evaluate the managers and ensure that they are doing a reasonable job, beyond the issue of is the project on schedule? Even for a relatively small project, researchers may need to work diligently to adapt to the managerial role, since many of the qualities that make for a good researcher are quite the opposite of those of a good manager:

Researcher	Manager
Optimal Solutions	Pragmatic Solutions
Accurate Solutions	Workable Solutions
Works with things	Works with people
Process Oriented	Outcome Oriented
Individually Successful	Group Successful

A good researcher requires creativity and may be considered a tinkerer, i.e., a person who is constantly changing things based on new ideas. A good manager is also creative but is less of a tinkerer. Constant change in a management situation results in confusion and a lack of consistency, which ultimately result in data of low quality. A few of the key management issues that directly affect the data management system are:

1.4.1.1 Two-way communication

Any person on the project can make a valuable contribution, comment, or observation. These contributions should be respected. Listening is an important aspect to learn what is truly happening and discover remedies. The person actually performing the task will often know more about the nuances and what is effective than anyone else. People have different degrees of readiness to express their ideas and concerns, and opportunities for informal one-on-one discussions (e.g., over coffee) are important.

1.4.1.2 Consistency

Consistency is essential in the implementation of the protocol, in the data collection process, and with regards to decisions made during the project. Lack of consistency may result from different decisions among the principal investigators, a lack of communication in relaying problems and solutions, and different decisions made on the same issue at different times. Innumerable minor questions (and many major ones) will arise during the course of a project. A useful perspective to use in resolving these questions is how would outside investigators view the decision. Decisions often need to be made quickly and recorded in some form that they can be readily consulted when the same or a similar question arises at a later date.

An inability to implement the study protocol consistently may result in selection bias or information bias. Information bias in confounding variables may compromise the ability to correct fully for confounding in the analysis. Also, when the methods and results are presented to others (a fundamental part of academic research) and the inevitable questions arise, it is very embarassing to have to describe and explain inconsistent methods.

1.4.1.3 Lines of authority and responsibility

Authority and responsibility need to be clearly defined and the designated persons accessible to the staff. Accessibility is often a problem in an academic research project where many of the staff are part-time employees, and the full-time management staff has other commitments resulting in a lack of accessibility to the staff. Among other things it is generally desirable to designate a specific person to be responsible for authorizing all changes to the computerized data sets.

1.4.1.4 Flexibility

The data management system must be flexible to respond to changes in the protocol, survey instruments, and staff changes. The longer a project runs the more susceptible it is to changes. Even a project using secondary data is subject to changes. Every project will undergo some modifications. Thus the data management system must be flexible to allow for easy modification.

1.4.1.5 Simplicity

Keep the data management system as simple as possible and within the talents of the (potential) staff. Simplicity reduces errors by reducing dependency on "key" personnel and by making the system easier to learn and implement.

Computers are wonderful tools in data management, but it is easy to complicate things by their use. The use of non-user friendly software packages or uncommon packages breeds complexity. Computerized systems actually increase the cost and technical support of systems. Their benefits are in the realm of increased efficiency and (hopefully) a reduction of errors. A small project may benefit from a predominantly, manual system when properly designed and implemented.

1.4.2 Integration

Integrate the data management system throughout the entire study process from the idea and proposal stage to the final paper, storage of information, and storage of data until planned destruction. Obviously, some concern to data management is given at the proposal stage during the budget and staffing process. More than this is needed; a general flow of the system should be thought through. This will provide a preliminary assessment of its resource demands and feasibility.

1.4.3 Standardization

Standardization extends not just to the instruments but to the procedures for participant enrollment, procedures for records review, data entry mechanisms, documentation, and other facets. This is essential to obtain quality information.

1.4.4 Pilot testing

Pilot testing is routinely done with the survey instruments. Rarely, does one pilot test crucial parts of the data management system. Use the pilot test of the survey instruments as an opportunity to pilot test aspects of the data management system, e.g. coordination of interviewers, call backs, coordination with other sources (participant identification) etc. A key aspect is to make this as real as possible to avoid the "test run" syndrome and a lack of seriousness among the staff.

The data management system may be pilot tested when the preliminary versions of the survey instruments are under evaluation and during the evaluation of the laboratory methods. If the project is sufficiently large, then a pilot test of the entire system may be done on the first few participants (5, 10, or 20). The project is then briefly halted for review and modification prior to complete implementation. Large projects make use of a "vanguard" cohort that goes through all aspects of the study sufficiently in advance of the actual study population to enable adjustment of instruments and procedures.

1.4.5 Quality Control and Quality Assurance

1.4.5.1 Redundancy

Duplication of data collection items is a well-established quality control procedure. This applies equally to survey instruments, laboratory procedures, and the data flow system. Duplication may occur in series or in parallel.

1.4.5.1.1 Parallel

Runs in parallel are the simultaneously evaluation of two data collection items. With a laboratory, this is the blind submission of two identical items for evaluation. With a survey instrument, this is the repetition of a question, perhaps in slightly altered format.

*1.4.5.1.2 Serie***s**

Runs in series are the repetition of items at two different time points. With a laboratory, this is the blind submission of two identical items at different times. With a survey instrument, this is the repetition of all or part of the survey instrument at a different time. This may involve a call back of a selected subsample with a brief verification questionnaire asking some similar items. With data entry procedures, the verification of data entry is accomplished by the duplicate entry of the entire batch of items. These two entries are compared and non-matches are identified and re-entered. Double keying (also called key verification) though standard is not automatic, so must generally be specifically requested and budgeted.

1.4.5.2 Error introduction

A useful technique is to introduce errors into the data management system in order to evaluate the error trapping mechanism and consistency of error handling. This may be done by introducing erroneous data or identifying a particular problem and following it through the data management system.

1.4.6 Reduce the number of entry points

Each participant should enter the study the same way or each subject should have the same opportunity for entry. Different protocols should not be used to enroll different subjects of the same category (e.g. cases or controls, or the exposed or unexposed). This can be challenging when there are multiple sites.

Take a planned approach to every source of data, including logs, tracking forms, appointment systems. Hindsight can reveal usefulness of data sources not originally intended for analysis. Considerations in planning include design of the procedure for collecting, recording, coding, computerizing, verifying, ensuring security, and documenting. Try to limit situations where changes are made directly into a data base, with no audit trail. Without an audit trail it may be impossible to reconstruct data if an erroneous change is made or even to verify if a change has been made.

1.4.7 Monitor

Monitor the data management to ensure its proper implementation. For example, when a data collection activity is underway there should be frequent and regular review of the number of subjects participating, reasons for nonparticipation, problems encountered, etc. Data collection forms (or a sample, if the volume is large) should be scrutized promptly to identify problems (for example, an excess of missing items, misunderstood items), for which corrective action may be possible. A manual or computerized system is important to keep track of data forms.

Crucial elements to identify are:

1. Adherence to the study protocol (to ensure the objectives of the protocol are maintained);

2. Consistency of the protocol's implementation;

3. Deficiencies (and strengths) of the data management system;

4. Response to changes, problems, crises - how well is the data management system detecting and responding to changes, problems, and crises? This monitoring may be accomplished through the use of erroneous data or by tracking the dates and items that were identified as problems, the date of their identification and the date of the correction.

1.4.8 Document

Documentation is a special challenge because of its lack of glamour and the fact that at any particular point in the study (before the end) competing urgent priorities make documentation a very hard activity to keep abreast of. Yet it is absolutely essential and cannot always be satisfactorily reconstructed after the fact. Budget time and personnel for documenting events, decisions, changes, problems, solutions. Review documentation as it is being developed to show the importance you assign to it and to make sure it is in the form you need it.

Document investigator meetings and committee meetings with each item on the agenda (1 or 2 sentences should suffice), followed by the decision or action (open, closed, resolved and summary thereof). Recounting all the discussion is nice but tends to be too lengthy - strive for succinctness and key points. Having minutes published within a day or two of the meeting forces them to be succinct and gets them done before memory fades. Keeping a contemporaneous journal or word processing document with (dated) notes about things that should be included in progress reports is another technique.

Project documentation should include:

- A succinct narrative of the objectives and methods of the study.
- A detailed chronology of events and activities through the life of the project, showing starting and ending dates and numbers of subjects for each data collection activity.
- For each data collection activity, a record of the detailed procedures used and an accounting of the number of subjects in scope, the number selected for data collection, and the disposition for all subjects (by category, e.g., could not contact, refusal). This material should have cross references to original sources (e.g., computer runs) to enable verification when necessary.
- Compendium of all data collection instruments, including documentation of the sources for questionnaire items obtained from pre-existing instruments. Results of pretesting and validation analyses or cross-references to them should be included.
- Lists and descriptions of all interventions applied and materials used (e.g., intervention materials, training materials)
- Documentation on all computerized data and final analyses (information on datasets, variables, computer runs - see below).

Obviously it will be easier to assemble these materials if careful documentation is prepared along the way. As a minimum every document should bear a date and a notation of authorship. Documents retained in a word processing file should if possible contain a notation of where the document file resides, so that it can be located for later revision or adaptation in creating related documents.

1.4.9 Poke around

The amount of activity and detail in a large project can easily exceed what the (usually limited) staff (and time-urgent investigators) are able to handle comfortably. Despite the highest motivation and experience, communication will be incomplete and important items will be overlooked. Investigators should review data forms regularly to familiarize themselves with the data in its raw form and verify that data collection and coding are being carried out as stipulated. It may also be worthwhile even to poke around occasionally in file drawers, stacks of forms, and among computer files.

1.4.10 Repeat data analyses

The fact that debugging is a major component of commercial software development suggests that investigators need to make provisions to detect and correct programming errors or at least to minimize their impact. One strategy is to have a different programmer replicate analyses prior to publication. Typically only a small portion of analyses that have been carried out end up in a publication, so this strategy is more economical than many others, though of course if serious errors misled the direction of the analysis much work will have been lost. Replication begins as close as possible to raw data offers the greatest protection, but more often other methods are used to ensure the correctness of the creation of the first analysis dataset. An object lesson about the importance of verifying the accuracy of data analyses is offered by the following excerpt from a letter to the New England Journal of Medicine (Jan 14, 1999, p148):

> "In the February 5 issue, we reported the results of a study … We regretfully report that we have discovered an error in computer programming and that our previous results are incorrect. … After the error was corrected, a new analysis showed no significant increase …"

It is a good bet that error reports such as these are the tip of the iceberg.

2. Data conversion

Picture stacks of questionnaires, stacks of medical record abstract forms, lists of laboratory results, photocopies of examination results, and the like. Before they can be analyzed, these original data need to be coded for computerization (even if the volume is small enough and the intended analyses are simple enough for manual tabulation, coding is still required). This process can be a major and arduous undertaking, and may involve the following steps for each data "stream":

1. Preparation of a coding manual stating the codes to be used for each data item and the decisions to be made in every possible situation that occurs (see sample coding manual);

2. Coding of a sample of data forms to pilot test the coding manual (see sample coded questionnaire);
3. Revision of the coding manual, re-coding of the sample, and coding of the remainder of the forms (see sample guidelines);
4. Maintenance of a coding log (see sample) recounting the identification number and circumstances for any data item about which a question arose or a nonroutine decision was made, to enable backtracking and recoding if subsequently indicated;
5. Recoding by supervisory personnel of a percentage (e.g., 10%) of data forms as a quality check.

Depending on the source of the data, coding can be largely routine (e.g., verifying that a response category has been circled and perhaps writing the appropriate code to be keyed) or highly demanding (e.g., evaluating a transcript of a medical record to determine the diagnosis and presenting complaint).

Coding is a good opportunity to review each data form for any irregularity that can be easily detected, including verbal information written on the questionnaire. Correction of irregularities (multiple responses to an item, inconsistent values, missing response, etc.) generally requires access to the data form, so it is easier to take care of the problem at the coding stage rather than when the forms have been filed and the computerized dataset is in use.

2.1 Data cleaning / editing – objectives

After coding, forms are data entered with some type of verification to detect and correct keying errors (double keying, manual proofing, etc.). The computerized data are then "cleaned" and "edited". Data cleaning and editing may also be referred to as "damage control." It is at this stage where the initial screening of the collected information is made to assess its validity and useful. Ideally, data cleaning is an ongoing process; initiated when the first results arrive. Detecting errors early may assist in minimizing them in the future and assist in their correction.

2.1.1 Incomplete data

Incomplete data are missing values for a single data item, incomplete or incorrectly completed instruments. An incorrectly completed survey instrument may be one that had the "skip" patterns improperly followed. The identification of these issues and their correction, when possible, are both of importance.

2.1.2 Extreme values

Extreme values for a variable are generally referred to as "outliers". Outliers may also occur for a site (in a multi-site study) or an interviewer who is more "extreme", with regards to timeliness, interview time, or responses. Outliers may meet one or both of two possible criteria.

2.1.2.1 Statistical

There are formal statistical tests for outliers. This process is designed to identify those values that may unduly influence a statistical analysis. A visual inspection of the data is actually quite informative in gaining impressions about potential outliers.

2.1.2.2 Substantive

For an outlier to be truly an outlier it must not make substantive sense, for example a hemoglobin of 0.5 (though a hemoglobin of 0.5 may not be a statistical outlier in a small group of patients with severe anemia, whose expected range might be between 3.5 and 8) or a height of 9 feet 10 inches with a given weight of 95 pounds in an apparently healthy adult.

2.1.3 Expected Results

Examining the data with an idea of what is expected is helpful in determining how "good" these data are.

2.2 Placement of the data editing process

Some reduction of the time and distance between data collection and entry into the analysis system is helpful in error correction. The editing of data should occur during all aspects of the data collection and analysis. Some editing may occur during or shortly after data collection; this often involves manual means. Additional editing procedures will occur later during the formal coding and entry. Post-entry editing procedures will encompass the final aspect of the editing process.

2.2.1 Time of data collection

Examples of this are verification of respondents' identity, use of subject identification numbers with a check digit, clear marking of specimens with duplicate labels (including the caps), prompt reviewing of completed instruments, and provision of pre-labeled instruments.

2.2.2 Time of data entry (keying)

Many data entry programs enable checks for valid ranges and/or values as data are being keyed and can even include "hard" logic (consistency) checks. Modern large scale telephone surveys use computers to track and enter the data during the survey process. This ensures that the survey instrument is properly followed, responses are within tolerance range, and may even provide an opportunity for checks in consistency of response.

2.2.3 Post-entry

Most of the formal steps associated with data editing occur after the data have been keyed and verified. These involve the examination of individual records and their aggregates.

2.3 The steps in the Editing process

2.3.1 Manual Editing

As discussed above, manual checks are performed during coding of data forms. This stage checks for proper completion (skip patterns, etc.) of the questionnaire. Error correction may entail a return to the source of the original information. Or with an abstraction of medical (or other) records a photocopy of the source record may be obtained for comparison purposes.

2.3.2 Frequency distributions

SAS PROC FREQ (typically with the MISSPRINT or MISSING option selected) and PROC UNIVARIATE with the FREQ and PLOT options are useful in examining frequency distributions. Frequency distributions are helpful in identifying the extent and types of missing data, unusual patterns, and potential outliers. For example, birth weight in grams is generally recorded to the nearest ten grams, so the final digit should be zero. Blood pressure is generally recorded in mmHg, so the final digit should be uniformly distributed between 0 and 9. "The case of the missing eights" (Stellman SD, *Am J Epidemiol* 129:857-860, 1989) presents a case study of how an alert analyst noticed that a distribution of counts contained no counts with 8 as the least significant digit. Only after a great deal of checking and investigation was the problem traced to a programming error (a mixup of zero and the letter "oh").

2.3.3 Logic checks

These checks are evaluations of internal comparisons within a single observation.

"Hard" comparisons are made when there are obvious inconsistencies, for example pregnancy in males, or following the proper skip pattern of the survey instrument.

An example of a "hard" comparison is where sex is abstracted from two different sources. Those records with disagreement may be identified and handled according to the study protocol. This protocol may set those in disagreement to missing, or there may be preferred data source (e.g. respondent questionnaire versus medical record), or the original respondent may be contacted for verification. Disagreement across forms may reveal that the forms belong to different respondents.

"Soft" comparisons are possible but the exact values (cutpoints) will be study dependent. Marital status may be questioned if age at marriage is below a specified value, which may be gender-specific. The exact age is chosen dependent upon the population under investigation. Another check might include birth weight by gestational age.

2.3.4 Univariate displays

These statistics are useful in determining measures of central tendency (the familiar mean, median, and mode), measures of dispersion (standard deviation, variance, range, percentiles, skewness, and kurtosis). In addition, graphical representations (e.g. histograms, box plots, and normal probability plots) are helpful for describing data.

2.3.5 Bivariate displays

If the same data have been collected at multiple points, then agreement across the measurement points should be assessed. In addition, expected relationships can be examined: birth weight and gestational age, systolic blood pressure and diastolic blood pressure, height and weight. Unusual combinations should prompt examination for coding or entry errors.

Differences may be examined in the case of continuous variables and an acceptance protocol developed.

2.4 Treatment of missing values

Coding of missing and/or inconsistent responses merits careful thought. An item may have no response for a variety of reasons, and it is often useful to distinguish among those reasons. For example, an item may not be relevant for some types of respondents (e.g., a question about a prostate exam asked of a female participant, a question about the year in which injection drugs were last used asked of a participant who has not injected drugs), a "screening" question may have skipped the respondent around the item, the respondent may not recall the answer, the respondent may decline to answer, or the respondent may simply omit an answer (in a self-administered questionnaire) without giving a reason. An item such as "Have you had a Pap test in the past year?" included in a self-administered questionnaire may not have been answered for any one of these reasons.

If there are many missing responses and only a single missing value code is used, then a frequency distribution will often leave the analyst wondering about the usefulness of the item, since if there are too few analyzable responses the item conveys limited information. It is preferable to use a different missing value code for each situation. Then a frequency distribution can show the number of responses of each type.

2.4.1 Techniques for coding missing values

A widely-used convention for coding missing values uses out-of-range numeric codes, such as "999", "998", or "888", to represent missing values, with the number of digits equal to the field length. For example, in the public use datafile for the National Survey of Family Growth, responses of "don't know" are coded 9, 99, 999, or 9999; refusals are coded 8, 98, 998, or 9998; and "not ascertained" values are 7, 97, 997, or 9997, depending on the column length of the original data items. There are several limitations to this coding procedure. First, statistical packages may not recognize these as missing values, so that it is easy (and embarrassing) to have the actual values actually included in analyses. Second, it may be necessary to use different codes for the same missing reason due to different field lengths. Third, the numbers provide no useful mnemonic device.

One very useful facility in the Statistical Analysis System (SAS) is the provision for special missing value codes. A missing value for a numeric variable can be coded with one of 27 different missing value codes, consisting of a dot or a dot followed by a single letter.

Although it is rare to use more than two or three for a given variable, an analysis could differentiate among "not applicable", "does not know", and "refused" by coding these as .n, .k, and .r, respectively. A more elaborate coding for a variable such as menstrual discomfort might differentiate among not applicable due to gender (.m), not applicable due to having had a hysterectomy (".h"), not applicable due to natural cessation of menses (.c).

These values can be tabulated separately in frequency distributions, so that the extent and nature of "missingness" for a variable can be quickly assessed, and the analyst can keep track of denominators more easily. SAS generally does not include missing values coded in this way in calculations, which saves some programming effort and protects against programming lapses. The TABLES statement in PROC FREQ provides an option (MISSING) to treat missing values the same as all other values (useful to examine percentages of missing values) and an option (MISSPRINT) to display missing values in tables but not include them in the denominator for percentages (permitting the correct computation of percentage distributions for analysis while permitting easy verification of the number of data points and reasons for their absence.

2.5 Outliers

Examination for extreme values (range checks) is also a crucial preliminary step in the screening of data. First, outliers should be checked to the original data forms to verify accuracy of transcription. If the outlier cannot be dismissed as an error, then care must be taken to avoid distorting the analysis.

2.5.1 What to do with them?

Outliers may be replaced with a missing value, but then the observation is lost with regards to the analysis (and in a mathematical modeling procedure, the entire observation is unused). Moreover, if the outlier is a legitimate value, then simply deleting it is a questionnable procedure.

The analysis can be repeated with and without the outlier data to assess the impact of the outlier on the analysis. Or, the analysis can be repeated using (nonparametric) statistical procedures that are not affected by outliers and the results compared to parametric procedures - or nonparametric procedures can be used completely. Such procedures typically involve medians, rather than means, or focus on the ranks of the values of a variable rather than on the values themselves. Categorical procedures in which a variable is first categorized into groups will often be unaffected by extreme values.

2.6 Concern with Data Cleaning / Editing

There is a problem with the handling of these missing values, outliers, and other edit checks: more attention is given to the extreme problems and less attention to those errors that are not as visible. For example, a transcription error resulting in a birthweight of 2,000 grams being recorded as 3,000 grams may go completely undetected once data entry is completed. But this misclassification bias may have a substantial effect if it occurs in a large enough proportion of observations. The entire data management system should be designed to minimize and reduce these

errors. Related to this concern is the comparison with "expected" values. While this is a useful tool in inspecting and understanding the data, there is a concern in trying to force the data into an expected distribution. Thus the focus is on those errors of the extreme. An error in the opposite direction, from more extreme to less, is missed by this definition of data examination. This latter concern applies equally through the remainder of the data examination and analysis.

2.7 Documentation

Documentation covers all aspects of the data as well as all problems identified, their solutions, and all changes made to the data. Some techniques are:

- Keep a master copy of the questionnaire and record changes and decisions for each item. Cross index this questionnaire with the variable names in the computer files.
- Keep at least the original data and all programs leading to creation of the final dataset, so that any intermediate dataset may be recreated if need be. (This is the rationale for not making direct changes to the data base.)
- Document computer programs with a unique identifier (i.e., program name), title of project, brief description of the program, input and output, any dependencies (programs that MUST be run prior to this or essential data bases), date of request and person, date of development and analyst, including modifications).
- Document computer programs within the program (in comment statements and title statements), files and programs, and externally (i.e. notebooks).
- Maintain a notebook of program runs in chronological order, showing the (unique) program name, date run, programmer, history (e.g., rerun of an earlier version), data set used, and one-line description. Sometimes programs that create datasets are listed in a separate section from programs that analyze datasets.
- Try to use self-documenting methods. Adopt a system of naming conventions for datasets, computer runs, and variable names. Choose meaningful variable names if possible, and allow for suffixes (e.g., using only 7 characters in a SAS variable name leaves the 8th character for designating recodes of the original variable). Assign internally-stored labels to data sets and variables (SAS LABEL or ATTRIB statement). If more than 40 characters are needed, add a comment in the program that creates the variable or dataset. Consider using value labels (formats) to document the values for each variable.

3. Data analysis

With the availability of microcomputer statistical packages, it is easy to compute many statistics that previously required the assistance of someone with biostatistical training (and with fewer distractions from the task of data analysis), with an increase in the danger of uniformed, inappropriate, or incorrect use of statistical tests (W. Paul McKinney, Mark J. Young, Arthur Hartz, Martha Bi-Fong Lee, "The inexact use of Fisher's Exact Test in six major medical journals" *JAMA* 1989; 261:3430-3433).

The first stages of data analysis should emphasize obtaining a "feel" for the data, i.e., some familiarity with their essential features. The process of examining the data to understand them is integrated throughout the cleaning and analysis. Always question data and examine them with a critical view. The same concepts used in data cleaning and editing are applicable in trying to understand the data. Specifically, these are the expected values, missing values, and outliers. Now they are applied in a "multivariate sense."

Many of the methods of approaching a dataset are similar to those described above under data cleaning, such as examination of:

1. Univariate distributions (frequency distributions [PROC FREQ], summary statistics [PROC UNIVARIATE], graphs [PROC UNIVARIATE, PROC PLOT or other].
2. Crosstabulations (frequency distributions across important groupings, such as sex, race, exposure, disease, using PROC FREQ)
3. Scatterplots showing pairs of continuous variables
4. Correlation matrices

These analyses should include the assessment of agreement where it is expected to occur. It is often helpful to prepare summary tables of basic information from the above examination, that can be used for reference purposes during later stages of analysis and writing.

3.1 Data reduction:

Data reduction is an essential activity that, like data management, takes place at virtually every place where data are involved. In the data analysis phase, data reduction involves deciding whether and how continuous variables can be grouped into a limited number of categories and whether and how to combine individual variables into scales and indexes. There is also the need to derive conceptually more meaningful variables from individual data items.

3.2 Graphical representation

There are many graphical packages available that provide the ability to plot, view, and to an extent analyze data. Graphical representations of data are extremely useful throughout the examination of the data. Statisticians are often familiar with these techniques for examining the data, describing data, and evaluating statistical tests (e.g. plots of residuals). The visual impact of a graph is informative and will increase the understanding of the data and limit the surprises that may occur. There are few general principles, as each data set is different and will have an individual approach. Many of the modern statistical graphics packages available on personal computers have a variety of functions such as fitting curves, for example, linear, quadratic, other polynomial curves, and spline curves.

3.3 Expected values

Perhaps, the single most important concept to remember is to have an idea of what is expected. This concept has been applied during the editing and cleaning process. Understanding what is expected is a function of both the study design and the values of the parameters in the target population. For example, if randomized allocation has been used, then the randomized groups should be similar. If controls are selected from the general population via random digit dialing methods, then their demographics should reflect the population as a whole. When examining a table, first check the variables, labels, and N's for the total table and the subcategories that are not included to make sure that you understand the subset of observations represented. Second examine the marginal distributions to make sure they conform to what you expect. Then examine the internal distribution, particularly, with regards to the referent group. Finally proceed to assess the association or other information in the table.

3.4 Missing values

The impact of missing data is magnified for analyses involving large number of variables, since many analytic procedures require omitting any observation that lacks a value for even one of the variables in the analysis. Thus, if there are four variables, each with missing data for 10% of the observations, in a worst-case situation 40% of the observations could be omitted from the analysis. To assess the extent and nature of missing data for a variable, a complete "missing value" analysis should ideally be done. That means comparing the presence/absence of information for a variable with other key factors, e.g. age, race, gender, exposure status, and/or disease status. The goal is to identify correlates of missing information. Relationships are indicative, though not conclusive, of selection bias. This analysis may give insights into how to impute values for those missing (e.g., missing cholesterol could be estimated as a function of sex, age, race, and body mass). Strong relationships between one covariate and missing values for another indicate that imputed values should be stratified by levels of the first covariate.

Although they receive relatively little attention in introductory treatments of data analysis, missing values are the bane of the analyst. Examination of the data for missing values (e.g., via SAS PROC FREQ or PROC UNIVARIATE) is an essential first step prior to any formal analyses. Special missing value codes (see above) facilitate this examination. Missing values are a serious nuisance or impediment in data analysis and interpretation. One of the best motivations to designing data collection systems that minimize missing values is experience in trying to deal with them during analysis!

3.4.1 Effects of missing data

Two kinds of missing data can be distinguished: data-missing and case-missing. In the former case, information is available from a study participant, but some responses are missing. In case-missing, the prospective participant has declined to enroll or has dropped out. This discussion will address the situation of data-missing.

Missing data have a variety of effects. As a minimum, missing data decrease the effective sample size, so that estimates are less precise (have wider confidence intervals) and statistical tests

have less power to exclude the statistical null hypothesis for observed associations. This problem is compounded in multivariable analyses (e.g., stratified analysis or logistic regression), since most such procedures drop every observation which has a missing value for any of the variables in the analysis. Thus, a logistic model with eight variables can easily lose 30% of the observations even if none of the individual variables has more than 10% missing values.

In both univariate and multivariable analyses, missing data leads to what might be referred to as the problem of the "changing denominator". Each one-way or two-way table may have different numbers of participants, which is both disconcerting to readers and tedious to keep explaining. One workaround is to analyze only complete-data cases (i.e., observations with no missing values), but the price in number of observations lost may be unacceptable.

Missing data situations are characterized in terms of the degree and patterns of "missingness". If there is no systematic pattern to missing data for a particular item, i.e., all participants are equally likely to omit a response, then the missing values are missing completely at random (MCAR). When data are MCAR, then estimates from the nonmissing data will not be biased by the missing data, since the nonmissing data is essentially a simple random sample of the total (potential) data.

It is probably more often the case that different groups of participants have different rates of missing data. In this case, the data are missing at random (MAR) (assuming that missing data occur randomly within each group). If groups who differ in their rates of missing data also differ in their distributions of the characteristic being measured, then overall estimates of that characteristic will be biased.

For example, if persons with multiple sexual partners are more likely to decline to answer a question on that topic, then the estimate of the mean number of partners or the proportion of respondents with more than X partners will be biased downwards. Estimates of associations with other variables may also be distorted. Furthermore, attempts to control for the variable as a potential confounder may introduce bias (from selectively removing observations from the analysis) or due to incomplete control for confounding.

3.4.2 What to do about missing data?

As in so many other areas of public health, prevention is best. First, data collection forms and procedures should be designed and pretested to minimize missing data. Second, it may be possible to elicit a response from a hestitant or unsure respondent (but such elicitation must avoid the hazards of eliciting an inaccurate response or contravening in any way the participant's right to decline to answer), to recontact participants if questionnaire review turns up missing responses, or to obtain the data from some other source (e.g., missing information in a hospital medical record may be available from the patient's physician). Third, it may be possible to combine data from different sources to create a combined variable with fewer missing values (e.g., sex from a questionnaire and sex from an administrative record, though the issue of differential accuracy of the sources may be an issue).

Despite the best efforts, however, missing data are a fact of life, and it is the rare observational study that avoids them completely. Nevertheless, the smaller the percentage of missing data, the smaller a problem they will create and the less it will matter how they are dealt with during analysis.

3.4.3 Do not try to control for missing values of a confounder

The suggestion arose some years ago to treat missing values as a valid level of a variable being controlled as a potential confounder. For example, if an association was being stratified by smoking, there might be three strata: smoker, nonsmoker, smoking status not known. Recent work suggests that this practice may actually increase confounding and is not recommended.

3.4.4 Imputation for missing data

In recent years a great deal of work has gone into developing analytic methods for handling missing data to minimize their detrimental effects. These methods seek to impute values for the missing item responses in ways that attempt to increase statistical efficiency (by avoiding the loss of observations which have one or a few missing values) and to reduce bias that results when missing data are MAR, rather than MCAR (i.e., missing data rates vary by subgroup).

One simple method of imputation, now out of favor, is simply to replace missing values with the mean or median of the available responses. This practice enables observations with missing values to be used in multivariable analyses, while preserving the overall mean or median of the variable (as computed from the nonmissing responses). For categorical variables, however, the mean may fall between categories (e.g., the mean for a 0-1 variable may be .3), and for all variables substituting a single value for a large number of missing responses will change the shape of the distribution of responses (increasing its height at that value and reducing its variance), with effects on statistical tests. Moreover, if missing values are not MCAR, then the mean of the observed values may be biased and therefore so will the mean of the variable after imputation.

3.4.5 Randomized assignment of missing cases

A more sophisticated approach is to draw imputed values from a distribution, rather than to use a single value. Thus, observations without missing values (complete data cases) can be used to generate a frequency distribution for the variable. This frequency distribution can then be used as the basis for randomly generating a value for each observation lacking a response. For example, if education was measured in three categories -- "less than high school" (25% of complete data cases), "completed high school" (40%), or "more than high school" (35%) -- then for each observation with education missing, a random number between 0 and 1 could be drawn from a uniform distribution and the missing value replaced with "less than high school" if the random number was less than or equal to 0.25, "completed high school" if the number was greater than 0.25 but less than or equal to 0.65, or "more than high school" if greater than 0.65.

This method avoids introducing an additional response category and preserves the shape of the distribution. But if the missing data are not MCAR, the distribution will still be biased (e.g.,

greater nonresponse by heavy drinkers will still lower the estimate of alcohol consumption; greater nonresponse by men may also lower the estimate of alcohol consumption).

3.4.6 Conditional imputation

Modern imputation methods achieve more accurate imputations by taking advantage of relationships among variables. If, for example, female respondents are more likely to have a confidant than are male respondents, then imputing a value for "presence of a confidant" can be based on the respondent's sex. WIth this approach, confidant status among men will be imputed based on the proportion of men with a confidant; confidant status among women will be imputed based on the proportion of women with a confidant. In this way, the dataset that includes the imputed values will give a less biased estimate of the population values than will the complete-data cases alone.

A simple extension from imputation conditional on a single variable is imputation conditional on a set of strata formed from a number of variables simultaneously. If the number of strata is too large, a regression procedure can be used to "predict" the value of the variable to be imputed as a function of variables for which data are available. The coefficients in the regression model are estimated from complete-data cases.

Imputed values are then randomly assigned (using a procedure such as that outlined above) using the stratum-specific distributions or predicted values from the regression model. This strategy provides superior imputations for missing values and preserves associations between the variable being imputed and the other variables in the model or stratification. The stronger the associations among the variables, the more nearly accurate the imputation. There does remain, though, the problem of what to do when the value of more than one variable is missing. If in actuality two variables are associated with each other, then imputing values to one independently of the value of the other will weaken the observed association.

3.4.7 Joint imputation

Yet another step forward is joint imputation for all of the missing values in each observation. Picture an array which categorizes all complete-data observations according to their values of the variables being considered together and a second array categorizing all remaining observations according to their configuration of missing values. Suppose there are three dichotomous (0-1) variables, A, B, C and that A is known for all respondents but B and/or C can be missing. The arrays might look like this:

Complete data cases

Stratum #	A	B	C	Count	Percent of total	% distribution conditioning on A	% distribution conditioning on A & B	% distribution conditioning on A, C=0	% distribution conditioning on A, C=1
1	0	0	0	400	33	53	67	80	
2	0	0	1	200	17	27	33		75
							100		
3	0	1	0	100	8	13	67	20	
4	0	1	1	50	4	7	33		25
						100	100	100	100
5	1	0	0	240	20	53	62	83	
6	1	0	1	150	13	33	38		88
							100		
7	1	1	0	40	3	9	67	17	
8	1	1	1	20	2	4	33		12
						100	100	100	100
Total				1,200	100				

Missing value configurations

Configuration	A	B	C	Count
a.	0	0	.	12
b.	0	1	.	18
c.	1	0	.	10
d.	1	1	.	30
e.	0	.	0	40
f.	0	.	1	10
g.	1	.	0	15
h.	1	.	1	25
i.	0	.	.	20
j.	1	.	.	10

In this example, the eight strata in the cross-classification of the complete data cases are numbered 1 through 8, and the percentages for each stratum are computed in four different ways: unconditionally (i.e., the count as a percentage of all of the complete-data cases), conditionally based on the value of A only, conditionally based on the value of A and B, and conditionally on the value of A and C [the latter requires two columns for clarity]. Meanwhile, the 10 possible missing data configurations are arrayed in the second table and labeled a. through j.

Imputation is then carried out as follows. Missing value configuration a. has A=0 and B=0, so the 12 cases in this configuration belong in stratum 1 or stratum 2. To preserve the distribution of the complete data cases in those two strata (67% in stratum 1, 33% in stratum 2 – see column headed "A & B"), the 12 cases are randomly assigned to stratum 1 and stratum 2 with assignment probabilities in that proportion, so that stratum 1 is expected to receive 8 and stratum 2 to receive 4. The 18 cases in configuration b. have A=0 and B=1, so they belong in either stratum 3 or stratum 4. These 18 cases will be randomly allocated between these two strata with probabilities proportional to the distribution of the complete data cases across those two strata (which happens to be the same as the strata with A=0 and B=0). Configurations c. and d. will be handled in the same manner. Configuration e. has A=0 and C=0, so the 40 cases in this configuration belong in either stratum 1 or 3. These 40 cases will be randomly assigned to strata 1 or 3 in proportion to the distribution in the column headed "A, C=0". Thus the random assignment procedure will on average assign 32 cases (80%) to stratum 1, and 8 cases (20%) to stratum 3. The remaining configurations will be handled in the same manner. Configurations i. has A=0 but no restriction on B or C, so the 20 cases in this configuration will be randomly allocated across strata 1, 2, 3, or 4 according to the distribution in the column headed "A" conditional on A=0.

Joint, conditional, imputation makes maximum use of the available data on the three variables, adjusts the distribution of each variable to give a better estimate of that expected for the population as a whole and preserves many of the two-way associations involving variables being imputed. The procedure can be carried out using a modeling procedure instead of a cross-classification, which enables the inclusion of more variables.

Model-fitting using the EM ("Expectation Maximization") algorithm is the current state of the art. The BMDP AM procedure uses this algorithm, but it is designed for continuous variables with a multivariate normal distribution and imputes each variable independently, so that two-way associations are weakened. A new program by Joe Shafer at Pennsylvania State University uses the EM algorithm with categorical variables and jointly imputes data; however, it requires very powerful computer resources.

3.4.8 Multiple imputation

All of the above procedures result in a single dataset with imputed values in place of missing values. However, since the imputed values are derived from the rest of the dataset, analyses based on them will understate the variability in the data. As a corrective, the imputation process can be carried out repeatedly, yielding multiple datasets each with a (randomly) different set of imputed values. The availability of multiple imputations enables estimation of the additional variance introduced by the imputation procedure, which can then be used to correct variance estimates for the dataset as a whole.

[With thanks to Drs. Michael Berbaum, University of Alabama at Tuscaloosa and Ralph Foster, Research Triangle Institute (NC USA) for educating me on this topic and reviewing this section.]

3.5 Outliers

Outliers are now examined with respect to a multivariate approach, i.e. are there any extreme values. For example, you stratify the exposure - disease relationship by a factor with 4 levels. The observation is made of the 4 stratum specific odds ratios of 2.3, 3.2, 2.7, and 0.98. The fourth stratum indicates a potentially strong interaction. What if this stratum contains only 6 observations? Even though the association may be statistically significant, collapsing the strata is reasonable as the most extreme table may be a result of imprecision. Alternatively, the values of the most extreme table may be recategorized.

3.6 Creation of analysis variables

The variables defined to contain the data in the form it was collected (as responses on a questionnaire, codes on a medical abstraction form, etc.) do not always serve the purposes of the analysis. For example, a questionnaire on risk behaviors might use separate items to ask about use of crack, injected cocaine, injected heroin, and snorted heroin, but a single variable combining these behaviors ("yes" if used cocaine or heroin, "no" if used neither) might be more useful for the analyst. In that case a derived variable would be created (treatment of missing values becomes an issue here, as well). Similarly, a question about marital status and a question about living with a "romantic partner" might be combined into a variable indicating "living with spouse or partner".

3.7 Deciding which values to include in analyses

It is not always clear which values to include in analyses. For example, generally missing values are excluded from the denominator for computation of percentages, except when the purpose is an assessment of the extent of missing data. Sometimes, however, it is more meaningful to treat at least some categories of missing values in the same way as non-missing values. For example, a series of items about specific changes in behavior might be preceded with a screening question, such as "Have you make changes during the past year to reduce your risk of acquiring HIV?"

If the respondent answers "yes", s/he is asked about specific changes; otherwise the specific items are skipped. In this situation, a missing value due to the skip really means "no". This situation can be handled by creating a new variable for each of the specific items or by recoding the existing variables to "no" when the screening item was answered with "no" or other techniques. In contrast, a "true missing" would be present if the individual item was not answered even though the screening question was answered "yes". This "true missing" and would probably be excluded from the analysis. Similarly, if this type of behavior change was not relevant for the respondent, then the item is "not applicable" and the observation would probably be excluded as well ("probably", because the appropriate treatment depends upon the purpose of the analysis and intended interpretation).

3.8 Assessment of assumptions

During this stage, the assumptions underlying the statistical techniques are assessed. For example, a chi-square test has certain minimum expected cell sizes. A t-test assumes a Gaussian

(normal) distribution in the population. Other assumptions are those made about reality. For example, what if a person responds to the question on race by circling 3 responses, Black, Hispanic, and White. There is a study protocol to classify such an individual; however, this protocol may differ from other similar studies or the U.S. Census, or state birth certificates, etc. This may have an impact on the expected distribution and /or interpretation.

3.9 Examination of study questions

Data analyses may be approached in an exploratory fashion or in pursuit of answers to specific questions. Ideally the latter should have been specified in the research proposal or well before the analysis process has begun. Often new questions (or all questions) are formulated during the analysis process. In either case, it is highly desirable to articulate specific questions as a guide to how to proceed in the data analysis.

Besides their relevance to the questions at hand, analyses generally need to reflect the study design. For example, cross-sectional designs do not provide direct estimates of incidence, matched designs may warrant matched analyses.

Bibliography

Davidson, Fred. Principles of statistical data handling. Thousand Oaks, California, SAGE, 1996, 266pp.

Graham JW, Hofer SM, Piccinin AM. Analysis with missing data in drug prevention research. IN: Collins LM, Seitz LA (eds). Advances in data analysis for prevention intervention research. NIDA Research Monograph 142. U.S. D.H.H.S., N.I.H., National Institute on Drug Abuse, 1994, 13-63.

Marinez, YN, McMahan CA, Barnwell GM, and Wigodsky HS. Ensuring data quality in medical research through an integrated data management system. Statistics in Medicine 1984; 3:101-111.

Hybels, C. Data management outline. Presented at the American Geriatrics Society Summer Workshop. 1989.

Hse J. Missing values revisited. Presented at the all-Merck statisticians conference, October 23, 1989.

Hulley, Stephen B. and Steven R. Cummings. Designing clinical research: an epidemiologic approach. Baltimore, Williams & Wilkins, 1988. Chapter 15: Planning for data management and analysis.

Meinert, Curtis L.; Susan Tonascia. Clinical trials: design, conduct, and analysis. New York, Oxford, 1986.

Raymond, Mark R. Missing data in evaluation research. Evaluation & the health professions 1986;9:395-420.

Spilker, Bert; John Schoenfelder. Data collection forms in clinical trials. Raven Press, 1991.

Appendix

```
*************************************************************************;
*  The following SAS code can be adapted for use to create a check character
*  for ID numbers which can then be used to detect transcription errors when
*  the ID numbers are read again.  For example, numeric ID numbers can be
*  generated by any system and then suffixed or prefixed with a check character.
*  The ID's can be printed on questionnaire labels, specimen labels, coding
*  forms, or other data collection or tracking instruments.  When the ID numbers
*  are keyed along with the associated data, the data entry program can use
*  an adaptation of the code that follows to verify the accuracy of
*  transcription of the ID number itself.
*
*  The check character generated by the following code will detect transcription
*  errors involving a misrecording of any single digit of the ID number,
*  reversal of any two digits, or even multiple errors with relatively rare
*  exceptions.  Since errors in ID numbers can be among the most troublesome
*  to detect and correct, use of a check character is recommended.
*
*  The code on which this SAS routine was based was developed by
*  Robert Thorton at the Research Triangle Institute (RTI), based in turn
*  on an article by Joseph A. Gallian ("Assigning Driver's License Numbers",
*  Mathematics Magazine, February 1991, 64(1):13-22).  Thornton's code forms
*  the method for creating and checking ID numbers for the RSVPP study.
*  This SAS version was developed by Vic Schoenbach, 10/18/94, 10/24/94;
*
*  Here are some sample ID's and their corresponding check digits,
*  taken from a list of ID numbers provided by RTI to Project RAPP:
*
*         5-1120 -> S  (i.e., the complete ID number is 5-1120-S)
*         5-1111 -> T
*         5-1101 -> W
*         5-1011 -> A
*         5-1001 -> D
*         5-1002 -> B
*         5-2001 -> V
*         5-3001 -> Q
*
*************************************************************************;
*;

* This program reads a list of ID numbers and assigns check characters.
  The program also reads the check characters assigned by RTI so that these
  can be displayed alongside the calculated check characters to facilitate
  verification of the correctness of the calculated check characters,
  for testing purposes;

data;   * Create a SAS dataset with the original and calculated numbers;

  * Do not write the following variables into the dataset:      ;
    drop alphabet char1 lng sum i mod23 complem ;

* Define three variables:  the ID number, the check digit, & a 1 byte work area;
attrib strng length=$22 label='ID number needing check digit';
```

```
attrib ckd length=$1 label='Check digit to be calculated';
length char1 $1;  * For picking out one character at a time from the ID;
length sum 8;  * For calculation purposes;
alphabet= 'ABCDEFGHIJKLMNOPQRSTUVW';

infile cards;  * Input data file will be on "cards" (i.e., right after
                                                        the program);
input strng $  rti_ckd $ ;  * Read in data (consisting of ID's and the
                  RTI check digit, so that it can be printed in the output);

sum=0;  * Temporary variable to compute running sum;
lng=length(strng);  * Get length of ID to be processed;
if lng > 21 then do;  * Check that the ID is not too long;
        file print;
        put // '*** Error: ' strng= ' is too long = (' lng ')' //;
        file log;  return;  end;

do i = 1 to lng;    * Iterate through each digit of ID number ;
  char1 = substr(strng,lng-i+1,1);   * Extract a character from the ID;
                              * (Hyphens will be ignored);
  if char1 ^= '-' then
     if char1 < '0' or char1 > '9' then do;  * Must be a valid digit - if not
                                               then print error message;
       file print;
       put // '*** Error: Non-numeric character in ID: ' strng= char1=  //;
       file log; return;   * Go back for next ID number;
       end;   * End of then do;
     else do;  * (To get here, character must be a digit from 0-9);
       sum = sum + ((i+1) * char1);  * Take the sum of the digits of the ID
                                   number, weighting each digit by its position;
       end;   * End of else do;
end;    * End of do i = 1 to lng;

* Weighted sum has been obtained - now reduce it;
mod23 = mod(sum,23);  * Calculate the remainder after dividing by 23;
complem = 23 - mod23;  * Take the complement from 23;
ckd=substr(alphabet,complem,1);  * The check character is the
                                   corresponding letter of the alphabet;
return;

cards;   * Here come the test ID's -- note that one is invalid;
   5-1120  S
   5-1111  T
   5-11R1  W     (invalid ID number)
   5-1101  W
   5-1011  A
   5-1001  D
   5-1002  B
   5-2001  V
   5-3001  Q
run;               *(end of list of ID numbers);

* Display the results to verify correctness;
  proc print;  var _all_;  run;
```

17. Epidemiology and public health

Clinical versus the public health approaches

In their report of a major study conducted by the U.S. National Academy of Science's Institute of Medicine, the Committee for the Study of the Future of Public Health defined the mission of public health mission as:

> "the fulfillment of society's interest in assuring the conditions in which people can be healthy" (p 40)

The substance of public health was defined as:

> "organized community efforts aimed at the prevention of disease and the promotion of health. It links many disciplines and rests upon the scientific core of epidemiology." (p 41)

Public health focuses on the health of the community, but is a community an entity other than the people in a particular location or institutional unit? To begin exploring this question, let us first contrast two complementary approaches to maintaining and improving health – the clinical approach and the public health approach.

Clinical approach

The clinical approach deals with individuals, families. The provider's mission is to do what is best for the patient. Although it has been criticized for devoting insufficient attention to prevention, clinical medicine is not inherently tied to curative, rather than preventive approaches. In fact, in recent decades the time and resources devoted to preventing disease have greatly increased, especially in the realm of secondary prevention (e.g., management of hypertension and hypercholesterolemia). Pediatrics has long emphasized primary prevention.

What is more intrinsic to the clinical approach is the focus on the individual, or sometimes the family, in terms of diagnosis and intervention. Diagnostic inquiry is directed at the patient, e.g., her or his history, experiences, physiology, and so on. The scope of inquiry is primarily the prevention and treatment of medically recognized diseases, trauma, and psychiatric disorders.

Preparation of clinicians emphasizes core knowledge in biomedical sciences oriented towards understanding physiological and pathological processes, the effects of pharmacologic and surgical interventions, and techniques for investigation and intervention with the individual. In addition to allopathic medicine, numerous other approaches are offered in a clinical-type setting, including acupuncture, chiropractic, massage therapy, and many others. But the clinical encounter with an individual remains the framework.

Public health approach

The public health approach, in its ideal concept, deals with communities. The public health mission is to serve the community, even when particular individuals may well be disadvantaged in some way. There is some ambiguity in this statement, though, since any given population may be regarded as consisting of various "communities", whose interests are often perceived to differ. But typically public health focuses on a population or on subgroups within it.

The public health approach emphasizes prevention, though prevention in this context generally means preventing the occurrence of disease in individuals. At the level of the community, the distinction between prevention and cure may not be as clear.

The scope of public health is much broader than that of the clinical approach, because there is no framework of a clinical encounter to confine the time for diagnosis or intervention, and the variety of people and their situations in a community multiply the range of factors that can affect health. Therefore, in addition to specific and general causes of medically-recognized diseases, trauma, and psychiatric disorders, public health is concerned with the organization of society and the protection of the environment, and properly focuses on the future.

Public health providers have a small core of common training, due to the many fields of knowledge that become relevant when one deals with factors outside the individual. Channels for intervention are similarly broad, as they can deal with individuals, families, government organizations, the media, and the physical environment.

Contrasting the clinical and public health approaches

Two WHO reports on *in vitro* fertilization (IVF), published two years apart, illustrate the contrast in the clinical and public health approaches. The first (1990), issued by the WHO Regional Office for Europe in Copenhagen used a public health approach aimed at finding the best mix of curative and preventive health services, given existing resources, to maximize health status. The second (1992), issued by the WHO headquarters in Geneva, used a clinical approach to health policy development and focused on individual patients and their available treatment options. Here are some examples of these contrasting perspectives, taken from a commentary by Stephenson and Wagner (1993):

Prevention

- Copenhagen - options and recommendations for integration of preventive health services into an overall plan for the management of infertility in the community
- Geneva - no discussion of the prevention of infertility

Health services planning

- Copenhagen - a technology or procedure should have proven effectiveness, safety, and benefit as evaluated by clinical trials and other epidemiology methods, before acceptance as standard treatment.

- Geneva - " . . . IFV and allied procedures changed from being purely experimental in character to become accepted treatments for certain types of infertility and the numbers of centres offering them increased rapidly."

Rationing of health care

- Copenhagen - provision of services should be determined by the prevalence of the condition, the priority for infertility services within all human services, the medical and social options available to infertile people, and consumer views and choices. The public must have a voice in setting these priorities.
- Geneva - "Respect for the principle of quality of services requires the availability of medically assisted conception to the population requiring such service."

Standards of practice

- Copenhagen - recommendations for limits on age (40 years of age or younger), number of IVF treatment cycles per woman, and three eggs/embryos per IVF treatment cycle.
- Geneva - no recommendations

Research priorities

- Copenhagen - priority to epidemiological, social, and health services research
- Geneva - focuses on laboratory and clinical problems

The individual and population approaches have also been contrasted in regard to the epidemiology and prevention of sexually transmitted diseases and HIV (Aral et al., 1996).

Overlap

To be sure, there is considerable overlap between the two approaches, which at its best provides many opportunities for cooperation and complementary services and at its worst invites charges of duplication and turf wars. From the clinical side, the importance of prevention is being increasingly emphasized in primary care; from the public health side, interventions directed at the individual (e.g., inoculation, early detection and treatment, risk factor management) are typically carried out in one-on-one clinical settings. Pediatrics particularly has a strong orientation to prevention, and their are also disciplines of community medicine, community pediatrics, and social medicine.

There are also many activities and organizations that blend both clinical and public health approaches, as, for example, public health clinics, outreach services, patient education, clinical dietetics, clinical epidemiology, and questions of the availability, effectiveness, quality, and affordability of health services.

Obviously, both clinical and public health approaches are essential. Without health care at the individual level, much suffering occurs. Without public health, the brushfires of disease can easily overwhelm treatment resources. There is, however, a growing concern that the clinical approach has been gaining ascendency in confronting health needs out of proportion to the needs of public

health, particularly at the world level. Among the factors that favor the clinical approach over public health are:

- Symptoms and discomfort tend to motivate action much more than do theoretical concerns about low-level risks in the future.
- Individual victims of disease can be (or be made) highly visible and can elicit sympathy and a desire to help; by conrast, benefits from effective public health tend to be invisible and abstract.
- Effective treatment of a feared or disabling condition is highly visible and can be dramatic; by contrast, beneficiaries of effective public health measures typically do not think of themselves as being at risk nor as having benefited.
- Groups of individuals who have been affected by a disease can be highly influential in the political process; by contrast, public health benefits large groups, so specific individuals are not moved to action.
- Health care insurance systems provide an enormous revenue stream to support clinical services; by contrast, public health must compete with numerous other worthy constituencies for government appropriations.
- Clinical professions have many more people than do public health professions, which means more visibility, more potential letter-writers, and more membership dues for professional organizations.
- Much clinical care is delivered by the private sector, which has much greater ability to market its services and perspectives.

Thus, it is hardly surprising that resources devoted to health care services are orders of magnitude greater than those devoted to public health. Nevertheless, nations differ in their relative expenditure on public and private health services, and there are opportunities to influence the balance.through public education (a.k.a. marketing) campaigns.

Academic versus public health perspectives

As noted in an earlier chapter, the modern history of public health has been shaped by advances in scientific knowledge and technology, and growth in the public's acceptance that disease control is possible and a public responsibility. These advances have come from and contributed to a major expansion of epidemiologic research and training, including the development of epidemiology as an academic discipline. But the rise of academic epidemiology and its access to federal resources for research have had effects on the field that are not universally welcomed. To be sure, epidemiology continues to be the discipline that conducts surveillance for diseases in the population, identifies and prioritizes threats to health, designs control and preventive measures, and evaluates their effectiveness. In this role, epidemiologic research has strong links to the needs of public health authorities and direct applicability to important public health needs.

Since World War II, however, as the importance of scientific and biomedical research for modern societies has become apparent, epidemiology has developed a strong role as a "basic" science and a position of growing respect among academic researchers. This role has fundamental importance for

public health, since the best opportunities to prevent disease and improve health often come from advances in basic understanding of the causes of disease, the development of new methods to study them, and the assessment of preventive and control measures. Nevertheless, there is an abiding concern about the weakening of the link between public health practitioners and academic epidemiologists, imbalances between allocation of research funding and importance of public health problems, and the forces that draw epidemiologists' efforts toward what is perceived as scientifically and academically valuable and but further away from public health needs.

This concern has been expressed by major figures in epidemiology and public health. Nearly 20 years ago, Milton Terris (The epidemiologic tradition. *Public Health Reports* 1979;94(3):203-209) objected to the growing divide between academic epidemiology and public health practice, and Lilienfeld and Lilienfeld (1982:147-148) and Mervyn Susser have warned about the overemphasis on technique. The Committee for the Study of the Future of Public Health also made a number of strong criticisms of schools of public health. Cecil Sheps has warned about the "substitution of method for meaning".

How can teaching and research be in conflict with the mission of public health? There are many aspects to this question, but one is the familiar question of where to set priorities when not everything can be done. Although biomedical research has led to remarkable discoveries and capabilities, in many instances it is possible to accomplish a great deal of prevention without the full knowledge of the pathogenic agent. In the words of the late Ernst Wynder, ". . . as we reflect on the history of medicine, we may conclude that the complex disease entitles of the twentieth century, like the diseases of the past, will respond first to preventive strategies on the basis of new knowledge as well as of information already at hand." (EL Wynder, *Am J Epidemiol* 1994:549). Wynder provides these examples:

Comparison of the date of discovery of a measure to prevent a disease with the date of identification of its true causative or preventive agent

Disease	Discoverer of preventive measure	Year of discovery preventive measure	Year of discovery of agent	Discoverer of agent
Scurvy	J. Lind	1753	1928	A. Szent-Gyorgi
Pellagra	J. Goldberger	1755	1924	G. Casal et al.
Scrotal cancer	P. Pott	1775	1933	J.W. Cook et al.
Smallpox	E. Jenner	1798	1958	F. Fenner
Puerperal fever	I. Semmelweis	1847	1879	L. Pasteur
Cholera	J. Snow	1849	1893	R. Koch
Bladder cancer[a]	L. Rehn	1895	1938	W.C. Hueper et al.
Yellow fever	W. Reed et al.	1901	1928	A. Stokes et al.
Oral cancer[b]	R. Abbe	1915	1974	D. Hoffmann et al.

Causative or preventive agents

Scurvy	(Ascorbic acid)
Pellagra	(Niacin)
Scrotal cancer	Benzo(a)pyrene
Smallpox	Orthopxovirus
Puerperal fever	Streptococcus
Cholera	Vibrio cholerae
Bladder cancer[a]	2-Naththylamine
Yellow fever	Flavivirus
Oral cancer[b]	N-nitrosonomicotine

[a] associated with aniline dye; [b] associated with tobacco chewing

Source: Wynder EL. Invited commentary: studies of mechanism and prevention. *Am J Epidemiol* 1994:547-549, Table 1.

The current health profile of the people of the world as a whole and of the United States (especially among minority groups) highlights many health problems where the application of existing scientific and medical knowledge could bring major improvements. It has been argued that nearly half of deaths in the United States could be prevented by the application of existing medical knowledge.

Deaths from Preventable Causes in the United States in 1990

Cause	Estimated No. of Deaths	Percentage of Total Deaths
Tobacco	400,000	19
Dietary factors and activity patterns	300,000	14
Alcohol	100,000	5
Microbial agents	90,000	4
Toxic agents	60,000	3
Firearms	35,000	2
High-risk sexual behavior	30,000	1
Motor vehicle injuries	25,000	1
Illicit use of drugs	20,000	<1
Total	1,060,000	49

Source: Carl E. Bartecchi, Thomas D. MacKenzie, Robert W. Schrier. The human costs of tobacco use (first of two parts). *New Engl J Med* 330;1994:907-912, Table 1, page 908. Reprinted from McGinnis JM and Foege WH. Actual causes of death in the United States. *JAMA* 1993;270:2207-12. Values are composite approximations drawn from studies that use different approaches to derive estimates, ranging from actual counts (e.g., firearms) to calculations of population-attributable risk (e.g., tobacco). The numbers have been rounded.

Individual-level versus societal level perspectives

The reasons – behavioral, social, political, and economic factors – for the lack of application of existing knowledge are rarely the subject of epidemiologic inquiry. Moreover, these factors are also the major determinants of health in populations, so that their position outside of the scope of epidemiology greatly restricts epidemiology's potential for improving health.

Geoffrey Rose (1985) has argued that concentration on the person as a unit and on a lessening of personal risk has led to the neglect of populations and of the preventive goal of reducing incidence. Similarly, Nancy Krieger (1994) has criticized definitions of epidemiologic theory that emphasize concepts pertaining to study design and causal inference, and ignore issues of what drives societal patterns of health and disease.

Poole (1994) contrasts two perspectives on the nature and role of epidemiology. In the first viewpoint (which he identifies with Milton Terris and Mervyn Susser), health of a group, cohort, community, or a people is more than the summation of the health of its individual members. Public health's special province is this "more". From this viewpoint, epidemiology "is not so much the study of disease and health IN human populations as the study of disease and health OF human populations" (Poole). Epidemiology is seen as a social science (a population science) that focuses on the forest, rather than on the trees.

In what Poole refers to as the newer view (advanced by Ken Rothman and Sander Greenland), epidemiology is seen "as a type of medical research, as a way of using populations to obtain biologic knowledge about disease and health in individual persons". Here, epidemiology is seen as natural science, the health of the population is the summation of health of individuals, and public health is medicine for the masses with an emphasis on prevention. This view presents epidemiology as a dispassionate science, rather than an activist one.

Multilevel statistical models (also called hierarchical regression models and various other names) represent a partial answer to this conflict, since they allows for the inclusion of both individual-level and group-level variables in the same regression model. However, while multilevel modeling addresses the statistical issues of correct estimation when variables are measured at different levels, the conceptual model and theoretical aspects, which lies at the heart of the debate, remain.

While the first viewpoint described by Poole tends to be associated with public health activism, it is certainly possible to focus on societal level factors without endorsing or promoting any particular course of action. The societal perspective may be more congenial to activists in that it appears to invite advocacy more directly than does the individual-level perspective. But many individual level factors (e.g., immunization, nutrition, tobacco use, fitness) are powerfully influenced by the social environment, which argues for an activist stance in regard to individual-level relationships as well. In some respects, therefore, the debate between the two viewpoints contrasted by Poole is another version of the debate, discussed in the first chapter, about whether epidemiology is more properly a science or a public health profession that includes advocacy as part of the job description.

Human behavior is also biology

The debate about individual-level versus societal-level viewpoints is likely to evaporate for several reasons. Perhaps the most important of these is that as society and scientific knowledge evolve the interacting influences of individuals and the environment become increasingly apparent and important. Advances in genetic science and technology, including the mapping of the human genome, are greatly expanding the possibilities to understand disease processes at the individual level. But as this understanding unfolds it will, of course, disclose environmental (in the broadest meaning of the term) influences. Indeed, identification of susceptibility genes will increase the power of epidemiologic studies to identify environmental factors, since inclusion of nonsusceptible persons weaken associations. At the same time, advances in understanding of societal factors will make clear the need to understand the individuals whose individual and collective behavior creates and maintains those factors (Schoenbach 1995).

Since the human species is, after all, a part of the animal kingdom, full understanding of human behavior requires a biological perspective as well as the perspectives of the psychological, sociological, economic, and political sciences. That biological perspective must encompass influences related to genetic factors, environmental exposures (e.g., lead), prenatal exposures, nutritional factors, pharmacologic factors, and neuroanatomical/neuroendocrinological effects of past experiences (e.g., nurturing, violence). It must also take account of behavioral and cognitive tendencies that our species has acquired in our journey through evolutionary time.

As our population numbers and density increase, and the growth of technology and organizations magnifies our potential impact, human behavior becomes an increasingly important factor on society and on the environment. One area where this impact is evident is war and conflict. In addition to millions upon millions of deaths from political, ethnic, and religious violence in the past century (an illustrative list: Armenia, Bosnia, Cambodia, Chechnya, China, Congo, Egypt, Korea, Kosovo, Lebanon, India, Iran, Iraq, Ireland, Israel, Japan, Russia, Rwanda, Spain, Syria, Timor, Vietnam – plus World Wars I and II and innumerable colonial wars) represent a direct impact, armed conflict devastates public health infrastructures, physically and psychologically maims many of the survivors, destroys agriculture and industry, creates massive numbers of displaced persons, and harms the environment. Nuclear war, the most dramatic anti-social behavior, could render irrelevant virtually all epidemiologic achievements. The ability of individual or small groups of terrorists to harm large numbers of people is attracting heightened attention as a result of such incidents as the Oklahoma and World Trade Center bombings and the sarin gas attack in Tokyo (and the belief that the organization responsible for the latter was also trying to obtain specimens of ebola virus).

Even more profound than these blatant harms to human life and health, however, may be the growing imbalance between population and environmental resources. Such imbalances are a familiar phenomenon in nature – and a temporary one, since population size adjusts to fit within available resources.

World population growth and urbanization

By 2030, world population is expected to grow to over eight billion from the current six billion (Lutz, 1994). Meanwhile the industrialized countries' share of population is expected to shrink to

14%, so that the burden of the environments in developing countries will intensify greatly. The impacts of population size on life, the environment, and public health are manifold and sometimes complex. The age structure of the population, its geographical distribution, and many other factors all influence the impact of population size. The governments of the world have yet to accept fully that there is an upper limit to the earth's carrying capacity. In 1982 the United Nations Food and Agriculture Organization (FAO) estimated that under optimal conditions the world could support over 30 billion people, though a more realistic figure for food sufficiency is 10 to 15 billion, a range that the world is projected to reach by the year 2050 (Lutz, 1994).

Population growth rates are a function of birth and death rates. Crude death rates are very similar between the developing countries as a whole and the developed countries, because the former have a much younger age structure (average age in 1990 was 38 years in Western Europe, 22 years in sub-Saharan Africa) (Lutz, 1994). Birth rates in the developing world are much higher, with only China, Hong Kong, and Taiwan having birth rates below 20 per 1,000 persons. Both younger age structure and higher total fertility rates (lifetime number of births/woman) are responsible for the higher birth rates. Although there are many uncertainties that underlie projections of birth rates, mortality, and population growth, "'The question is not 'if' world population will grow, but rather 'how big' will it become." (Lutz 1994:34).

Birth rates in urban areas are generally smaller than those in rural areas, but urban areas also grow through rural-urban migration. Growing urbanization is bringing dramatic changes which are being largely ignored in thinking about the future (Melinda Meade, UNC Department of Geography, in a 1998 seminar). In 20 years, India will double in size, adding 900 million people to its cities. Lagos, Nigeria will grow to 25 million. According to Meade, we are approaching a qualitative change.

Historically, Meade explains, many communicable diseases flourished when the development of cities created adequate population density for microbes like measles. But urbanization in the U.S. was "stepped migration", the classical pattern – people move from farm to town, then to a nearby city, then to a distant, larger city, acquiring an urban lifestyle in the process. In contrast, urbanization in the developing world is "chain migration" – people go directly from villages to cities, sometimes even bringing their farm animals with them. U.S. cities grew at 1%, doubling in 70 years. Many Asian and African cities are growing at 7%, doubling in 10 years!

Meade explains further that urbanization, especially rapid urbanization, provides a larger host population for communicable diseases, more interaction (especially in a service economy), and shortages of pure water and sewage treatment. Urbanization brings changes in the host population (genes, gender, age), habitat (natural → built, social), and behavior (beliefs, social organization, technology). Urbanization leads to draining marshes, introducing artificial irrigation, and deforestation, all of which promote different species of vectors. For example, new disease vectors are developing that "like" organically polluted water. Bubonic plague had come to Europe before the Black Death but did not spread wildly because of the absence of rats in Europe. Enormous population growth in Europe in the Middle Ages overwhelmed the habitat – agriculture, sewage, grain storage, fluctuating yields – led to a large rat population and poor/malnourished human population, creating the conditions for the spread of plague. In fact, outbreaks of threatening communicable diseases, including plague itself, are a present reality (and if it can be characterized as such, a fascinating saga – see Laurie Garrett's *The coming plague*). Besides communicable diseases,

crowded, under-resourced urbanized areas spawn massive shantytowns and high rates of unemployment, desperation and crime. Unbreathable air and depletion of water supplies are major issues. For a vivid and disquieting portrait of some of these situations, see Robert D. Kaplan, The coming anarchy (*Atlantic monthly*, February 1994; 273:44-76; available at `http://www.theatlantic.com/politics/foreign/anarchy.htm`).

Global epidemiology?

Accurate knowledge is an essential for effective action. As illustrated by Ernst Wynder's examples, even partial knowledge can lead to successful prevention. However, partial knowledge can also lead to exchanging one set of problems for another, perhaps worse than those that motivated the original actions. Sir Austin Bradford Hill (1968: 300) wrote that the incomplete and tentative nature of scientific knowledge "… does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time." But the judgment of what action is demanded by existing knowledge is often complex and controversial.

The debate between contrasting views of epidemiology outlined earlier reflects to some extent the conflict between the desire to be confident in one's methods and data on the one hand and the need to tackle the major problems that confront public health. But that conflict is one for individuals to resolve in choosing where to work and what to work on, rather than a decision for the field. If epidemiology confines itself to studying biomedical questions that it has the tools for studying, to whom does it leave the other problems that confront public health? If the study of health in human populations is epidemiology, then whether the people who tackle these problems call themselves medical geographers, biological anthropologists, or epidemiologists, they will be practicing epidemiology. Challenges to human health are not constrained by the availability of methodologies to study them.

In principle, and increasingly in practice, the purview of epidemiology extends to the fauna and flora of the planet and their global environment. The importance of developing a global perspective becomes clearer every decade, as advances in science, production, transportation, and communication, with the accompanying changes in human activity, have created the conditions for global epidemics, global contamination, conflict between peoples separated by great distances, and even modification of the planet (McMichael 1993). In his book *Planetary Overload*, Anthony McMichael (1993) identifes international inequality as the key issue that must be addressed in order to protect the global environment on which human health depends:

1. The "one underlying problem is the entrenched inequality between rich and poor countries, which predominantly reflects recent imperial history, power relationships and the global dominance of Western industrial technology and economic values." (p. 7)
2. The "two central manifestations of this inequality are:
    1. rapid, poverty-related, population growth and land degradation in poor countries, and
    2. excessive consumption of energy and materials, with high production of wastes, in rich countries." (p. 7)
3. The "three possible (perhaps coexistent) adverse outcomes of those manifestations are:

1. exhausting various non-renewable materials,
2. toxic contamination of localised environments, and
3. impairment of the stability and productivity of the biosphere's natural systems." (p. 7)

Although the study of the world's people and our environment, living and nonliving, can neither be claimed by nor contained within any discipline or field, epidemiology's multidisciplinary perspective draws, as a matter of course, from all fields of knowledge. In that respect, epidemiology is as logical a field as any to include the study of global health, in its broadest interpretation, within its scope. John Last made this very point in accepting the Abraham Lilienfeld Award from the American College of Epidemiology: "There is a need for innovative, transdisciplinary approaches. Epidemiology is already transdisciplinary. Epidemiology is well placed to take leadership." (American College of Epidemiology Annual Meeting, Boston, September 22, 1997).

Bibliography

Annas, George J.; Leonard H. Glantz, Norman A. Scotch. Back to the future: the IOM Report reconsidered. Editorial. *Am J Public Health* 1991; 81:835-837.

Aral, Sevgi; King K. Holmes, Nancy S. Padian, Willard Cates, Jr. Overview: individual and population approaches to the epidemiology and prevention of sexually transmitted diseases and human immunodeficiency virus infection. *J Infectious Dis* 1996;174(Suppl 2):S127-33.

Colditz GA. Epidemiology – future directions. Intl J Epidemiol 1997; 26:693-697.

Frank SM. Changing demographics in the United States. Implications for health professionals. *Cancer* 1991;67(6 Suppl):1772-8.

Hill, Austin Bradford. The environment and disease: association or causation? *Proceedings Royal Society Medicine* 1965;58:295-300.

Institute of Medicine (U.S.) Committee for the Study of the Future of Public Health. *The future of public health*. Washington, D.C., National Academy of Sciences, 1988.

Krieger, Nancy. Epidemiology and the web of causation: has anyone seen the spider? *Soc Sci Med* 1994;39:887-903.

Lilienfeld, Abraham M. and David E. Lilienfeld. Epidemiology and the public health movement: a historical perspective. *Journal of Public Health Policy* 1982; 3:140-149:

Lutz, Wolfgang. The future of world population. *Population Bulletin* June 1994;49(1):2-47.

McMichael, Anthony J. *Planetary overload: global environmental change and the health of the human species*. NY, Cambridge, 1993.

Murray, Christopher J. L., Alan D. Lopez. *The global burden of disease*. Harvard School of Public Health on behalf of the World Health Organization and the World Bank, 1996. Summary by Phyllida Brown. Distributed by Harvard University Press.

Murray, Christopher J. L., Alan D. Lopez. *Global health statistics: a compendium of incidence, prevalence, and mortality estimates for over 200 conditions*. Harvard School of Public Health on behalf of the World Health Organization and the World Bank, 1996. Distributed by Harvard University Press.

National Center for Health Statistics. *Health, United States, 1993*. Hyattsville, MD: Public Health Service, 1994.

O'Hare WP, Pollard KM, Mann TL, Kent MM. African Americans in the 1990s. *Population Bulletin* July 1991; 46(1):2-40, p4.

O'Hare WP. America's Minorities--the demographics of diversity. *Population Bulletin* December 1992;47(4):2-47 (Washington, D.C.: Population Reference Bureau, Inc.)

Poole C. Ecologic analysis as outlook and method. Editorial. *Am J Public Health* (May) 1994;84(5):715-716

Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985;14:32-38.

Schoenbach VJ. Behavioral epidemiology: expanding the boundaries. Poster presentation, American College of Epidemiology, Annual Meeting, St. Louis, MO, September 1995.

Shy, Carl M. The failure of academic epidemiology. Am J Epidemiol 1997;145:479-484. Also, commentary by Alexander M. Walker and Dr. Shy's reply, 485-487.

Smillie, Wilson G. *Public health: its promise for the future*. New York, Macmillan, 1955.

Stephenson PA, Wagner MG. WHO recommendations for IVF: do they fit with "Health for All"? *Lancet* 1993;341:1648-1649.

Susser M. Epidemiology in the US after World War II: the evolution of technique. *Epidemiologic Reviews* 1985;7:174-177.

Terris, Milton. The epidemiologic tradition. *Public Health Reports* 1979;94(3):203-209. (See Topic I2 Orientation and Introduction)

18. Overview and Conclusion

A look backward on what has been covered and forward towards where the field is going.

What is an "epidemiologic perspective"?

- Population orientation – increasingly, a global orientation
- Problem-oriented / multidisciplinary
- Breadth/challenge – in principle can address any health-related problem – or *any* problem?
- Prevention emphasis
- Sees society as the organism – interconnectedness among all facets

Epidemiology "successes" and "not-yet-successes"

- Cholera - John Snow
- Pellagra - Joseph Goldberger
- Rubella and birth defects - Gregg
- Fluoride and dental caries
- Cigarette smoking and lung cancer
- Blood pressure, cholesterol, smoking, fitness and CHD/CVD
- Legionnaire's disease
- Breast cancer
- Prostate cancer
- Pancreatic cancer
- Adolescent pregnancy
- Sexually transmitted diseases
- HIV/AIDS
- Drug use
- Violence
- Environmental protection

- Health for all

What contributes to success?

- Specific disease definition
- Biological reasoning and measures, laboratory research
- Individual-level measures of disease
- Heterogeneity of exposure within groups
- The epidemiologic perspective handed down over generations. But the practice of epidemiology as we know it is largely a product of the last 50 years.

Epidemiologic methods have developed rapidly

- Elaboration of epidemiologic theory - case-control studies, epidemiologic measures, randomized trials
- Computing revolution - data management, database linkage, mapping and geographical databases, computer-based data collection
- Statistical analysis methods - many new techniques, e.g., logistic regression, proportional hazards, longitudinal analysis, simulations
- More accurate and precise measures - revolutions in biochemical and molecular biology
- Communications, organizational, and management innovations - for large studies
- But there is also a recurrent concern about the effects on the field of some of these advances, e.g.:

  "Perhaps the most dangerous aspect of the state of our discipline today is that there is an unhealthy emphasis on HOW one conducts an epidemiologic study and not WHY and WHAT one does in such a study. Simply put, we are training technocrats. As Lionel Beak so aptly stated (14): 'In teaching, there is often excessive emphasis on how rather than what or why. Efforts are made to train men [sic!] who are technically competent. The end has been more vocationalism. … And many administrators, and faculty, who have played a significant role in bringing this about readily assume that this is how it must or should be.' This trend has been further emphasized by technologic developments in computation which allow one to deal automatically with masses of data in a mechanical and thoughtless manner. More attention and emphasis must be given to reasoning about the various types of data that are collected and analyzed."

  147-148 in Abraham M. and David E. Lilienfeld. Epidemiology and the public health movement: a historical perspective. *Journal of Public Health Policy* 1982; 3:140-149.

Epidemiology is expanding

- In recent decades, epidemiology has enjoyed enormous growth, expanding opportunities and horizons, and growing recognition from other disciplines:
- Growing awareness of public health-related issues and acceptance (after World War II) of role of government
- Chronic diseases - substantial National Institute of Health funding for epidemiologic research
- Growth of environmental and occupational health regulations - epidemiology is a major source of the evidence
- Litigation - epidemiology in the courtroom (Benedectin, breast implants, tobacco liability, …)
- Epidemiology is increasingly seen as source of research skills and techniques for clinical research
- Corporate as well as public sponsors
- Surge of managed care is creating new demands
- New schools of public health, new epidemiology departments, new research units
- International expansion

Is epidemiologic research becoming more difficult?

- Rarer conditions, larger studies
- Very low level exposures
- Subtle relationships/weak effects
- Constructs difficult to define and measure (psychiatric, behavioral, psychosocial, community) as outcomes and exposures
- Understudied populations - researchers unfamiliar, populations disaffected and distrustful, ethical and political concerns
- Greater sensitivity to human subjects issues - truly informed, truly consenting, privacy protection
- Intervention studies [" . . . I think that we need to face up to the difficulties of doing intervention trials. We talk of experimental epidemiology, but we do very little of it. It is extremely difficult. … I think that we just need to face up to the need for doing more experimental epidemiology." (Sir Richard Doll, interview with *Epidemiology Monitor*)]

Challenges in the environment for epidemiology

- Rising expectations of what epidemiology can do and how quickly – the public (and sponsors) wants not just leads, but answers.
- Media interest and publicity – too much and too soon? (abetted by marketing imperatives and fund-raising).
- Link between academic epidemiology and public health practice has weakened – academic epidemiology has its own perspectives and objectives – Milton Terris argues that the rise of academic epidemiology has led to an overemphasis on statistics, analysis, and hypothesis tests at the expense of biological thinking and hypothesis creation.
- Universities increasingly dependent upon research project funding. → Not "funding for what?", but "what for funding?": The ivory tower → The ivy-covered corporate tower?
- Competing priorities for public funds
- All sponsors are looking for marketable results, impact

What sets priorities for research funding

- Public health policy process (Objectives for the Nation)
- New, expanding, and feared diseases (HIV, TB, Alzheimer's disease)
- Increased recognition for existing problems (injury, teen pregnancy)
- Political process (cancer, HIV, minority health, women's health, aging)

"Academic-Industrial Complex" (cf. Eisenhower's warning about the Military-Industrial Complex)

- Peer review, peer influence
- Research institutions
- Drug industry sponsorships of research, conferences, publications
- Insurance industry
- Corporate health care
- American Medical Association – political contributions and lobbying

Limits on funding

- Era of limits
- Costs are rising - inflation, technology, expectations, quality, Big Science
- More investigators, more institutions

- More reliance on "soft money" – research funding as an engine of growth

Growing ability to meet challenges

- Researchers - more and better trained
- Increasing diversity (gradual!) in the profession
- Theoretical and methodologic development (EPID leisure class)
- Record keeping and bureaucratization - megagovernment, megacorporations, megahealthcare - computerized information
- Computers and software - more powerful, more available, more friendly, more customized, more intelligent
- Measurement (automatic readout, continuous monitoring)
- Pattern recognition (e.g., ECG reading, CT scan)
- Record linkage
- Routine surveillance/follow-up
- Larger datasets
- New analytic procedures
- Molecular biology revolution
- New assays
- New understanding ["So I think that epidemiologists have to become much more biochemically and biologically minded than some are nowadays." (Sir Richard Doll, interview with the Epidemiology Monitor]

Epidemiologists' wish list

- Biological markers of past exposure (e.g., diet) (need a "C14 for epidemiology")
- Ways to measure social and behavioral variables
- Ways to understand social factors and disease in the context of social as well as physical environment

Some fundamental questions

What is epidemiology?

- Do epidemiologists compromise their scientific credibility if they become advocates?
- Must epidemiology deal with "disease" or can it address any event, condition, or characteristic?

- What kind of population is required to be "population-based" – geographic?, worksite?, health care provider?, patients with a medical condition?, . . .?

What are the goals of epidemiology / public health?

- Prolong life? How long? Life expectancy 80 years?, 90 years?, 100 years? 150 years?
- Should we extend life as long as we can consistent with good quality of life? How much life does a generation have a "right" to? What is our generation's "fair share"?
- Does the effect on the environment matter? Can the earth become too crowded?
- What do we think about a guest who never leaves?

What determines health?

- What we don't know (e.g., Alzheimer's disease, arthritis, breast cancer, pancreatic cancer, prostate cancer) or don't know enough (cardiovascular disease, stroke, . . .)
- What we do know but don't know how to change (e.g., smoking, drugs, violence, risky sexual behavior, . . .)
- What we know how to change but do not (e.g., pure drinking water, good sanitation, immunization, breast feeding, preventive health care, environmental protection, unplanned pregnancy, sexually transmitted diseases, food, housing, transportation, physical security, . . .)
- Collective consciousness?

Is public health a noble calling?

Many people pursue self-aggrandizement. Public health professionals pursue a better life for all. But we also want to be paid to do that. Thus we experience diverse and sometimes conflicting attractions, responsibilities, and demands:

Science	Management
Curiosity	Quality control
Imagination	Personnel
Creativity	Regulations
Collegiality	Money
Dissemination	Public relations
Idealism	Practicality
Pursue knowledge	Get a job
and understanding	Get grants
Improve public health	Get publications
Help the disadvantaged	Get more grants
Share freely	Get known
Assist others	Get ahead

This is not a new challenge*.

Epidemiology seeks knowledge to improve health for all. Knowledge may not be enough to improve health. Powerful forces – geologic, meteorologic, microbiologic, economic, cultural, political -- work to counter changes that would advance public health (e.g., lead, tobacco, global warming, handguns, reproductive health, political extremism, pollution, war and violence). But knowledge is certainly key in alerting us that change would be beneficial and can help to build a consensus to bring about change. Can knowledge reveal how to reconcile conflicting imperatives among economics, politics, religion, culture, ecology, and health? That may be the ultimate challenge for epidemiology.

---

* My teacher, Bert Kaplan, is fond of quoting the renowned rabbi, Hillel: "He used to say, If I am not for myself, who will be for me? And if I am only for myself, what am I? And if not now, when?" *Sayings of the Fathers (*or *Pirke Aboth)*, translated by Joseph H. Hertz. NY, Behrman House, 1945, I-14. [The commentary adds that "for myself" is "far more than merely a rule of worldly wisdom. 'If I do not rouse my soul to higher things, who will rouse it?' (Maimonides)."]

Bibliography

American College of Epidemiology. Epidemiology and minority populations: statement of principles. *Ann Epidemiol* 1995; 5(4).

Evans, Alfred S. Subclinical epidemiology: the first Harry A. Feldman Memorial Lecture. *Am J Epidemiol* 1987; 125:545-556.

Gordis L.: Challenges to epidemiology in the next decade. *Am J Epidemiol* 1988; 128:1-9.

Gordis L.: Challenges to epidemiology in the coming decade. *Am J Epidemiol* 112:315-321, 1980.

Graham, Saxon. Enhancing creativity in epidemiology. *Am J Epidemiol* 1988; 128:249-253.

Greenberg, Bernard G. The future of epidemiology. *J Chron Dis* 1983; 36:353-359.

Greenhouse, Samuel W. Some epidemiologic issues for the 1980s. *Am J Epidemiol* 1980; 112:269-273

Kuller, Lewis H. Epidemiology is the study of "epidemics" and their prevention. *Am J Epidemiol* 1991 (Nov 15); 134(10):1051-

Last, John M. Acceptance speech for the Abraham Lilienfeld Award. American College of Epidemiology Annual Meeting, Boston, September 22, 1997.

"Outgoing SER President addresses group on faith, evidence, and the epidemiologist." *Epidemiology Monitor* July 1983; 4(7):3-4.

Susser, Mervyn. Epidemiology today: a 'thought-tormented world'. *International Journal of Epidemiology* 1989; 18(3):481-487.

Terris, Milton. The changing relationships of epidemiology and society: the Robert Cruikshank Lecture. *Journal of Public Health Policy* 1985; 15-36.

Terris, Milton. The Society for Epidemiologic Research (SER) and the future of epidemiology. *Am J Epidemiol* 1992;136(8):909-915.