Moyses Szklo | Javier Nieto

Epidemiology

Beyond the Basics

FOURTH EDITION

Epidemiology
Beyond the Basics

FOURTH EDITION

Moyses Szklo, MD, DrPH
University Distinguished Service Professor
of Epidemiology and Medicine
Johns Hopkins University
Editor in Chief, *American Journal of Epidemiology*
Baltimore, Maryland

F. Javier Nieto, MD, PhD
Dean and Professor of Epidemiology
College of Public Health and Human Sciences
Oregon State University
Corvallis, Oregon

JONES & BARTLETT
LEARNING

World Headquarters
Jones & Bartlett Learning
5 Wall Street
Burlington, MA 01803
978-443-5000
info@jblearning.com
www.jblearning.com

Jones & Bartlett Learning books and products are available through most bookstores and online booksellers. To contact Jones & Bartlett Learning directly, call 800-832-0034, fax 978-443-8000, or visit our website, www.jblearning.com.

Substantial discounts on bulk quantities of Jones & Bartlett Learning publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones & Bartlett Learning via the above contact information or send an email to specialsales@jblearning.com.

Production Credits
VP, Product Management: David D. Cella
Director of Product Management: Michael Brown
Product Specialist: Carter McAlister
Production Editor: Vanessa Richards
Senior Marketing Manager: Sophie Fleck Teague
Manufacturing and Inventory Control Supervisor: Amy Bacus
Composition: S4Carlisle Publishing Services
Cover Design: Scott Moden
Rights & Media Specialist: Merideth Tumasz
Media Development Editor: Shannon Sheehan
Cover Image (Title Page, Part Opener, Chapter Opener): © Enculescu Marian Vladut/Shutterstock
Printing and Binding: Edwards Brothers Malloy
Cover Printing: Edwards Brothers Malloy

Library of Congress Cataloging-in-Publication Data
Names: Szklo, M. (Moyses), author. | Nieto, F. Javier, author.
Title: Epidemiology : beyond the basics / Moyses Szklo, F. Javier Nieto.
Description: Fourth edition. | Burlington, Massachusetts : Jones & Bartlett Learning, [2019] | Includes bibliographical references and index.
Identifiers: LCCN 2017061666 | ISBN 9781284116595 (pbk.)
Subjects: | MESH: Epidemiologic Methods | Epidemiologic Factors | Epidemiology
Classification: LCC RA651 | NLM WA 950 | DDC 614.4--dc23
LC record available at https://lccn.loc.gov/2017061666

6048

Printed in the United States of America
22 21 20 19 18 10 9 8 7 6 5 4 3 2 1

Contents

Preface

This book was conceived as an intermediate epidemiology textbook. Similar to previous editions, the fourth edition discusses key epidemiologic concepts and basic methods in more depth than that found in basic textbooks on epidemiology. For the fourth edition, new examples and exercises have been added to all chapters. In addition, several new topics have been introduced, including the use of negative controls to evaluate for the presence of confounding, a simple way to understand adjustment using multiple regression, use of a single analytic unit, and translational epidemiology. Some concepts that were discussed in previous editions have been expanded, including efficacy and ways to conceptualize a control group.

As an intermediate methods text, this book is expected to have a heterogeneous readership. Epidemiology students may wish to use it as a bridge between basic and more advanced epidemiologic methods. Other readers may desire to advance their knowledge beyond basic epidemiologic principles and methods but are not statistically minded and, therefore, reluctant to tackle the many excellent textbooks that strongly focus on epidemiology's quantitative aspects. The demonstration of several epidemiologic concepts and methods needs to rely on statistical formulations, and this text extensively supports these formulations with real-life examples, thereby making their logic intuitively easier to follow. The practicing epidemiologist may find selected portions of this book useful for an understanding of concepts beyond the basics. Thus, the common denominators for the intended readers are familiarity with the basic strategies of analytic epidemiology and a desire to increase their level of understanding of several notions that are insufficiently covered (and naturally so) in many basic textbooks. The way in which this textbook is organized makes this readily apparent.

In Chapter 1, the basic observational epidemiologic research strategies are reviewed, including those based on studies of both groups and individuals. Although descriptive epidemiology is not the focus of this book, birth cohort analysis is discussed in some depth in this chapter because this approach is rarely covered in detail in basic textbooks. Another topic in the interface between descriptive and analytical epidemiology—namely, ecological studies—is also discussed, with a view toward extending its discussion beyond the possibility of inferential (ecological) bias. Next, the chapter reviews observational studies based on individuals as units of observation—that is, cohort and case-control studies. Different types of case-control design are reviewed. The strategy of *matching* as an approach by which to achieve comparability prior to data collection is also briefly discussed.

Chapters 2 and 3 cover issues of measurement of outcome frequency and measures of association. In Chapter 2, absolute measures of outcome frequency and their calculation methods are reviewed, including the person-time approach for the calculation of incidence density and both the classic life-table and the Kaplan-Meier methods for the calculation of cumulative incidence. Chapter 3 deals with measures of association, including those based on relative (e.g., relative risk, odds ratio) and absolute (attributable risk) differences. The connections between measures of association obtained in cohort and case-control studies are emphasized. In particular, a description is given of the different measures of association (i.e., odds ratio, relative risk, rate ratio) that can be obtained in case-control studies as a function of the control selection strategies that were introduced in Chapter 1.

Chapters 4 and 5 are devoted to threats to the validity of epidemiologic studies—namely, bias and confounding. The "natural history" of a study is discussed, which allows distinguishing between these two concepts. In Chapter 4, the most common types of bias are discussed, including selection bias and information bias. In the discussion of information bias, simple examples are given to improve the understanding of the phenomenon of misclassification resulting from less-than-perfect sensitivity and specificity of the approaches used for ascertaining exposure, outcome, and/or confounding variables. This chapter also provides a discussion of cross-sectional biases and biases associated with evaluation of screening procedures; for the latter, a simple approach to estimate lead time bias is given, which may be useful for those involved in evaluative studies of this sort. In Chapter 5, the concept of confounding is introduced, and approaches to evaluate confounding are reviewed. Special issues related to confounding are discussed, including the distinction between confounders and intermediate variables, residual confounding, the role of statistical significance in the evaluation of confounding effects, and the use of negative controls.

Interaction (effect modification) is discussed in Chapter 6. The chapter presents the concept of interaction, emphasizing its pragmatic application as well as the strategies used to evaluate the presence of additive and multiplicative interactions. Practical issues discussed in this chapter include whether to adjust when interaction is suspected and the importance of the additive model in public health. A new flow chart is presented at the end of the chapter summarizing the main steps in the evaluation of interaction.

The next three chapters are devoted to the approaches used to handle threats to the validity of epidemiologic results. In Chapter 7, strategies for the adjustment of confounding factors are presented, including the more parsimonious approaches (e.g., direct adjustment, Mantel-Haenszel) as well as the more complex approaches (i.e., multiple regression, instrumental variables, Mendelian randomization, and propensity scores). Emphasis is placed on the selection of the method that is most appropriate for the study design used (e.g., Cox proportional hazards for the analysis of survival data and Poisson regression for the analysis of rates per person-time). Chapter 8 reviews the basic quality control strategies for the prevention and control of measurement error and bias. Both qualitative and quantitative approaches used in quality control are discussed. The most-often used analytic strategies for estimating validity and reliability of data obtained in epidemiologic studies are reviewed (e.g., unweighted and weighted kappa, correlation coefficients) in this chapter. In Chapter 9, the key issue of communication of results of epidemiologic studies is discussed. Examples of common mistakes made when reporting epidemiologic data are given as a way to stress the importance of clarity in such reports.

Chapter 10 discusses—from the epidemiologist's viewpoint—issues relevant to the interface between epidemiology, health policy, and public health, such as Rothman's causality model, proximal and distal causes, and Hill's guidelines. This chapter also includes brief discussions of three topics pertinent to causal inference—sensitivity analysis, meta-analysis, and publication bias—and consideration of the decision tree as a tool to evaluate interventions. A new section reviews the process of translational epidemiology.

As in the previous editions, Appendices A, B, C, and E describe selected statistical procedures (e.g., standard errors and confidence levels, trend test, test of heterogeneity of effects, intraclass correlation) to help the reader more thoroughly evaluate the measures of risk and association discussed in the text and to expose him or her to procedures that, although relatively simple, are not available in many statistical packages used by epidemiology students and practitioners. Appendix D includes two sections on quality assurance and control procedures taken from the corresponding manual of the Atherosclerosis Risk in Communities (ARIC) Study as examples of real-life applications of some of the procedures discussed in Chapter 8. Finally, Appendix F provides the answers to the exercises.

We encourage readers to advise us of any errors or unclear passages and to suggest improvements. Please email any such suggestions or comments to info@jblearning.com. All significant contributions will be acknowledged in the next edition.

Acknowledgments

This book is an outgrowth of an intermediate epidemiology course taught by the authors at the Johns Hopkins Bloomberg School of Public Health. Over the years, this course has benefited from significant intellectual input of many faculty members, including, among others, George W. Comstock, Helen Abbey, James Tonascia, Leon Gordis, and Mary Meyer. The authors especially acknowledge the late George W. Comstock, a mentor to both of us, who was involved with the course for several decades. His in-depth knowledge of epidemiologic methods and his wisdom over the years have been instrumental to our professional growth. Dr. Comstock also kindly provided many of the materials and examples used in Chapter 9 of this book. The original idea for developing a textbook on intermediate epidemiologic methods came from Michel Ibrahim, to whom we are very grateful.

We are indebted to many colleagues, including Leonelo Bautista, Daniel Brotman, Woody Chambless, Steve Cole, Josef Coresh, Rosa Maria Corona, Ana Diez-Roux, Jingzhong Ding, Manning Feinleib, Leon Gordis, Eliseo Guallar, Jay Kaufman, Kristen Malecki, Alfonso Mele, Paolo Pasquini, Paul Peppard, Patrick Remington, Jonathan Samet, Eyal Shahar, Richey Sharrett, and Michael Silverberg. These colleagues reviewed partial sections of this or previous editions or provided guidance in solving conceptual or statistical riddles. We are especially grateful to Blake Buchalter, Gabrielle Rude, Salwa Massad, Margarete (Grete) Wichmann, Hannah Yang, and Jennifer Deal for their careful review of some of the exercises and portions of the text. The authors are also grateful to Lauren Wisk for creating the ancillary instructor materials for this text. Finally, we would like to extend our appreciation to Patty Grubb, Michelle Mahana, and Jennifer Seltzer for their administrative help.

Having enjoyed the privilege of teaching intermediate epidemiology for so many years made us realize how much we have learned from our students, to whom we are deeply grateful. Finally, without the support and extraordinary patience of all members of our families, particularly our wives, Hilda and Marion, we could not have devoted so much time and effort to writing the four editions of this text.

About the Authors

Moyses Szklo, MD, DrPH, is University Distinguished Service Professor of Epidemiology and Medicine (Cardiology) at Johns Hopkins University. His current research focuses on risk factors for subclinical and clinical atherosclerosis. He is also editor in chief of *American Journal of Epidemiology*.

F. Javier Nieto, MD, PhD, is Dean and Professor of Epidemiology at the College of Public Health and Human Sciences at Oregon State University. His current research focuses on epidemiology of cardiovascular and sleep disorders, population-based survey methods, and global health.

PART 1

Introduction

CHAPTER 1

Basic Study Designs in Analytical Epidemiology

1.1 Introduction: Descriptive and Analytical Epidemiology

Epidemiology is traditionally defined as the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control health problems.[1] Epidemiology can be classified as either "descriptive" or "analytical." In general terms, *descriptive epidemiology* makes use of available data to examine how rates (e.g., mortality) vary according to demographic variables (e.g., those obtained from census data). When the distribution of rates is not uniform according to person, time, and place, the epidemiologist is able to define high-risk groups for prevention purposes (e.g., hypertension is more prevalent in U.S. blacks than in U.S. whites, thus defining blacks as a high-risk group). In addition, disparities in the distribution of rates serve to generate causal hypotheses based on the classic agent–host–environment paradigm (e.g., the hypothesis that environmental factors to which blacks are exposed, such as excessive salt intake or psychosocial stress, are responsible for their higher risk of hypertension).

A thorough review of descriptive epidemiologic approaches can be readily found in numerous sources.[2,3] For this reason and given the overall scope of this book, this chapter focuses on study designs that are relevant to *analytical epidemiology*, that is, designs that allow assessment of hypotheses of associations of suspected risk factor exposures with health outcomes. Moreover, the main focus of this textbook is *observational epidemiology*, even though many of the concepts discussed in subsequent chapters, such as measures of risk, measures of association, interaction/effect modification, and quality assurance/control, are also relevant to experimental studies (randomized clinical trials).

In this chapter, the two general strategies used for the assessment of associations in observational studies are discussed: (1) studies using populations or groups of individuals as units of observation—the so-called ecologic studies—and (2) studies using individuals as observation units, which include the prospective (or cohort), the case-control, the case-crossover, and the cross-sectional study designs.

Before that, however, the next section briefly discusses the *analysis of birth cohorts*. The reason for including this descriptive technique here is that it often requires the application of an analytical approach with a level of complexity usually not found in descriptive epidemiology; furthermore, this type of analysis is frequently important for understanding the patterns of association between age (a key determinant of health status) and disease in cross-sectional analyses. (An additional, more pragmatic reason for including a discussion of birth cohort analysis here is that it is usually not discussed in detail in basic textbooks.)

1.2 Analysis of Age, Birth Cohort, and Period Effects

Health surveys conducted in population samples usually include participants over a broad age range. Age is a strong risk factor for many health outcomes and is frequently associated with numerous exposures. Thus, even if the effect of age is not among the primary objectives of the study, given its potential confounding effects, it is often important to assess its relationship with exposures and outcomes.

TABLE 1-1 shows the results of a hypothetical cross-sectional study conducted in 2005 to assess the prevalence rates of a disease Y according to age. (A more strict use of the term *rate* as a measure of the occurrence of incident events is defined in Chapter 2, Section 2.2.2. This term is also widely used in a less precise sense to refer to proportions, such as prevalence.[1] It is in this more general sense that the term is used here and in other parts of the book.)

In **FIGURE 1-1**, these results are plotted at the midpoints of 10-year age groups (e.g., for ages 30–39, at 35 years; for ages 40–49, at 45 years; and so on). These data show that the prevalence of Y in this population decreases with age. Does this mean that the prevalence rates of Y decrease as individuals age? Not necessarily. For many disease processes, exposures have cumulative effects that are expressed over long periods of time. Long latency periods and cumulative effects characterize, for example, numerous exposure/disease associations, including smoking–lung cancer, radiation–thyroid cancer, and saturated fat intake–atherosclerotic disease. Thus, the health status of a person who is 50 years old at the time of the survey may be partially dependent on this person's past exposures (e.g., smoking during early adulthood). Variability of past exposures across successive generations

TABLE 1-1 Hypothetical data from a cross-sectional study of prevalence of disease Y in a population, by age, 2005.

Age group (years)	Midpoint (years)	2005 Prevalence (per 1000)
30–39	35	45
40–49	45	40
50–59	55	36
60–69	65	31
70–79	75	27

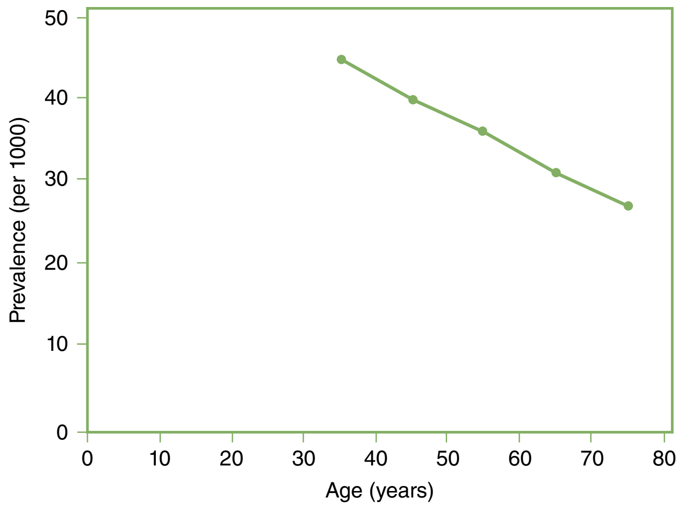

FIGURE 1-1 Hypothetical data from a cross-sectional study of prevalence of disease Y in a population, by age, 2005 (based on data from Table 1-1).

(birth cohorts*) can distort the apparent associations between age and health outcomes that are observed at any given point in time. This concept can be illustrated as follows.

Suppose that the same investigator who collected the data shown in Table 1-1 is able to recover data from previous surveys conducted in the same population in 1975, 1985, and 1995. The resulting data, presented in **TABLE 1-2** and **FIGURE 1-2**, show consistent trends of decreasing prevalence of Y with age in each of these surveys. Consider now plotting these data using a different approach, as shown in **FIGURE 1-3**. The dots in Figure 1-3 are at the same places as in Figure 1-2 except the lines are connected by *birth cohort* (the 2005 survey data are also plotted in Figure 1-3). Each of the dotted lines represents a birth cohort converging to the 2005 survey. For example, the "youngest" age point in the 2005 cross-sectional curve represents the rate of disease Y for individuals aged 30 to 39 years (average of 35 years) who were born between 1965 and 1974, that is, in 1970 on average (the "1970 birth cohort"). Individuals in this 1970 birth cohort were on average 10 years younger, that is, 25 years of age at the time of the 1995 survey and 15 years of age at the time of the 1985 survey. The line for the 1970 birth cohort thus represents how the prevalence of Y changes with increasing age for individuals born, on average, in 1970. Evidently, the cohort pattern shown in Figure 1-3 is very different from that suggested by the cross-sectional data and is consistent for all birth cohorts shown in Figure 1-3 in that it suggests that the prevalence of Y actually *increases* as people age. The fact that the inverse trend is observed in the cross-sectional data is due to a strong "cohort effect" in this example; that is, the prevalence of Y is strongly determined by the year of birth of the person. For any given age, the prevalence rate is higher in younger (more recent) than

***Birth cohort*: From Latin *cohors*, warriors, the 10th part of a legion. The component of the population born during a particular period and identified by period of birth so that its characteristics (e.g., causes of death and numbers still living) can be ascertained as it enters successive time and age periods.[1]

TABLE 1-2 Hypothetical data from a series of cross-sectional studies of prevalence of disease Y in a population, by age and survey date (calendar time), 1975–2005.

Age group (years)	Midpoint (years)	Survey date			
		1975	1985	1995	2005
		Prevalence (per 1000)			
10–19	15	17	28		
20–29	25	14	23	35	
30–39	35	12	19	30	45
40–49	45	10	18	26	40
50–59	55		15	22	36
60–69	65			20	31
70–79	75				27

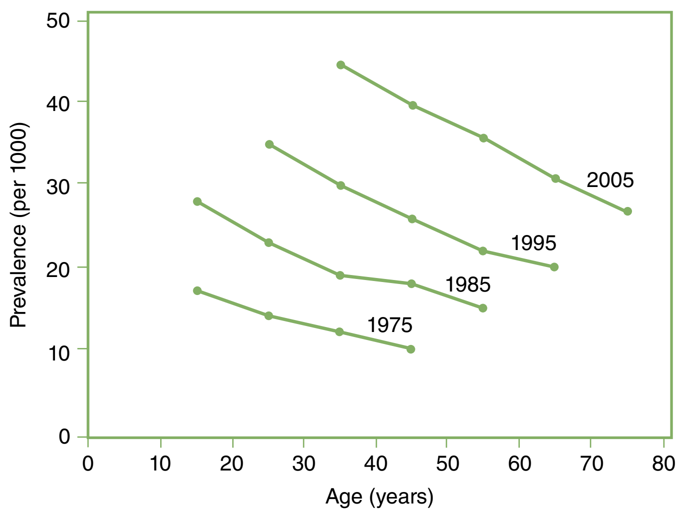

FIGURE 1-2 Hypothetical data from a series of cross-sectional studies of prevalence of disease Y (per 1000) in a population, by age and survey date (calendar time), 1975, 1985, 1995, and 2005 (based on data from Table 1-2).

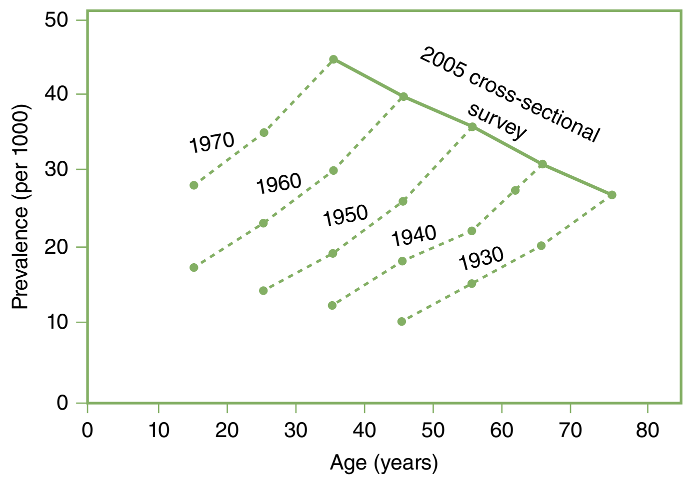

FIGURE 1-3 Plotting of the data in Figure 1-2 by birth cohort (see also Table 1-3). The dotted lines represent the different birth cohorts (from 1930 to 1970) as they converge to the 2005 cross-sectional survey (solid line, as in Figure 1-1).

in older cohorts. Thus, in the 2005 cross-sectional survey (Figure 1-1), the older subjects come from birth cohorts with relatively lower rates, whereas the youngest come from the cohorts with higher rates. This can be seen clearly in Figure 1-3 by selecting one age (e.g., 45 years) and observing that the rate is lowest for the 1930 birth cohort and increases for each subsequent birth cohort (i.e., the 1940, 1950, and 1960 cohorts, respectively).

Although the cross-sectional analysis of prevalence rates in this example gives a distorted view of the disease behavior as a birth cohort ages, it is still useful for planning purposes because, regardless of the mix of birth cohorts, cross-sectional data inform the public health authorities about the burden of disease as it exists currently (e.g., the age distribution of disease Y prevalence in 2005).

An alternative display of the data from Table 1-2 is shown in **FIGURE 1-4**. Instead of age (as in Figures 1-1 to 1-3), the scale in the abscissa (*x*-axis) corresponds to the birth cohort and each line to an age group; thus, the slope of the lines represents the change across birth cohorts for a given age group.

Often the choice among these alternative graphical representations is a matter of personal preference (i.e., which pattern the investigator wishes to emphasize). Whereas Figure 1-4 shows trends according to birth cohorts more explicitly (e.g., for any given age group, there is an increasing prevalence from older to more recent cohorts), Figure 1-3 has an intuitive appeal in that each line represents a birth cohort as it ages. As long as one pays careful attention to the labeling of the graph, any of these displays is appropriate for identifying age and birth cohort patterns. The same patterns displayed in Figures 1-3 and 1-4 can be seen in Table 1-2, moving downward to examine cross-sectional trends and diagonally from left to right to examine birth cohort trends. As an example, for the cohort born between 1955 and 1964 (midpoint in 1960), the prevalence rates per 1000 are 17, 23, 30, and 40 for ages (midpoint) 15, 25, 35, and 45 years, respectively. An alternative and somewhat more readable display of the same data for the purpose of detecting trends according to birth cohort is shown in **TABLE 1-3**, which allows the examination of trends according to age ("age effect") within each birth cohort (horizontal lines in Table 1-3). Additionally, and in agreement with Figure 1-4, Table 1-3 shows

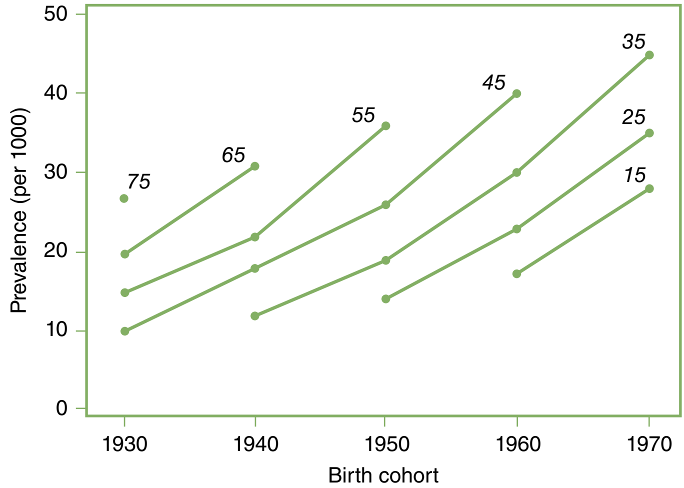

FIGURE 1-4 Alternative display of the data in Figure 1-3. Birth cohorts are represented in the *x*-axis. The lines represent age groups (labeled using italicized numbers representing the midpoints, in years).

TABLE 1-3 Rearrangement of the data shown in Table 1-2 by birth cohort.

		Age group (midpoint, in years)						
		15	25	35	45	55	65	75
Birth cohort range	Midpoint	Prevalence (per 1000)						
1925–1934	1930				10	15	20	27
1935–1944	1940			12	18	22	31	
1945–1954	1950		14	19	26	36		
1955–1964	1960	17	23	30	40			
1965–1974	1970	28	35	45				

how prevalence rates increase from older to more recent cohorts (cohort effect)—readily visualized by moving one's eyes from the top to the bottom of each age group column in Table 1-3.

Thus, the data in the previous example are simultaneously affected by two strong effects: "cohort effect" and "age effect" (for definitions, see **EXHIBIT 1-1**). These two trends are jointly responsible for the seemingly paradoxical trend observed in the cross-sectional analyses in this hypothetical example

(Figures 1-1 and 1-2) in which the rates seem to *decrease* with age. The fact that more recent cohorts have substantially higher rates (cohort effect) overwhelms the increase in prevalence associated with age and explains the observed cross-sectional pattern. In other words, in cross-sectional data, the rates in the older ages are those from the earlier cohorts, whose rates were lower than those of the more recently born cohorts.

In addition to cohort and age effects, patterns of rates can be influenced by the so-called period effect. The term *period effect* is frequently used to refer to a global shift or change in trends that affects the rates across all birth cohorts and age groups (Exhibit 1-1). Any phenomenon occurring at a specific point in time (or during a specific period) that affects an entire population (or a significant segment of it), such as a war, a new treatment, or massive migration, can produce this change independently of age and birth cohort effects. A hypothetical example is shown in **FIGURE 1-5**. This figure shows data similar to those used in the previous example (Figure 1-3) except, in this case, the rates level off in 1995 for all cohorts (i.e., when the 1970 cohort is 25 years old on average, when the 1960 cohort is 35 years old, and so on).

EXHIBIT 1-1 Definitions of age, cohort, and period effects.

Age effect	Change in the rate of a condition according to age regardless of birth cohort and calendar time
Cohort effect	Change in the rate of a condition according to year of birth regardless of age and calendar time
Period effect	Change in the rate of a condition affecting an entire population at some point in time regardless of age and birth cohort

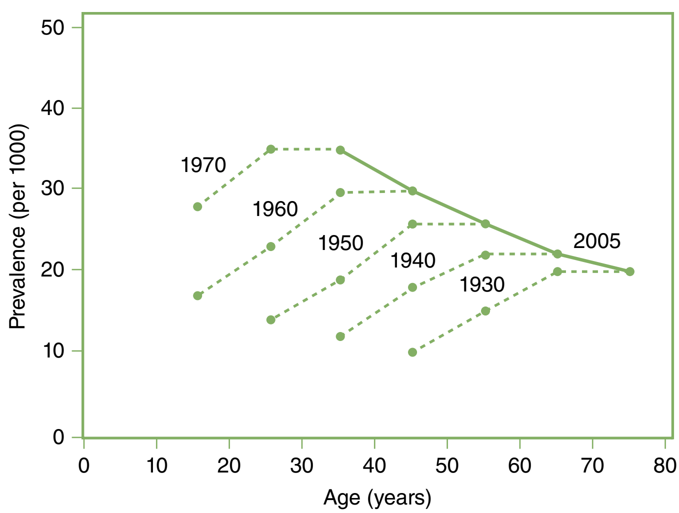

FIGURE 1-5 Hypothetical example of period effect. An event happened in 1995 that affected all birth cohorts (1930–1970) in a similar way and slowed down the rate of increase with age. The solid line represents the observed cross-sectional age pattern in 2005.