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

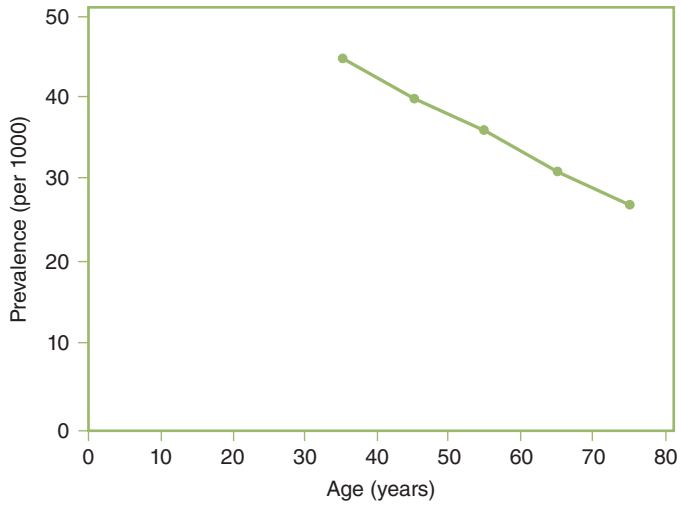


FIGURE 1-1 Hypothetical data from a cross-sectional study of prevalence of disease Y in a population, by age, 2005 (based on data from Table 1-1).

(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

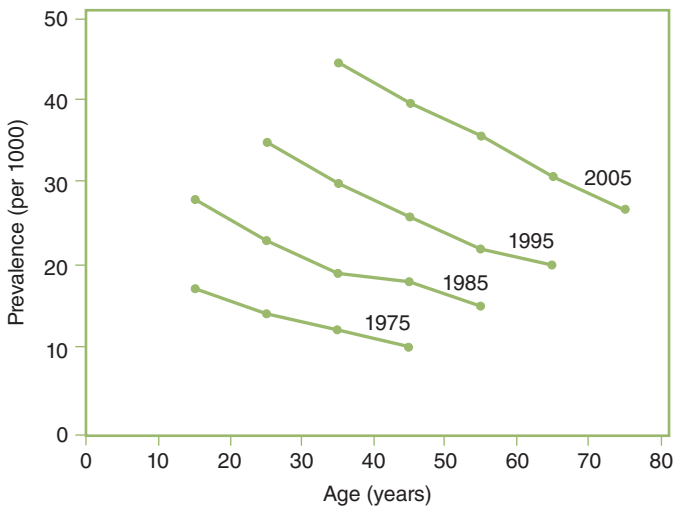


FIGURE 1-2 Hypothetical data from a series of cross-sectional studies of prevalence of disease Y (per 1000) in a population, by age and survey date (calendar time), 1975, 1985, 1995, and 2005 (based on data from Table 1-2).

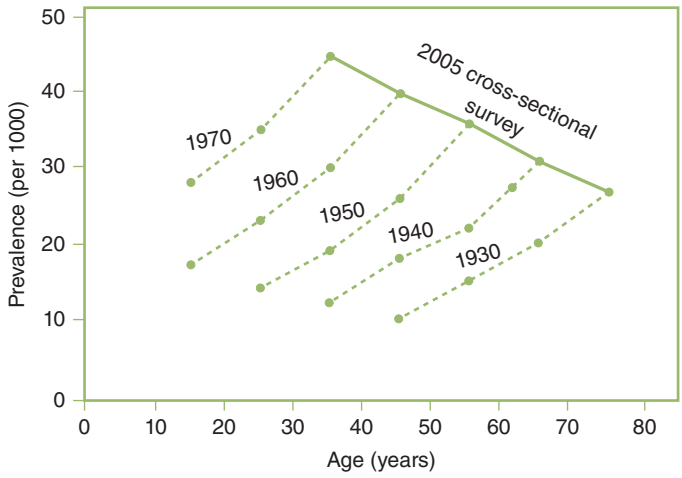


FIGURE 1-3 Plotting of the data in Figure 1-2 by birth cohort (see also Table 1-3). The dotted lines represent the different birth cohorts (from 1930 to 1970) as they converge to the 2005 cross-sectional survey (solid line, as in Figure 1-1).

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

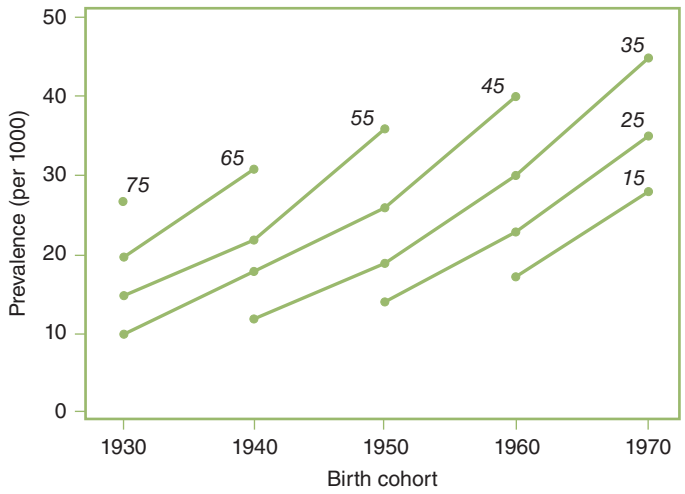


FIGURE 1-4 Alternative display of the data in Figure 1-3. Birth cohorts are represented in the *x*-axis. The lines represent age groups (labeled using italicized numbers representing the midpoints, in years).

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

Birth cohort range	Midpoint	Age group (midpoint, in years)						
		15	25	35	45	55	65	75
		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

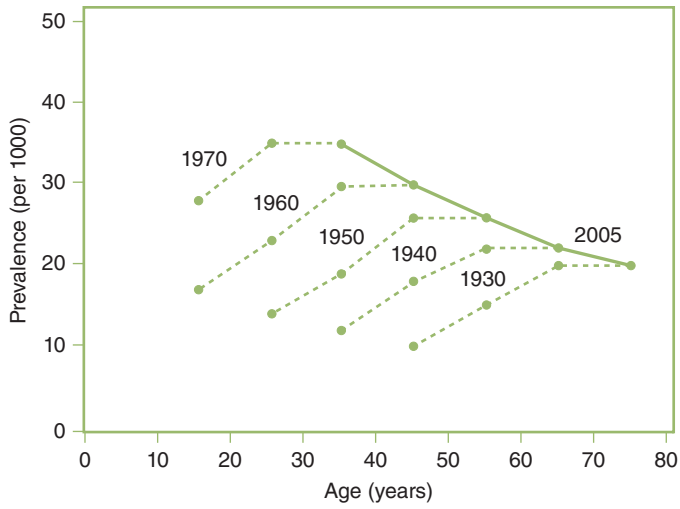


FIGURE 1-5 Hypothetical example of period effect. An event happened in 1995 that affected all birth cohorts (1930–1970) in a similar way and slowed down the rate of increase with age. The solid line represents the observed cross-sectional age pattern in 2005.

Period effects on prevalence rates can occur, for example, when new medications or preventive interventions are introduced for diseases that previously had poor prognoses, as in the case of the introduction of insulin, antibiotics, and the polio vaccine.

It is important to understand that the so-called birth cohort effects may have little to do with the circumstances surrounding the time of birth of a given cohort of individuals. Rather, cohort effects may result from the lifetime experience (including, but not limited to, those surrounding birth) of the individuals born at a given point in time that influences the disease or outcome of interest. For example, currently observed patterns of association between age and coronary heart disease (CHD) may have resulted from cohort effects related to changes in diet (e.g., fat intake) or smoking habits of adolescents and young adults over time. It is well known that coronary atherosclerotic markers, such as thickening of the arterial intima, frequently develop early in life.[4] In middle and older ages, some of these early intimal changes may evolve into raised atherosclerotic lesions, eventually leading to thrombosis, lumen occlusion, and the resulting clinically manifest acute ischemic events. Thus, a young adult's dietary and/or smoking habits may influence atherosclerosis development and subsequent coronary risk. If changes in these habits occur in the population over time, successive birth cohorts will be subjected to changing degrees of exposure to early atherogenic factors, which will in part determine future cross-sectional patterns of the association of age with CHD.

Another way to understand the concept of cohort effects is that they are the result of an *interaction* between age and calendar time. The concept of interaction is discussed in detail in Chapter 6. In simple terms, it means that a given variable (e.g., calendar time in the case of a cohort effect) *modifies* the strength or the nature of an association between another variable (e.g., age) and an outcome (e.g., coronary atherosclerosis). In the previous example, it means that the way age relates to the development of atherosclerosis changes over time as a result of changes in the population prevalence of key risk factors (e.g., dietary/smoking habits of young adults). In other words, calendar time–related changes in risk factors *modify* the association between age and atherosclerosis.

Cohort–age–period analyses can be applied not only to prevalence data but also to incidence and mortality data. A classic example is Wade Hampton Frost's study of age patterns of tuberculosis mortality.[5] **FIGURE 1-6** presents two graphs from Frost's landmark paper. With regard to Figure 1-6A, Frost[5(p94)] noted that "looking at the 1930 curve, the impression given is that nowadays an individual

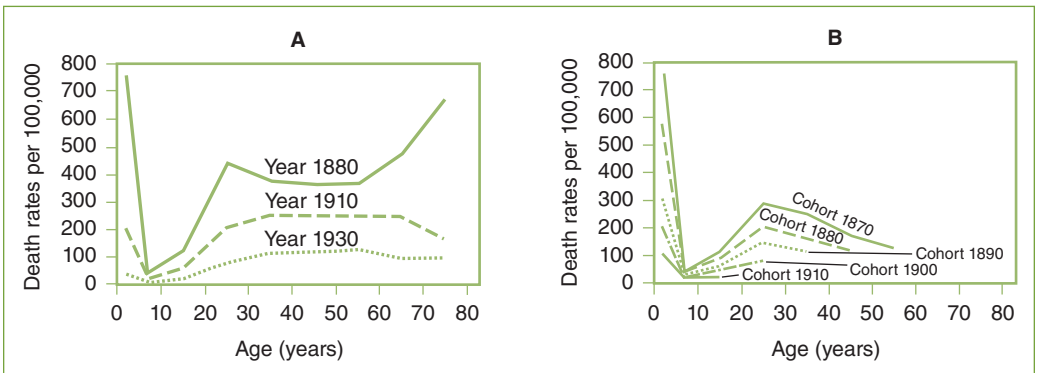


FIGURE 1-6 Frost's analysis of age in relation to tuberculosis mortality (males only). (A) Massachusetts death rates from tuberculosis, by age, 1880, 1910, 1930. (B) Massachusetts death rates from tuberculosis, by age, in successive 10-year cohorts.

Reproduced from Frost WH. The age-selection of tuberculosis mortality in successive decades. *Am J Hyg.* 1939;30:91-96.[5] By permission of Oxford University Press.

encounters his greatest risk of death from tuberculosis between the ages of 50 and 60. But this is not really so; the people making up the 1930 age group 30 to 60 have, in earlier life, passed through *greater* mortality risk" (emphasis in original). This is demonstrated in Figure 1-6B, aptly used by Frost to show how the risk of tuberculosis death after the first few years of life is actually highest at ages 20 to 30 years for cohorts born in 1870 through 1890.

Another example is shown in **FIGURE 1-7**, based on an analysis of age, cohort, and period effects on the incidence of colorectal cancer in a region of Spain.[6] In these figures, birth cohorts are placed on the *x*-axis (as in Figure 1-4). These figures show strong cohort effects: For each age group, the incidence rates of colorectal cancer tend to increase from older to more recent birth cohorts. An age effect is also evident, as for each birth cohort (for any given year-of-birth value in the horizontal axis) the rates are higher for older than for younger individuals. Note that a logarithmic scale was used in the ordinate in this figure in part because of the wide range of rates needed to be plotted. (For further discussion of the use of logarithmic vs arithmetic scales, see Chapter 9, Section 9.3.5.)

An additional example of age and birth cohort analysis of incidence data is shown in **FIGURE 1-8**. This figure shows the incidence of ovarian cancer in Mumbai, India, by age and year of birth

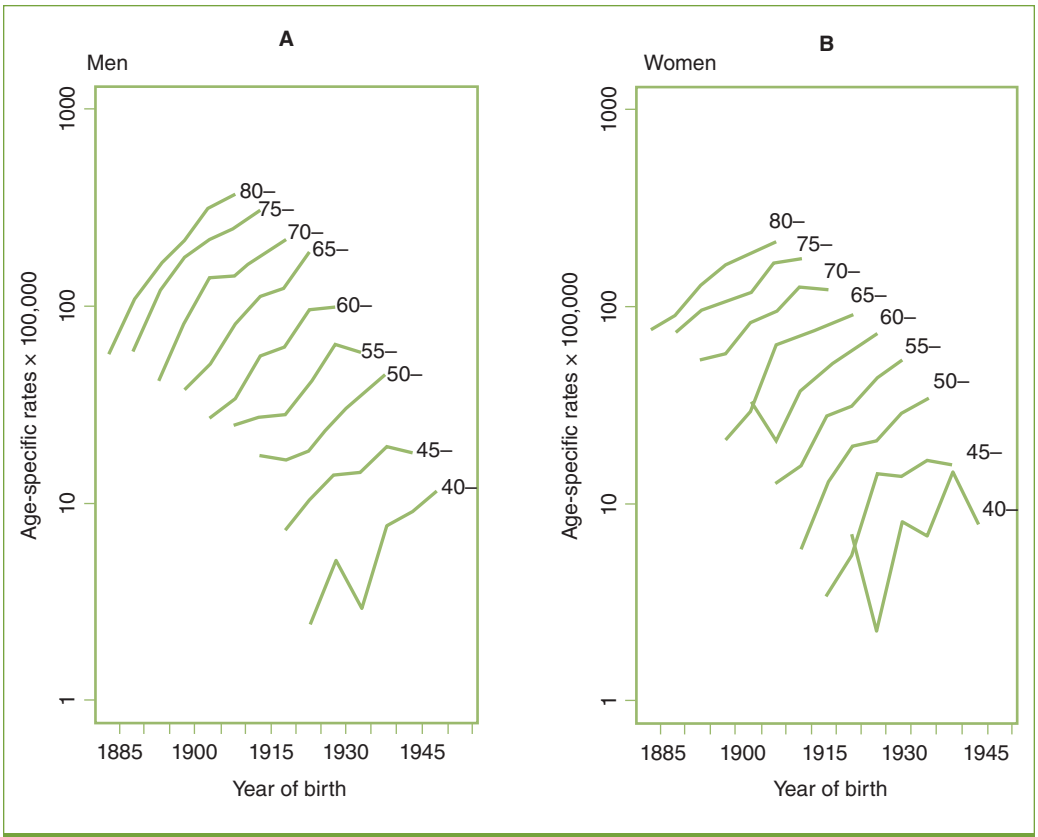


FIGURE 1-7 Trends in age-specific incidence rates of colorectal cancer in Navarra and Zaragoza (Spain). The number next to each line represents the initial year of the corresponding 5-year age group.

Reproduced from López-Abente G, Pollán M, Vergara A, et al. Age-period-cohort modeling of colorectal cancer incidence and mortality in Spain. *Cancer Epidemiol Biomarkers Prev*. 1997;6:999-1005.[6] With permission from AACR.

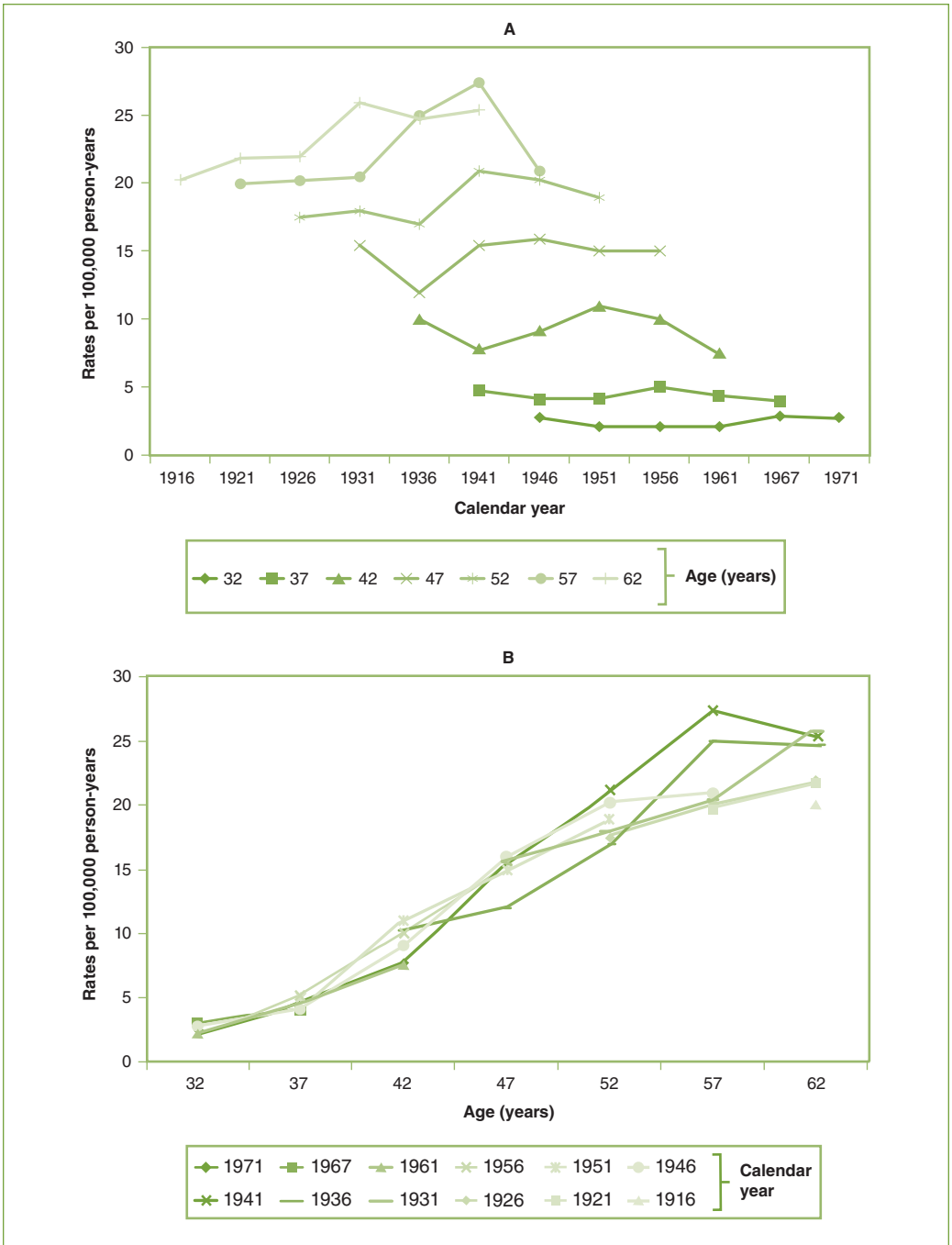


FIGURE 1-8 Incidence rates of ovarian cancer per 100,000 person-years, by birth cohort (A) and by age (B), Mumbai, India, 1976–2005.

Modified from Dhillon PK, Yeole BB, Dikshit R, Kurkure AP, Bray F. Trends in breast, ovarian and cervical cancer incidence in Mumbai, India over a 30-year period, 1976-2005: an age-period-cohort analysis. *Br J Cancer.* 2011;105:723-730.[7]

cohort.[7] This is an example in which there is a strong age effect, particularly for the cohorts born from 1940 through 1970; that is, rates increase dramatically with age through age 52 years but with very little cohort effect, as indicated by the approximate flat pattern for the successive birth cohorts for each age group (the figure shows the midpoint of each age group). It should be manifest that, when there is little cohort effect, as in this situation, the cross-sectional curves and cohort curves will essentially show the same pattern, with the cross-sectional curves practically overlapping each other (Figure 1-8B).

Period effects associated with incidence rates tend to be more prominent for diseases for which the cumulative effects of previous exposures are relatively unimportant, such as infectious diseases and injuries. Conversely, in chronic diseases, such as cancer and cardiovascular disease, cumulative effects are usually important, and thus, cohort effects tend to affect incidence rates to a greater extent than period effects.

These methods can also be used to study variables other than disease rates. An example is the analysis of age-related changes in serum cholesterol levels shown in **FIGURE 1-9**, based on data from the Florida Geriatric Research Program.[8] This figure reveals a slight cohort effect in that serum cholesterol levels tend to be lower in older than in more recent birth cohorts for most age groups. A J- or U-shaped age pattern is also seen; that is, for each birth cohort, serum cholesterol tends to first decrease or remain stable with increasing age and then increase to achieve its maximum value in the oldest members of the cohort. Although at first glance this pattern might be considered an "age effect," for each cohort the maximum cholesterol values in the oldest age group coincide with a single point in calendar time: 1985 through 1987 (i.e., for the 1909–1911 birth cohort at 76 years of age, for the 1906–1908 cohort at 79 years of age, and so on), leading Newschaffer et al. to observe that "a period effect is suggested by a consistent change in curve height at a given time point over all cohorts. . . . Therefore, based on simple visual inspection of the curves, it is not possible to attribute the consistent U-shaped increase in cholesterol to aging, since some of this shape may be accounted for by period effects."[8(p26)]

In complex situations, it may be difficult to differentiate age, cohort, and period effects. In a complex situation, such as that illustrated in the preceding discussion, multiple regression techniques can be used to disentangle these effects. Describing these techniques in detail is beyond the scope of this book. (A general discussion of multiple regression methods is presented in Chapter 7, Section 7.4.) The interested reader can find examples and further references in the original papers from the previously cited examples (e.g., López-Abente et al.[6] and Newschaffer et al[8]).

Finally, it should be emphasized that birth cohort effects may affect associations between disease outcomes and variables other than age. Consider, for example, a case-control study (see Section 1.4.2) in which cases and controls are closely matched by age (see Section 1.4.5). Assume that this is a study of a rare disease in which cases are identified over a 10-year span (e.g., from 2000 through 2009) and controls at the end of the accrual of cases (such as may happen when frequency matching is used—see section 1.4.5). In this study, age per se does not act as a confounder, as cases and controls are matched on age (see Chapter 5, Section 5.2.2); however, the fact that cases and controls are identified from different birth cohorts may affect the assessment of variables, such as educational level, that may have changed rapidly across birth cohorts. In this case, birth cohort, but not age, would confound the association between education and the disease of interest.

---

[*] A mean value can be calculated for both continuous and discrete (e.g., binary) variables. A proportion is a mean of individual binary values (e.g., 1 for presence of a certain characteristic, 0 if the characteristic is absent).

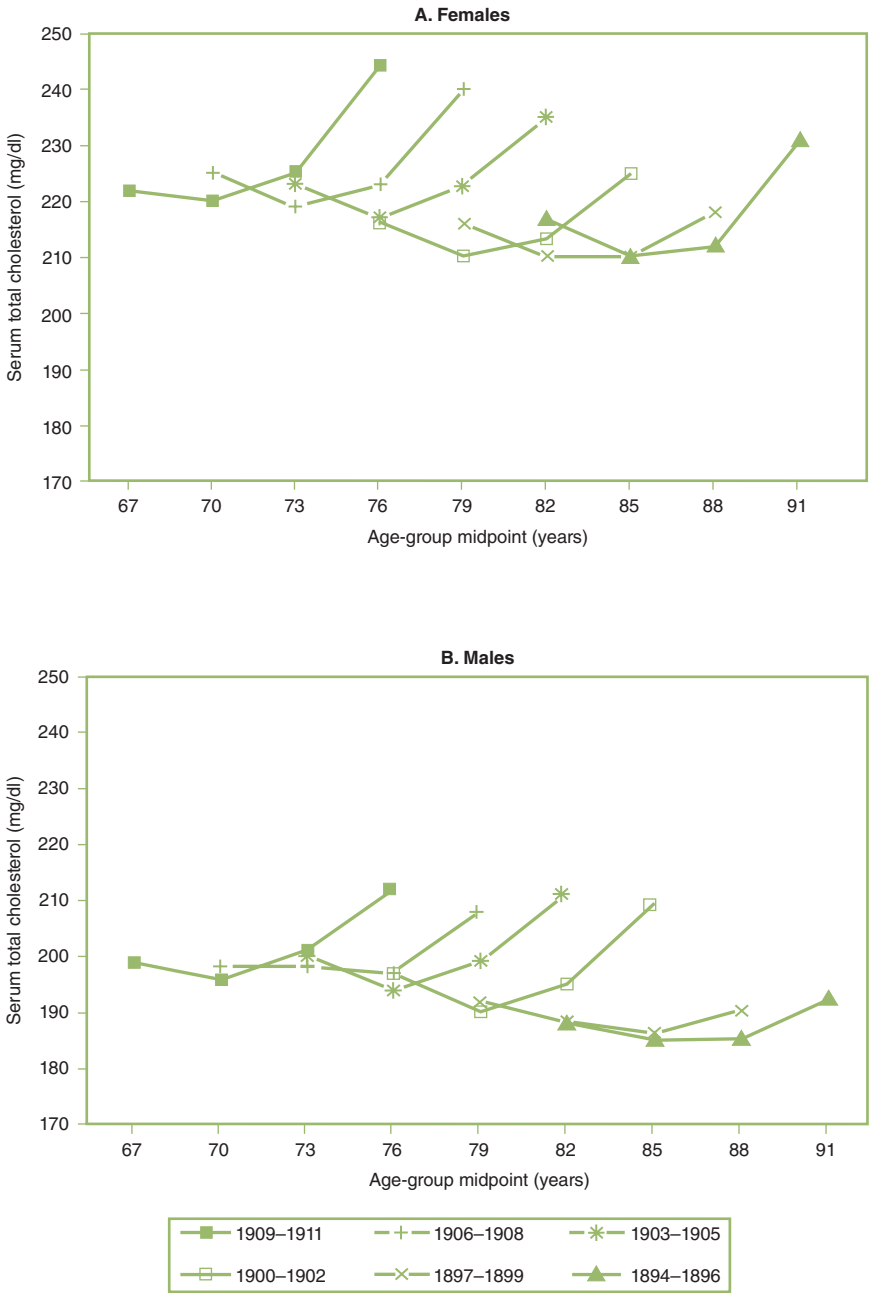


FIGURE 1-9 Sex-specific mean serum cholesterol levels by age and birth cohort. Longitudinal data from the Florida Geriatric Research Program, Dunedin County, Florida, 1976–1987.

Reprinted with permission from Newschaffer CJ, Bush TL, Hale WE. Aging and total cholesterol levels: cohort, period, and survivorship effects. *Am J Epidemiol*. 1992;136:23-34.[8] By permission of Oxford University Press.

1.3 Ecologic Studies

The units of observation in an ecologic study are usually geographically defined populations (such as countries or regions within a country) or the same geographically defined population at different points in time. Mean values* for both a given postulated risk factor and the outcome of interest are obtained for each observation unit for comparison purposes. Typically, the analysis of ecologic data involves plotting the risk factor and outcome values for all observation units to assess whether a relationship is evident. For example, **FIGURE 1-10** displays the death rates for CHD in men from 16 cohorts included in the Seven Countries Study plotted against the corresponding estimates of mean fat intake (percent calories from fat).[9] A positive relationship between these two variables is suggested by these data, as there is a tendency for the death rates to be higher in countries having higher average saturated fat intakes.

Different types of variables can be used in ecologic studies,[10] which are briefly summarized as follows:

- *Aggregate measures* that summarize the characteristics of individuals within a group as the mean value of a certain parameter or the proportion of the population or group of interest with

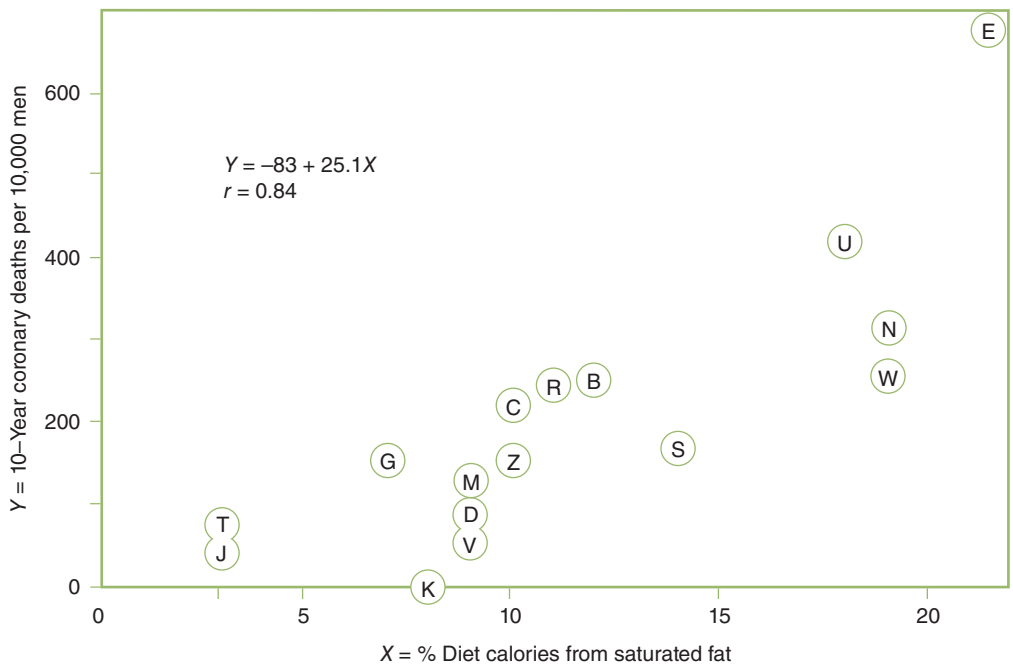


FIGURE 1-10 Example of an ecologic study. Ten-year coronary death rates of the cohorts from the Seven Countries Study plotted against the percentage of dietary calories from saturated fatty acids. Cohorts: B, Belgrade; C, Crevalcore; D, Dalmatia; E, East Finland; G, Corfu; J, Ushibuka; K, Crete; M, Montegiorgio; N, Zuphen; R, Rome railroad; S, Slavonia; T, Tanushimaru; U, American railroad; V, Velika Krsna; W, West Finland; Z, Zrenjanin. Shown in the figure are the linear regression coefficients (see Chapter 7, Section 7.4.1) and the correlation coefficient *r* corresponding to this plot.

Reprinted with permission from Keys A. *Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease.* Cambridge, MA: Harvard University Press; 1980.[9] By the President and Fellows of Harvard College.

a certain characteristic. Examples include the prevalence of a given disease, average amount of fat intake (Figure 1-10), proportion of smokers, and median income.

- *Environmental measures* that represent physical characteristics of the geographic location for the group of interest. Individuals within the group may have different degrees of exposure to a given characteristic, which could theoretically be measured. Examples include air pollution intensity and hours of sunlight.
- *Global measures* that represent characteristics of the group that are not reducible to characteristics of individuals (i.e., that are not analogues at the individual level). Examples include the type of political or healthcare system in a given region, a certain regulation or law, and the presence and magnitude of health inequalities.

In a traditional ecologic study, two ecologic variables are contrasted to examine their possible association. Typically, an ecologic measure of exposure and an aggregate measure of disease or mortality are compared (Figure 1-10). These ecologic measures can also be used in studies of individuals (see Section 1.4) in which the investigator chooses to define exposure using an ecologic criterion on the basis of its expected superior construct validity.* For example, in a cross-sectional study of the relationship between socioeconomic status and prevalent cardiovascular disease, the investigator may choose to define study participants' socioeconomic status using an aggregate indicator (e.g., median family income in the neighborhood) rather than, for example, his or her own (individual) educational level or income. Furthermore, both individual and aggregate measures can be simultaneously considered in *multilevel analyses*, as when examining the joint role of individuals' and aggregate levels of income and education in relation to prevalent cardiovascular disease.[11]

An ecologic association may accurately reflect a causal connection between a suspected risk factor and a disease (e.g., the positive association between fat intake and CHD depicted in Figure 1-10). However, the phenomenon of *ecologic fallacy* is often invoked as an important limitation for the use of ecologic correlations as *bona fide* tests of etiologic hypotheses. The ecologic fallacy (or *aggregation bias*) has been defined as the bias that may occur because an association observed between variables on an aggregate level does not necessarily represent the association that exists at an individual level.[1(p88)] The phenomenon of ecologic fallacy is schematically illustrated in **FIGURE 1-11**, based on an example proposed by Diez-Roux.[12] In a hypothetical ecologic study examining the relationship between per capita income and the risk of motor vehicle injuries in three populations composed of seven individuals each, a positive correlation between mean income and risk of injuries is observed; however, a close inspection of *individual* values reveals that cases occur exclusively in persons with low income (less than U.S. $20,000). In this extreme example of ecologic fallacy, the association detected when using populations as observation units (e.g., higher mean income relates to a higher risk of motor vehicle injuries) has a direction diametrically opposed to the relationship between income and motor vehicle injuries in individuals—in whom higher individual income relates to a lower injury risk. Thus, the conclusion from the ecologic analysis that a higher income is a risk factor for motor vehicle injuries may be fallacious (discussed later in this section).

Another example of a situation in which this type of fallacy may have occurred is given by an ecologic study that showed a direct correlation between the percentage of the population that was Protestant and suicide rates in a number of Prussian communities in the late 19th century.[10,13] Concluding from this observation that being Protestant is a risk factor for suicide may well be

*Construct validity is the extent to which an operational variable (e.g., body weight) accurately represents the phenomenon it purports to represent (e.g., nutritional status).

FIGURE 1-11 Schematic representation of a hypothetical study in which ecologic fallacy occurs. Boxes represent hypothetical individuals, darker boxes represent incident cases of motor vehicle (MV) injuries, and the numbers inside the boxes indicate individuals' annual incomes (in thousands of U.S. dollars). Ecologically, the correlation is positive: Population A has the highest values of both mean income and incidence of MV injuries, population B has intermediate values of both mean income and incidence of MV injuries, and population C has the lowest values of both mean income and incidence of MV injuries. In individuals, however, the relationship is negative: For all three populations combined, mean income is U.S. $13,456 for cases and U.S. $29,617 for noncases.

wrong (i.e., may result from an ecologic fallacy). For example, it is possible that most of the suicides within these communities were committed by Catholic individuals who, when in the minority (i.e., in communities predominantly Protestant), tended to be more socially isolated and therefore at a higher risk of suicide.

As illustrated in these examples, errors associated with ecologic studies are the result of *cross-level inference*, which occurs when aggregate data are used to make inferences at the individual level.[10] The mistake in the example just discussed is to use the correlation between the proportion of Protestants (which is an aggregate measure) and suicide rate to infer that the risk of suicide is higher in Protestant than in Catholic *individuals*. If one were to make an inference at the *population* level, however, the conclusion that predominantly Protestant communities with Catholic minorities have higher rates of suicide would still be valid (provided that other biases and confounding factors were not present). Similarly, in the income/injuries example, the inference from the ecologic analysis is wrong only if intended for the understanding of determinants at the level of the individual. The ecologic information may be valuable if the investigator's purpose is to understand fully the complex web of causality[14] involved in motor vehicle injuries, as it may yield clues regarding the causes of motor vehicle injuries that are not provided by individual-level data. In the previous example (Figure 1-11), higher mean population income may truly be associated with increased traffic volume and, consequently, with higher rates of motor vehicle injuries. At the individual level, however, the inverse association between

income and motor vehicle injuries may result from the higher frequency of use of unsafe vehicles among low-income individuals, particularly in the context of high traffic volume.

Because of the prevalent view that inference at the individual level is the gold standard when studying disease causation,[15] as well as the possibility of ecologic fallacy, ecologic studies are often considered imperfect surrogates for studies in which individuals are the observation units. Essentially, ecologic studies are seen as preliminary studies that "can suggest avenues of research that may be promising in casting light on etiological relationships."[3(p210)] That this is often but not always true has been underscored by the examples discussed previously here. Furthermore, the following three situations demonstrate that an ecologic analysis may on occasion lead to *more* accurate conclusions than an analysis using individual-level data—even if the level of inference in the ecologic study is at the individual level.

1. The first situation is when the within-population variability of the exposure of interest is low but the between-population variability is high. For example, if salt intake of individuals in a given population were above the threshold needed to cause hypertension, a relationship between salt and hypertension might not be apparent in an observational study of individuals in this population, but it could be seen in an ecologic study including populations with diverse dietary habits.[16] (A similar phenomenon has been postulated to explain why ecologic correlations, but not studies of individuals, have detected a relationship between fat intake and risk of breast cancer.[17])
2. The second situation is when, even if the intended level of inference is the individual, the *implications for prevention or intervention are at the population level*. Some examples of the latter situation are as follows:
   - In the classic studies on pellagra, Goldberger et al.[18] assessed not only individual indicators of income but also food availability in the area markets. They found that, independently of individual socioeconomic indicators, food availability in local markets in the villages was strongly related to the occurrence of pellagra, leading these authors to conclude the following:

     > *The most potent factors influencing pellagra incidence in the villages studied were (a) low family income, and (b) unfavorable conditions regarding the availability of food supplies, suggesting that under conditions obtaining [sic] in some of these villages in the spring of 1916 many families were without sufficient income to enable them to procure an adequate diet, and that improvement in food availability (particularly of milk and fresh meat) is urgently needed in such localities.*[18(p2712)]

   It should be readily apparent in this example that an important (and potentially modifiable) link in the causal chain of pellagra occurrence—namely, food availability—may have been missed if the investigators had focused exclusively on individual income measures.
   - Studies of risk factors for smoking initiation and/or smoking cessation may focus on community-level cigarette taxes or regulation of cigarette advertising. Although individual factors may influence the individual's predisposition to smoking (e.g., psychological profile, smoking habits of relatives or friends), regulatory "ecologic" factors may be strong determinants and modifiers of the individual behaviors. Thus, an investigator may choose to focus on these global factors rather than on (or in addition to) individual behaviors.
   - When studying the determinants of transmission of certain infectious diseases with complex nonlinear infection dynamics (e.g., attenuated exposure–infection relationship at the individual level), ecologic studies may be more appropriate than studies using individuals as observation units.[19]

3. The third situation is when testing a hypothesis exclusively at the population level. An example is given by a study of Lobato et al.[20] aimed at testing Rose and Day's theory that, as the distribution of a particular health-related characteristic in a population shifts, if its dispersion is unchanged, the mean and prevalence of extreme values are expected to be correlated.[21] In agreement with this concept, Lobato et al. found a strong correlation between mean body mass index (weight in kilograms/square of height in meters) and prevalence of obesity in 26 Brazilian capitals in adult women ($r = 0.86$).[20]

Because ultimately all risk factors must operate at the individual level, the quintessential reductionistic* approach would focus on only the causal pathways at the biochemical or intracellular level. For example, the study of the carcinogenic effects of tobacco smoking could focus on the effects of tobacco by-products at the cellular level, that is, alteration of the cell's DNA. However, will that make the study of smoking habits irrelevant? Obviously not. Indeed, from a public health perspective, the use of a comprehensive theoretical model of causality—one that considers all factors influencing the occurrence of disease—often requires taking into account the role of upstream and ecologic factors (including environmental, sociopolitical, and cultural) in the causal chain (see also Chapter 10, Sections 10.2.2 and 10.2.3). As stated at the beginning of this chapter, the ultimate goal of epidemiology is to be effectively used as a tool to improve the health conditions of the public; in this context, the factors that operate at a global level may represent important links in the chain of causation, particularly when they are amenable to intervention (e.g., improving access to fresh foods in villages, establishing laws that limit cigarette advertising, or improving the built environment in cities to eliminate barriers to individuals' physical activity habits). As a result, studies focusing on factors at the individual level may be insufficient in that they fail to address these ecologic links in the causal chain. This important concept can be illustrated using the previously discussed example of religion and suicide. A study based on individuals would "correctly" find that the risk of suicide is higher in Catholics than in Protestants.[10] This finding would logically suggest explanations for why the suicide rate differs between these religious groups. For example, is the higher rate in Catholics caused by Catholicism per se? Alternatively, is it because of some ethnic difference between Catholics and Protestants? If so, is it due to some genetic component that distinguishes these ethnic groups? The problem is that these questions, which attempt to characterize risk at the individual level, although important, are insufficient to explain fully the "web of causality,"[14] for they fail to consider the ecologic dimension of whether minority status explains and determines the increased risk of suicide. This example underscores the concept that both individual and ecologic studies are often necessary to study the complex causal determination not only of suicide but also of many other health and disease processes.[12] The combination of individual and ecologic levels of analysis poses analytical challenges for which statistical models (hierarchical models) have been developed. Difficult conceptual challenges remain, however, such as the development of causal models that include all relevant risk factors operating from the social to the biological level and that consider their possible multilevel interaction.[12]

1.4 Studies Based on Individuals as Observation Units

There are three basic types of nonexperimental (observational) study designs in which individuals are the units of observation: the cohort or prospective study, the case-control study, and the cross-sectional study. In this section, key aspects of these study designs are reviewed. The case-crossover study, a special type of case-control study, is also briefly discussed. For a more comprehensive discussion

*Reductionism is a theory that postulates that all complex systems can be completely understood in terms of their components, basically ignoring interactions between these components.

of the operational and analytical issues related to observational epidemiologic studies, the reader is referred to specialized texts.[22-26]

From a conceptual standpoint, the fundamental study design in observational epidemiology—that is, the design from which the others derive and that can be considered the gold standard—is the cohort or prospective study. Cohort data, if unbiased, reflect the "real-life" cause–effect temporal sequence of events, a *sine qua non* criterion to establish causality (see Chapter 10, Section 10.2.4). From this point of view, the case-control and the cross-sectional designs are mere variations of the cohort study design and are primarily justified by feasibility, logistical ease, and efficiency.

1.4.1 Cohort Study

In a cohort study, a group of healthy people, or a *cohort*,* is identified and followed for a certain time period to ascertain the occurrence of health-related events (**FIGURE 1-12**). The usual objective of a cohort study is to investigate whether the incidence of an event is related to a suspected exposure.

Study populations in cohort studies can be quite diverse and may include a sample of the general population of a certain geographic area (e.g., the Framingham Study[27]); an occupational cohort, typically defined as a group of workers in a given occupation or industry who are classified according to exposure to agents thought to be occupational hazards; or a group of people who, because of certain characteristics, are at an unusually high risk for a given disease (e.g., the cohort of homosexual men who are followed in the Multicenter AIDS Cohort Study[28]). Alternatively, cohorts can be formed by "convenience" samples, or groups gathered because of their willingness to participate or because of other logistical advantages, such as ease of follow-up; examples include the Nurses Health Study cohort,[29] the Health Professionals Study cohort,[30] and the American Cancer Society cohort studies of volunteers.[31]

After the cohort is defined and the participants are selected, a critical element in a cohort study is the ascertainment of events during the follow-up time (when the event of interest is a newly developed disease, prevalent cases are excluded from the cohort at baseline). This is the reason these studies are also known as *prospective studies*.[30] A schematic depiction of a cohort of 1000 individuals

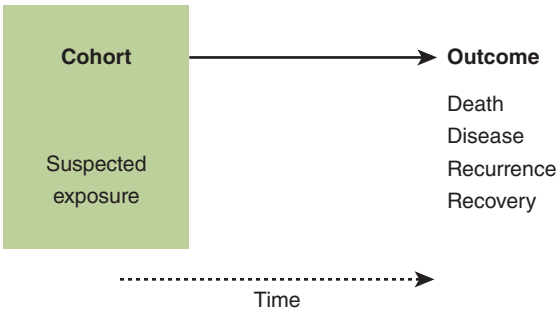


FIGURE 1-12 Basic components of a cohort study: exposure, time, and outcome.

*A definition of the term *cohort* broader than that found in the footnote in Section 1.2 is any designated and defined group of individuals who are followed or tracked over a given time period.[1]

is shown in **FIGURE 1-13**. In this hypothetical example, cohort members are followed for a given time period during which four events, such as incident disease cases or deaths (which appear in Figure 1-13 as lines ending with a dot), occur. In addition to these four events, seven individuals are lost to follow-up during the study period. These losses (represented in Figure 1-13 as arrows) are usually designated as *censored observations* or *withdrawals* and need to be taken into account for the calculation of incidence. Using the actuarial life-table approach as an example (see Chapter 2, Section 2.2.1), incidence can be estimated as the number of events occurring during the follow-up period divided by the number of subjects in the cohort at baseline minus one-half of the losses. Thus, for the hypothetical cohort in Figure 1-13, the incidence of disease is $4/[1000 - (1/2 \times 7)] = 4.01/1000$.

In the *cohort study*'s most representative format, a defined population is identified. Its subjects are classified according to exposure status, and the incidence of the disease (or any other health outcome of interest) is ascertained and compared across exposure categories (**FIGURE 1-14**). For example, based on the hypothetical cohort schematically represented in Figure 1-13 and assuming that the prevalence of the exposure of interest is 50%, **FIGURE 1-15** outlines the follow-up separately for exposed ($n = 500$) and unexposed ($n = 500$) individuals. Data analysis in this simple situation is straightforward, involving a comparison of the incidence of disease between exposed and unexposed persons, using as the denominator the "population at risk." For example, using the actuarial life-table approach previously mentioned for the hypothetical cohort study depicted in Figure 1-15, the incidence of disease in exposed individuals is $3/[500 - (1/2 \times 4)] = 6.02/1000$

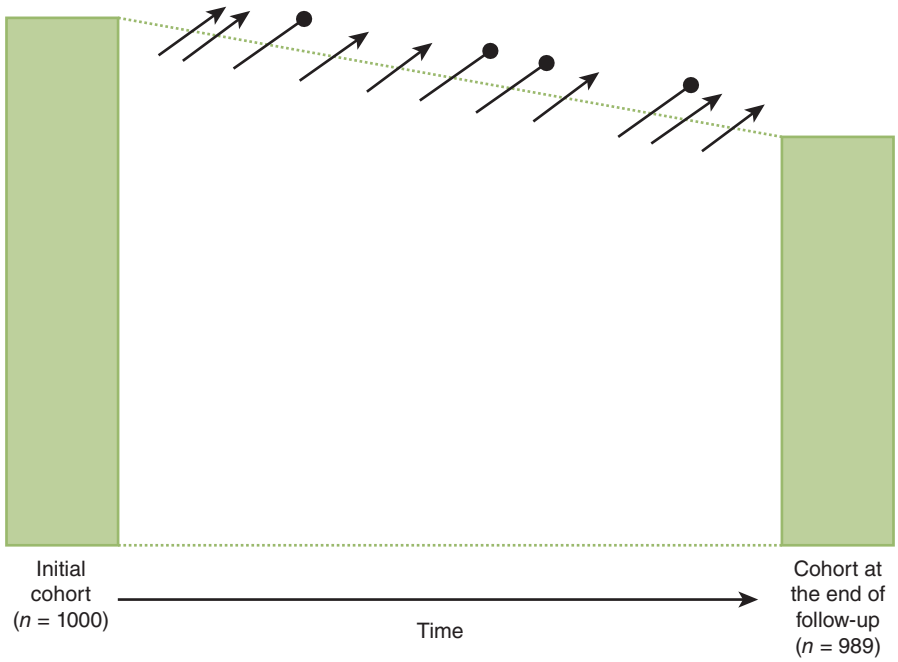


FIGURE 1-13 Diagram of a hypothetical cohort of 1000 subjects. During the follow-up, four disease events (line segments ending in dots) and seven losses to follow-up (arrows) occur so that the number of subjects under observation at the end of the follow-up is 989.

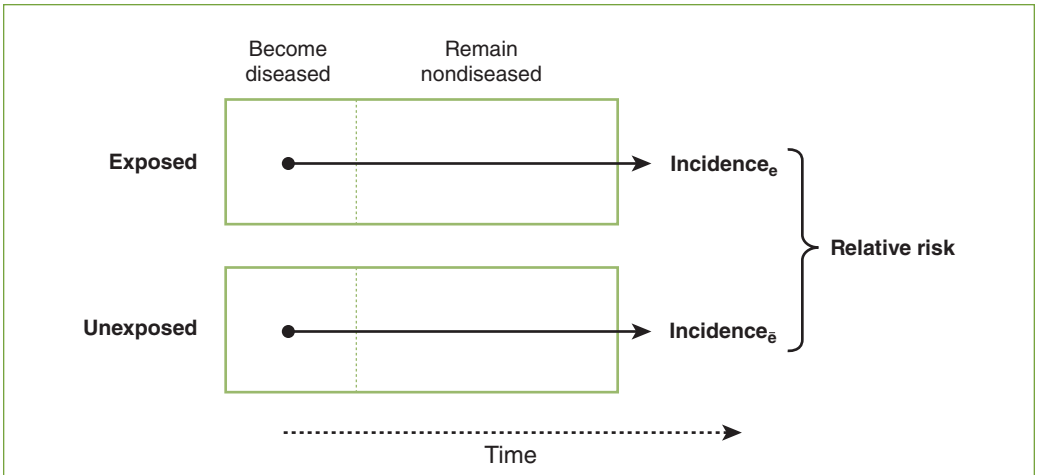


FIGURE 1-14 Basic analytical approach in a cohort study.

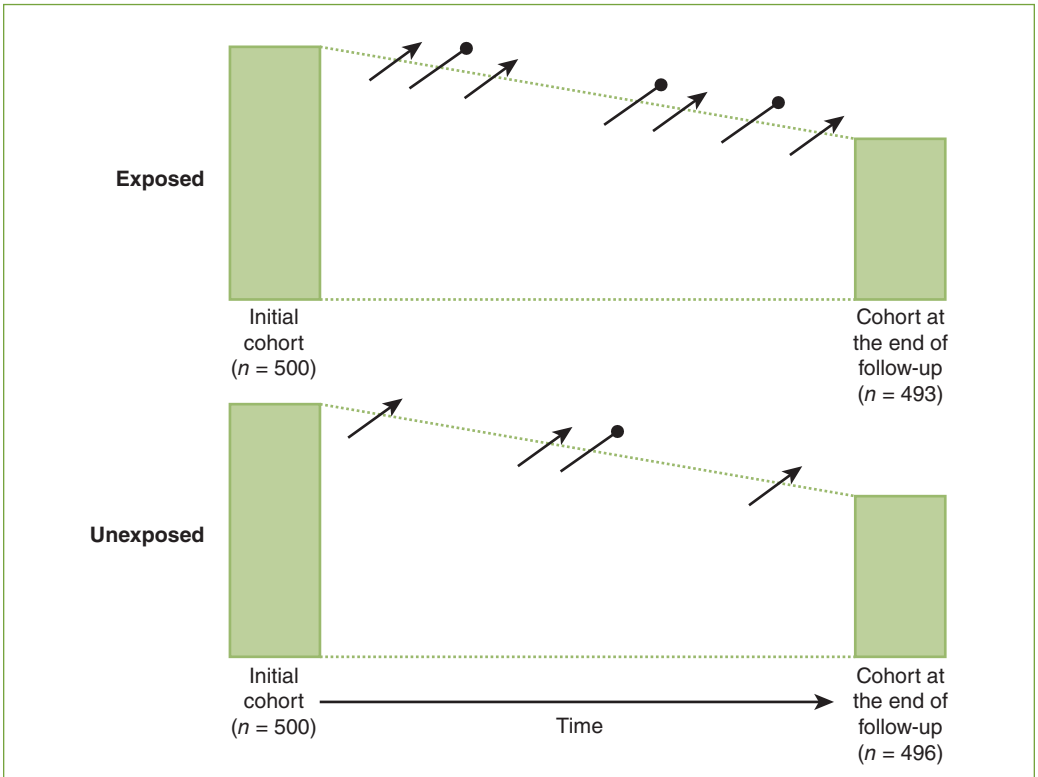


FIGURE 1-15 Same cohort study as in Figure 1-13, but the ascertainment of events and losses to follow-up is done separately among those exposed and unexposed.

and in unexposed is $1/[500 - (1/2 \times 3)] = 2.01/1000$. After obtaining incidence in exposed and unexposed, typically the relative risk is estimated (Chapter 3, Section 3.2.1); that is, these results would suggest that exposed individuals in this cohort have a risk approximately three times higher than that of unexposed individuals (relative risk $= 6.02/2.01 \approx 3.0$).

An important assumption for the calculation of incidence in a cohort study is that individuals who are lost to follow-up (the arrows in Figures 1-13 and 1-15) are similar to those who remain under observation with regard to characteristics affecting the outcome of interest (see Chapter 2). The reason is that even though techniques to "correct" the denominator for the number (and timing) of losses are available (see Chapter 2, Section 2.2), if the average risk of those who are lost differs from that of those remaining in the cohort, the incidence based on the latter will not represent accurately the true incidence in the initial cohort (see Chapter 2, Section 2.2.1). If, however, the objective of the study is an *internal comparison* of the incidence between exposed and unexposed subjects, even if those lost to follow-up differ from the remaining cohort members, as long as the biases caused by losses are similar in the exposed and the unexposed, they will cancel out when the relative risk is calculated (see Chapter 4, Section 4.2). Thus, a biased relative risk caused by losses to follow-up is present only when losses are differential in exposed and unexposed subjects with regard to the characteristics influencing the outcome of interest—in other words, when losses are affected by both exposure and disease status.

Cohort studies are defined as *concurrent*[3] (or truly "prospective"[32]) when the cohort is assembled at the present time—that is, the calendar time when the study starts—and is followed up toward the future (**FIGURE 1-16**). The main advantage of concurrent cohort studies is that the baseline exam,

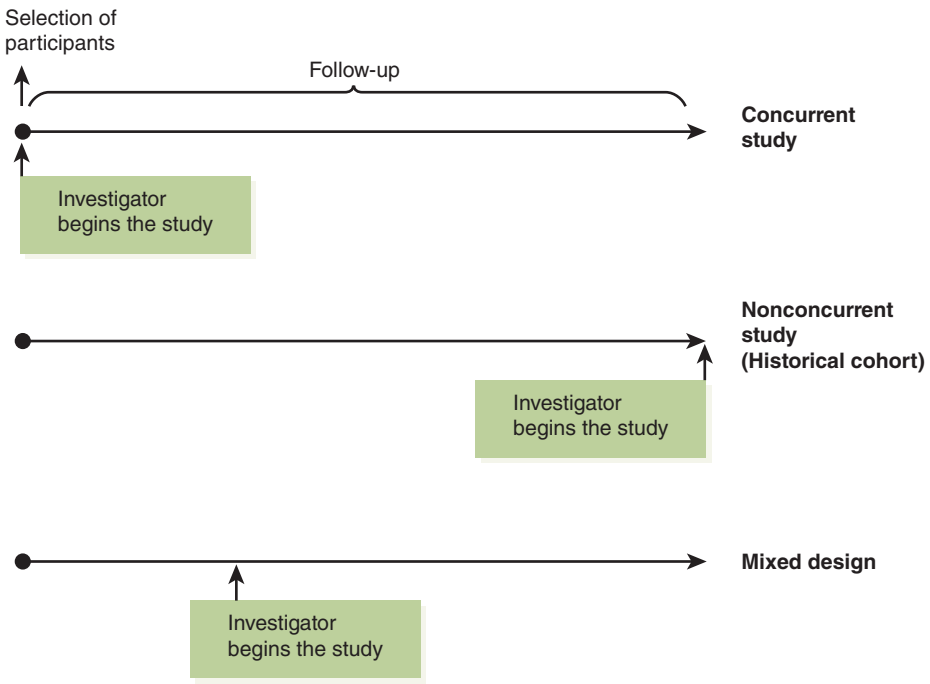


FIGURE 1-16 Types of cohort studies.

methods of follow-up, and ascertainment of events are planned and implemented for the purposes of the study, thus best fitting the study objectives; in addition, quality control measures can be implemented as needed (see Chapter 8). The disadvantages of concurrent studies relate to the amount of time needed to conduct them (results are available only after a sufficient number of events have been accumulated) and their usually elevated costs. Alternatively, in *nonconcurrent* cohort studies (also known as *historical* or *retrospective* cohort studies), a cohort is identified and assembled in the past on the basis of existing records and is "followed" to the present time (i.e., the time when the study is conducted) (Figure 1-16). An example of this type of design is a 1992 study in which the relationship between childhood body weight and subsequent adult mortality was examined nonconcurrently on the basis of existing records of weight and height values obtained from 1933 through 1945 in school-age children who were linked to adult death records.[33] The nonconcurrent design is also useful in occupational epidemiology, as occupational records can be linked to mortality or cancer registries. For example, a cohort of all electricians working in Norway in 1960 was followed nonconcurrently through 1990 to study the relationship of electromagnetic radiation to cancer incidence.[34] *Mixed designs* with both nonconcurrent and concurrent follow-up components are also possible (Figure 1-16). Nonconcurrent cohort studies are obviously less expensive and can be done more expeditiously than concurrent studies. Their main disadvantage is an obligatory reliance on available information; as a result, the type or quality of exposure or outcome data may not be well suited to fulfill the study objectives.

1.4.2 Case-Control Study

As demonstrated by Cornfield[35] and discussed in basic epidemiology textbooks (e.g., Gordis[3]), the case-control design is an alternative to the cohort study for investigating exposure–disease associations. In contrast to a cohort study, in which exposed and unexposed individuals are compared with regard to the disease incidence (or some other mean value for the outcome) (Figure 1-14), a case-control study compares cases (usually diseased individuals) and controls (e.g., nondiseased individuals) with respect to their level of exposure to a suspected risk factor. When the risk factor of interest is a categorical characteristic (e.g., former, current, and never smokers), the typical analytical approach in case-control studies is to compare the odds of exposure in cases with that in controls by calculating the exposure *odds ratio* (**FIGURE 1-17**), which is often an appropriate estimate of the relative risk (see Chapter 3, Section 3.2.1). When the exposure of interest is a continuous trait, its mean levels (e.g., mean blood pressure) can be compared between cases and controls.

The case-control study design has important advantages over the cohort design, particularly over the concurrent cohort study, as the need for a follow-up time is avoided, thus optimizing speed and efficiency.[3]

Case-Based Case-Control Study

In the simplest strategy for the selection of groups in a case-control study, cases occurring over a specified time period and noncases are identified. An example of this strategy, sometimes called *case-based case-control study*, is a study in which incident cases are identified as the individuals in whom the disease of interest was diagnosed (e.g., breast cancer) in a certain hospital during a given year and controls are selected from among members of the community served by this hospital who did not have a diagnosis of the disease of interest by the end of that same year. If exposure data are obtained through interviews, it is necessary to assume that recall or other biases will not distort the findings (see Chapter 4). If only living cases are included in the study, it must be also

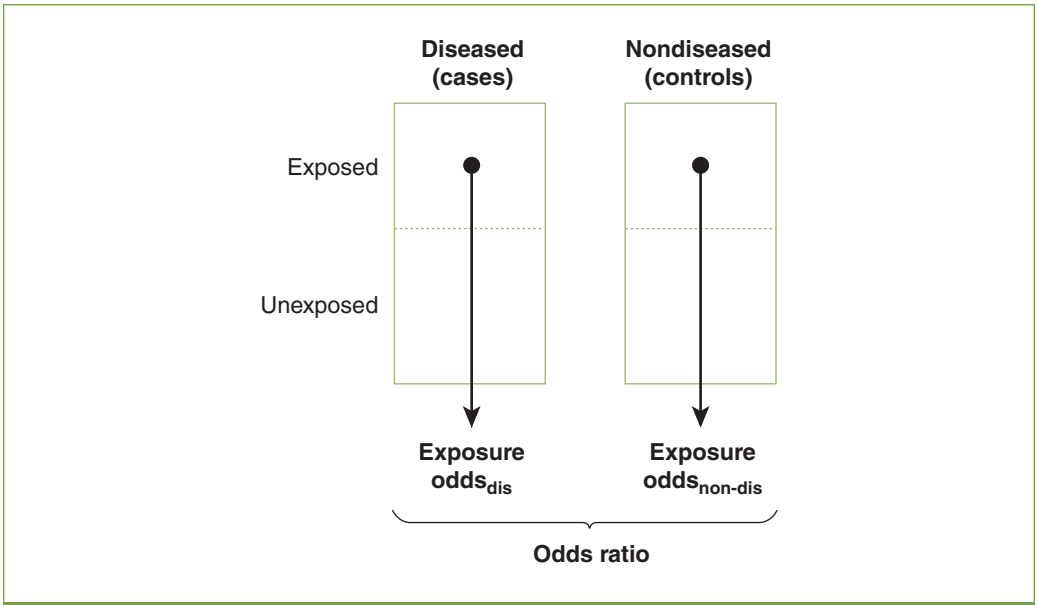


FIGURE 1-17 Basic design and analytical approach in a case-control study.

assumed that cases who survive through the time when the study is done are representative of all cases with regard to the exposure experience (**FIGURE 1-18**). Furthermore, to validly compare cases and controls regarding their exposure status, it is necessary to assume that they originate from the same reference population, that is, from a more or less explicitly identified cohort, as depicted in Figure 1-18. In other words, for a case-control comparison to represent a valid alternative to a cohort study analysis, cases and controls are expected to belong to a common reference population (or to a similar reference population or study base, discussed later in this chapter). An example is given by a population-based study of acoustic neuroma conducted by Fisher and colleagues.[36] These authors identified all 451 diagnosed acoustic neuroma cases in Sweden and selected age-, sex-, and region-matching controls (more on matching later in this chapter). In general, however, it is frequently difficult to define the source cohort in a case-control study, as, for example, in a case-based study in which the cases are ascertained in a single hospital A but controls are selected from a population sample. In this example, it is important to consider the correspondence between the patient population of hospital A and the population of the geographic area from which controls are sampled. Thus, for example, if hospital A is the only institution to which area residents can be admitted and all cases are hospitalized, a sample of the same area population represents a valid control group. If, however, residents use hospitals outside of the area and hospital A admits patients from other areas, alternative strategies have to be considered to select controls who are representative of the theoretical population from which cases originate (e.g., matching controls to cases by neighborhood of residency).

The assumption that cases and controls originate from the same hypothetical source cohort (even if undefined) is critical when judging the internal validity of case-control data. Ideally, controls should have been eligible for inclusion in the case group had they developed the disease of interest. Pragmatically, although it is not strictly necessary that cases and controls be chosen from exactly the *same* reference population, both groups must originate from populations having *similar*

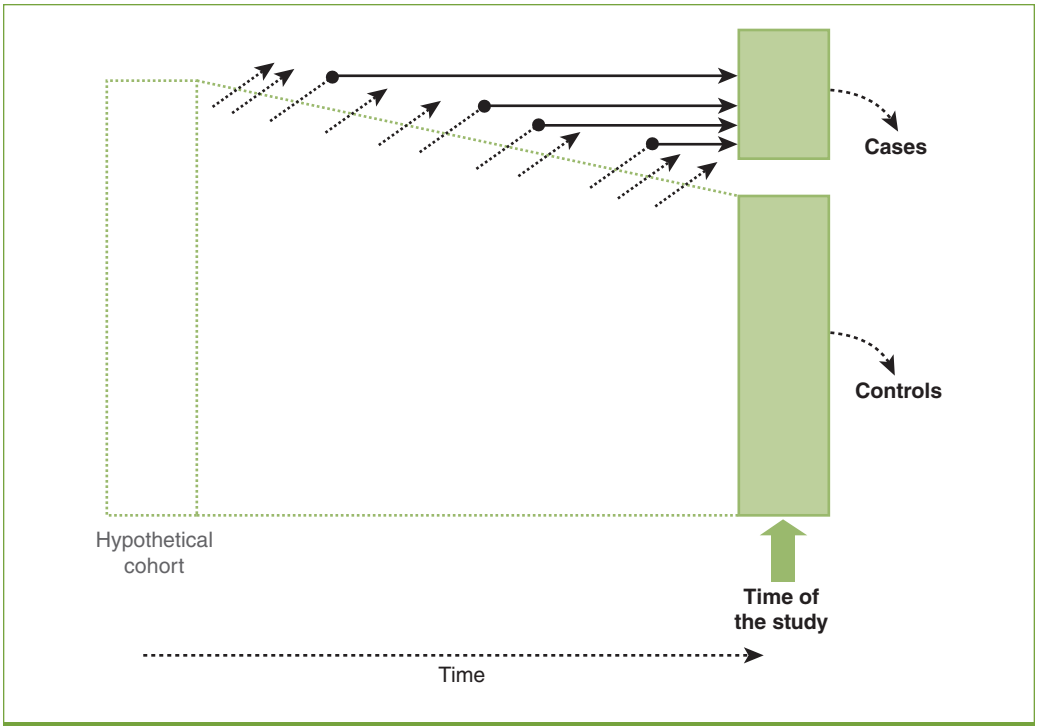


FIGURE 1-18 Hypothetical case-based case-control study assuming that cases and controls are selected from a hypothetical cohort, as in Figure 1-13. The case group is assumed to include all cases that occurred in that hypothetical cohort up to the time when the study is conducted (dots with horizontal arrows ending at the "case" bar); that is, all of them are assumed to be alive and available to participate in the study. Controls are selected from among those without the disease of interest (noncases) at the time when the cases are identified and assembled. Broken diagonal lines with arrows represent losses to follow-up.

EXHIBIT 1-2 Different ways to conceptualize an ideal control group.

- Sample of individuals without the disease selected from the same reference population (study base) from which cases were selected
- Group of individuals who, if they had developed the case disease, would have been included in the case group
- Group of individuals without the case disease who are subjected to the same selection process as the cases

relevant characteristics. Under these circumstances, the control group can be regarded as a reasonably representative sample of the case reference population. **EXHIBIT 1-2** summarizes the essential features of an ideal control group (i.e., one that maximizes its comparability with the case group).

When cases and controls are not selected from the same (or similar) reference population(s), *selection bias* may ensue (see Chapter 4). Selection bias may occur even if cases and controls are from the same "hypothetical" cohort; this happens when "losses" occurring *before* the study groups are

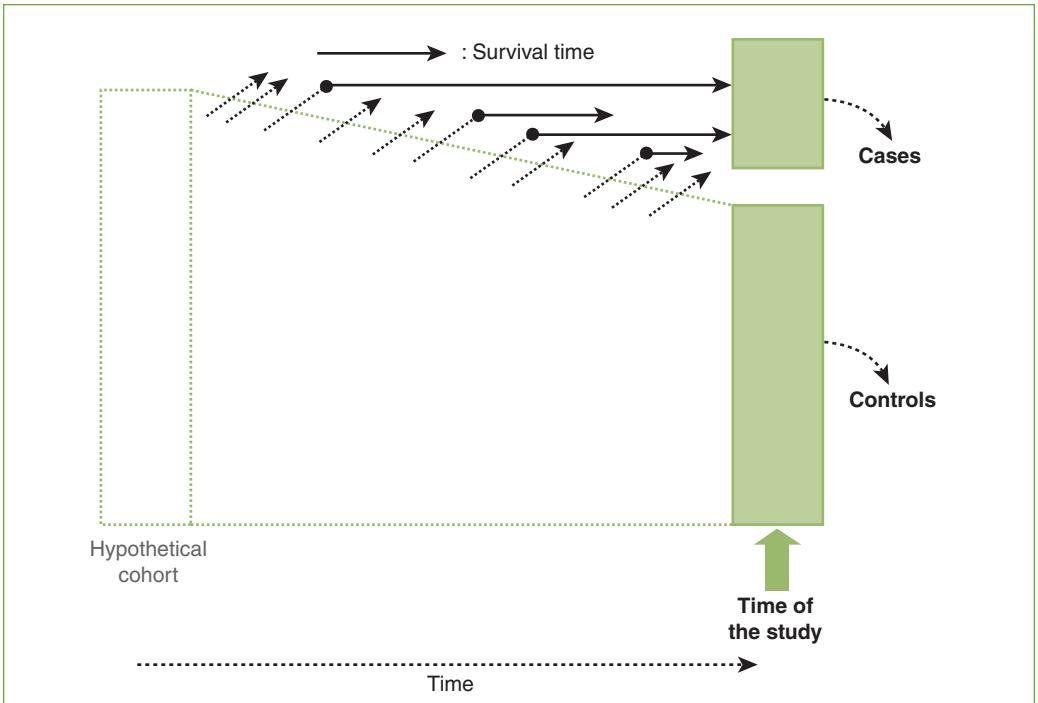


FIGURE 1-19 Survival bias in a case-based case-control study carried out cross-sectionally. Only cases with long survival after diagnosis (best prognosis) are included in the case group. In this hypothetical example, the horizontal lines starting in the cases' dot symbols represent survival times; note that only two of the four cases are included in the study. Broken diagonal lines with arrows represent losses to follow-up.

selected affect their comparability. For example, if losses among potential controls include a higher proportion of individuals with low socioeconomic status than losses among cases, biased associations may be found with exposures related to socioeconomic status. This example underscores the close relationship between selection bias in case-control studies and differential losses to follow-up in cohort studies. In this context, consider the similarity between Figures 1-18 and 1-13 in that the validity of the comparisons made in both cohort (Figure 1-13) and case-control (Figure 1-18) studies depends on whether the losses (represented by diagonal arrows in both figures) affect the representativeness of the baseline cohort (well defined in Figure 1-13, hypothetical in Figure 1-18) with regard to both exposure and outcome variables.

Deaths caused by either other diseases or the disease of interest comprise a particular type of (prior) "loss" that may affect comparability of cases and controls. For the type of design represented in Figure 1-18, characterized by cross-sectional ascertainment of study subjects, those who die before they can be included in the study may have a different exposure experience compared to the rest of the source population. In addition, by definition, this design identifies only cases that are prevalent at the time of the study, which will tend to be those with the longest survival (**FIGURE 1-19**). These types of selection bias constitute generic problems affecting cross-sectional ascertainment of study participants; another problem is recall bias, which results from obtaining past exposure data long after disease onset. (For a detailed discussion of these and other biases in case-control studies, see Chapter 4.)

It should be emphasized that although cross-sectional ascertainment of cases and controls is often carried out, it is not a *sine qua non* feature of case-based case-control studies. An alternative strategy, which aims at minimizing selection and recall biases and should be used whenever possible, is to ascertain cases concurrently (i.e., to identify and obtain exposure information on incident cases as soon as possible after disease onset). An example of this strategy is a study of risk factors for oral clefts conducted in Denmark.[37] In this study, case mothers were women who were hospitalized and gave birth to a live child with cleft lip and/or palate (without other malformations) between 1991 and 1994. Controls were the mothers of the two preceding births in the hospital where the case mother had given birth. Both case and control mothers were concurrently interviewed by trained nurses with regard to previous pregnancies, medications, smoking, diet, and other environmental and occupational exposures.

Case-Control Studies Within a Defined Cohort

When cases are identified within a well-defined cohort, it is possible to carry out *nested case-control* or *case-cohort studies*. These designs have received considerable attention in recent years[38,39] in part because of the increasing number of well-established large cohort studies that have been initiated and continued during the past few decades and in part because of recent methodological and analytical advances.

Case-control studies within a cohort are also known as *hybrid* or *ambidirectional designs*[40] because they combine some of the features and advantages of both cohort and case-control designs. In these studies, although the selection of the participants is carried out using a case-control approach (Figure 1-17), it takes place within a well-defined cohort. The case group consists of all (or a representative sample of) individuals with the outcome of interest occurring in the defined cohort over a specified follow-up period (diagonal lines ending with a dot in Figure 1-13). The control group can be selected from either individuals at risk at the time each case occurs or the baseline cohort. These two alternatives, respectively known as *nested case-control* and *case-cohort designs*, are described in the next paragraphs.

- Controls are a random sample of the individuals remaining in the cohort at the time each case occurs (**FIGURE 1-20**). This *nested case-control design* is based on a sampling approach known as *incidence density sampling*[40,41] or *risk-set sampling*.[23] Cases are compared with a subset (a sample) of the "risk set," that is, the cohort members who are at risk (i.e., who could become a case) at the time when each case occurs. By using this strategy, cases occurring later in the follow-up are eligible to be controls for earlier cases. Incidence density sampling is the equivalent of matching cases and controls on duration of follow-up (see Section 1.4.5) and permits the use of straightforward statistical analysis techniques (e.g., standard multiple regression procedures for the analysis of matched and survival data) (see Chapter 7, Section 7.4.6).
- Controls are selected as a random sample of the total cohort at baseline (**FIGURE 1-21**). In this design, known as *case-cohort*, the control group may include individuals who become cases during the follow-up (diagonal lines ending with a dot in Figure 1-21). Because of the potential overlap between the case and the cohort random sample (control) groups, special techniques are needed for the analysis of this type of study (see Chapter 7, Section 7.4.6).[39] An important advantage of the case-cohort design is that a sample of the baseline cohort can serve as a control group for different sets of cases occurring in the same cohort. For example, in a report from the Atherosclerosis Risk in Communities (ARIC) Study, Dekker et al.[42] used a case-cohort approach to analyze the relationship between heart rate variability (a marker of autonomic

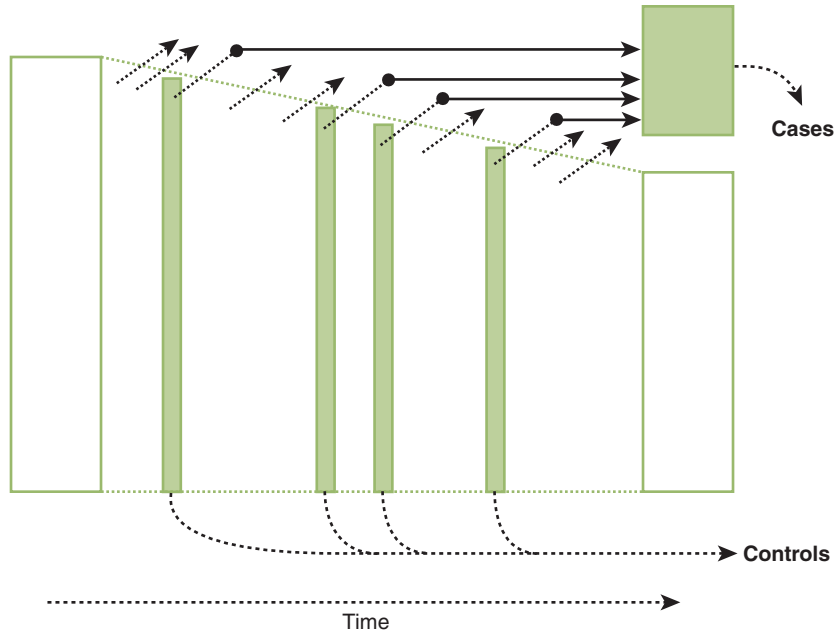


FIGURE 1-20 Nested case-control study in which the controls are selected at each time when a case occurs (incidence density sampling). Cases are represented by a dot connected to a horizontal arrow. Broken diagonal lines with arrows represent losses to follow-up.

nervous system function) and several outcomes. ARIC is a cohort study of approximately 15,800 men and women aged 45 to 64 years at the study's outset (1987–1989). During a 6-year follow-up period, 443 deaths from all causes, 140 cardiovascular deaths, 173 cancer deaths, and 395 incident CHD cases were identified. As a comparison group for all four of these case groups, a single sample of 900 baseline cohort participants was identified. Heart rate variability was thus measured at baseline in electrocardiography (ECG) records of these 900 controls and on the records of the individuals in each of the four case groups. (An incidence density-type nested case-control design would have required that, for each case group, a separate control group be selected, for a total of four different control groups.)

An additional practical advantage of the case-cohort approach is that if the baseline cohort sample is representative of the source population, risk factor distributions and prevalence rates needed for population attributable risk estimates (Chapter 3, Section 3.2.2) can be obtained.

Another consideration in these types of designs is whether to include or exclude the cases from the pool of eligible controls, that is, the baseline cohort sample or the risk sets in case-cohort and nested case-control designs, respectively. The analytical implications of this choice are discussed in Chapter 3, Section 3.4.1.

In general and regardless of which of the previously mentioned control selection strategies is used (e.g., nested case-control or case-cohort), the likelihood of selection bias tends to be diminished in comparison with the traditional case-based case-control study. This is because cases and controls

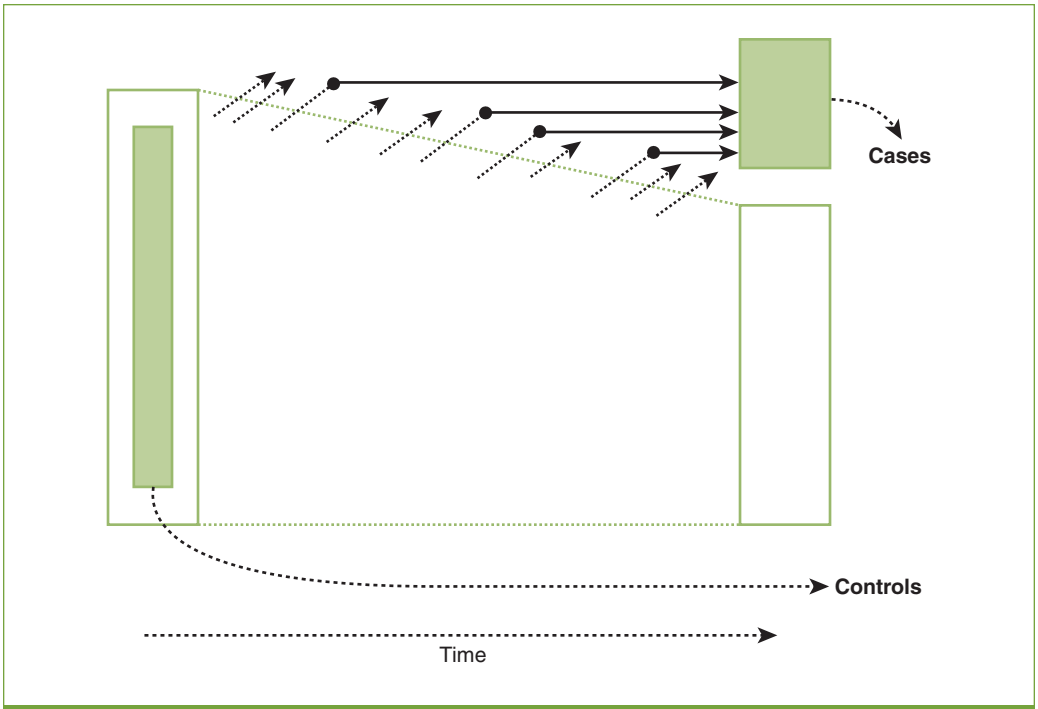


FIGURE 1-21 Case-control study in which the controls are selected from the baseline cohort (case-cohort study). Cases are represented by a dot connected to a horizontal arrow. Broken diagonal lines with arrows represent losses to follow-up.

are selected from the same (defined) source cohort, thus essentially guaranteeing that the ideal features listed in Exhibit 1-2 are met. Furthermore, other types of biases (such as ascertainment bias and temporal bias; see Chapter 4) are also less likely because (as in any traditional cohort study) exposures are assessed *before* the disease occurs.

When Should a Case-Control Study Within a Cohort Be Used Instead of a Comparison of Exposed and Unexposed in the Full Cohort?

If a well-defined cohort with prospectively collected follow-up data are available, why not simply analyze the data from the entire cohort (as in Figure 1-15)? What would be the advantage of limiting the study to a comparison of incident cases and a subset of the cohort (controls)? The answer is that the nested case-control and case-cohort designs are fundamentally efficient when *additional information that was not obtained* or measured for the whole cohort is needed. A typical situation is a concurrent cohort study in which biological (e.g., serum) samples are collected at baseline and stored in freezers. After a sufficient number of cases have been accrued during the follow-up, the frozen serum samples for cases and for a sample of controls can be thawed and analyzed. This strategy not only reduces the cost that would have been incurred if the analyte(s) of interest had been assayed in the entire cohort but also preserves serum samples for future analyses. A similar situation arises when the assessment of key exposures or confounding variables (see Chapter 5)

requires labor-intensive data collection activities, such as data abstraction from medical or occupational records. Collecting this additional information in cases and a sample of the total cohort (or of the noncases) is a cost-effective alternative to using the entire cohort. Thus, case-control studies within a cohort combine and take advantage of both the methodological soundness of the cohort design (i.e., limiting selection bias) and the efficiency of the case-control approach. Some examples follow:

- A study was conducted to examine the relationship of military rank and radiation exposure to brain tumor risk within a cohort of male members of the U.S. Air Force who had had at least one year of service between 1970 and 1989.[43] In this study, for each of the 230 cases of brain tumor identified in that 20-year period, four race- and year-of-birth–matched controls (*noncases*) were randomly selected among Air Force employees who were active *at the time the case was diagnosed* (for a total of 920 controls). The reason for choosing a nested case-control design (i.e., a design based on incidence density sampling; see Figure 1-20) instead of using the entire cohort of 880,000 U.S. Air Force members in this study was that labor-intensive abstraction of occupational records was required to obtain accurate data on electromagnetic radiation exposure as well as other relevant information on potentially confounding variables. An alternative strategy would have been not to exclude cases from the eligible control sampling frame (discussed previously). Yet another strategy would have been to use a case-cohort design whereby controls would have been sampled from among Air Force cohort members at the beginning of their employment (i.e., at baseline; see Figure 1-21).
- Graham et al.'s[44] study of the relationship of nonsteroidal anti-inflammatory drugs (NSAIDs) to acute myocardial infarction and sudden cardiac deaths is an additional example of a nested case-control study. The study base comprised NSAID-treated patients aged 18 to 84 years who had at least 12 months of Kaiser Permanente coverage without a diagnosis of cancer, renal failure, liver failure, severe respiratory disease, organ transplantation, or HIV/AIDS. Cases were identified during a 3-year follow-up. Controls were selected by risk-set (density) sampling using a 4:1 ratio. In addition to the built-in matching for date of case event, which is characteristic of the nested case-control design, controls were matched to cases on age, sex, and health plan region (North or South). Note that, because the study was conducted within a managed care organization membership, the pool of potential controls was very large, which allowed the authors to conduct fine risk-set sampling, that is, find controls who matched the cases almost exactly according to the date and all the other variables.
- Dekker et al.'s[42] study on heart rate variability in relation to mortality and CHD incidence in the ARIC Study (discussed previously) is an example of the application of a case-cohort design (Figure 1-21). In this study, an elaborate and time-consuming coding of the participant's ECG was required to characterize heart rate variability. Conducting such coding in the entire cohort (approximately 15,800 subjects) would have been prohibitively expensive. By using a case-cohort design, the authors were able to limit the ECG coding to only 900 controls and the individuals in the four case groups.
- Another example of sampling controls from the total baseline cohort (i.e., a case-cohort design) (Figure 1-21) is given by a study conducted by Nieto et al.[45] assessing the relationship of *Chlamydia pneumoniae* antibodies in serum collected at baseline to incident CHD in the ARIC Study. Over a 5-year follow-up period, a total of 246 cases of incident CHD (myocardial infarctions or coronary deaths) were identified. The comparison group in this study consisted of a sample of 550 participants of the total baseline cohort (known as a subcohort), which included 10 of the 246 individuals who later developed CHD (incident cases), a fact that needs

to be considered in the statistical analyses of these data (also see Chapter 7, Section 7.4.6). For this study, *C. pneumoniae* IgG antibody levels were determined in sera of only the cases and cohort sample, that is, in only approximately 800 individuals rather than in the approximately 15,800 cohort participants required for a full cohort analysis. In addition to the estimation of risk ratios expressing the relationship between *C. pneumoniae* antibodies and incident CHD, the selection of a random sample of the cohort in Nieto et al.'s study has the advantage over the incidence density nested case-control approach of allowing the estimation of the prevalence of *C. pneumoniae* infection in the cohort (and, by extension, in the reference population) and thus also of population attributable risk (Chapter 3, Section 3.2.2). As in the previous example, the control group could have been used for the assessment of the role of *C. pneumoniae* for a different outcome. For example, after a sufficient number of stroke cases were accrued, a study of the relationship between *C. pneumoniae* infection and stroke incidence could have been conducted; the only additional costs would have been those related to measuring *C. pneumoniae* antibodies in the stroke cases, as the measurements would have already been available in the control group.

- A third example of a case-cohort design is a study of the relationship of serum estrogens and estrogen metabolites to breast cancer in postmenopausal women.[46] This study is an example of an observational analysis of data from a clinical trial that was originally designed to address a different question. The data came from the Fracture Intervention Trial (FIT), a randomized, placebo-controlled trial designed to test whether alendronate would reduce the rate of fractures in women with low bone mineral density.[47] About 15,500 individuals who were screened for participation in FIT were considered for inclusion in this ancillary study, of which almost 14,000 were eligible (**FIGURE 1-22**). After a mean follow-up of about 8 years, 505 cases were identified in the whole cohort. After a few exclusions, 407 cases were compared with a random sample of 496 cohort participants (the "subcohort"). Note that the subcohort includes both 487 noncases and 9 incident cases developing during the follow-up period.

1.4.3 Cross-Sectional Studies

In a cross-sectional study design, a sample of the (or the total) reference population is examined at a given point in time. Like the case-control study, the cross-sectional study can be conceptualized as a way to analyze cohort data, albeit an often flawed one, in that it consists of taking a "snapshot" of a cohort by recording information on disease outcomes and exposures at a single point in time (**FIGURE 1-23**).* Accordingly, the case-based case-control study represented schematically in Figure 1-19 can also be regarded as a cross-sectional study, as it includes cross-sectionally ascertained prevalent cases and noncases (i.e., cohort participants who survived long enough to be alive at the time of the study). It follows that when cross-sectional data are obtained from a defined reference population or cohort, the analytical approach may consist of either comparing point prevalence rates for the outcome of interest between exposed and unexposed individuals or using a "case-control" strategy, in which prevalent cases and noncases are compared with regard to odds of exposure (see Chapters 2 and 3).

Even though any population-based cross-sectional morbidity survey could (at least theoretically) offer the opportunity to examine exposure/outcome associations,[50] cross-sectional analyses

*Cross-sectional studies can also be done periodically for the purpose of monitoring trends in prevalence of diseases or prevalence or distributions of risk factors, as in the case of the U.S. National Health Surveys.[48,49]

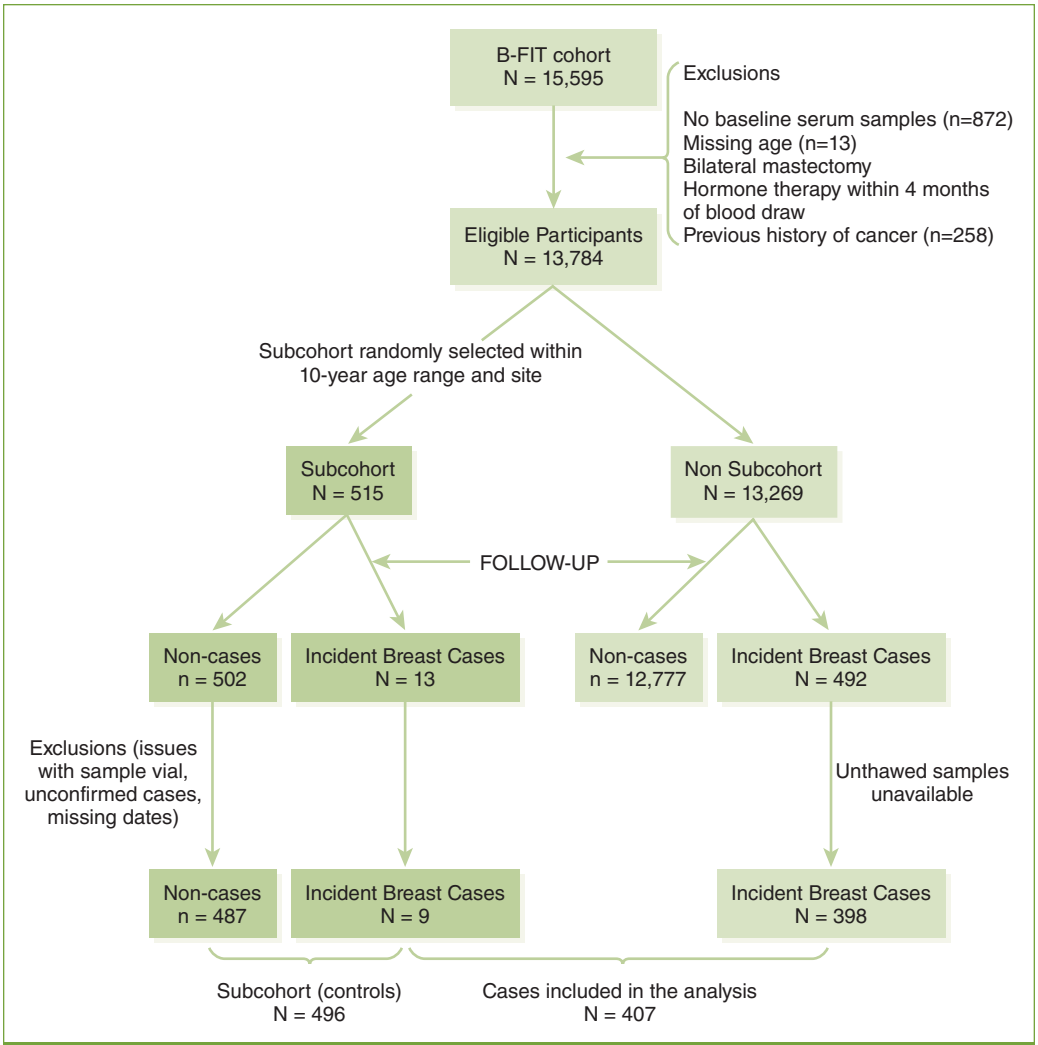


FIGURE 1-22 Design of the Breast and Bone Follow-Up to the Fracture Intervention Trial (B-FIT) case-cohort study.

Reprinted with permission from Dallal CM, Tice JA, Buist DSM, et al. Estrogen metabolism and breast cancer risk among postmenopausal women: a case-cohort study with B-FIT. *Carcinogenesis.* 2014;35:346-355.[46] By permission of Oxford University Press.

of baseline information in cohort studies are especially advantageous. This is particularly the case when examining subclinical outcomes less amenable to survival bias. In the context of baseline data from a cohort study, it may be of interest to verify whether results from cross-sectional analyses are consistent with subsequent analyses of longitudinal data. For example, in the ARIC Study, the cross-sectional associations found at baseline of both active and passive smoking with asymptomatic carotid artery atherosclerosis (defined by B-mode ultrasound)[51] were subsequently confirmed by assessing progression of atherosclerosis.[52]

The conditions influencing the validity of associations inferred from cross-sectional data are discussed in detail in Chapter 4 (Section 4.4.2).

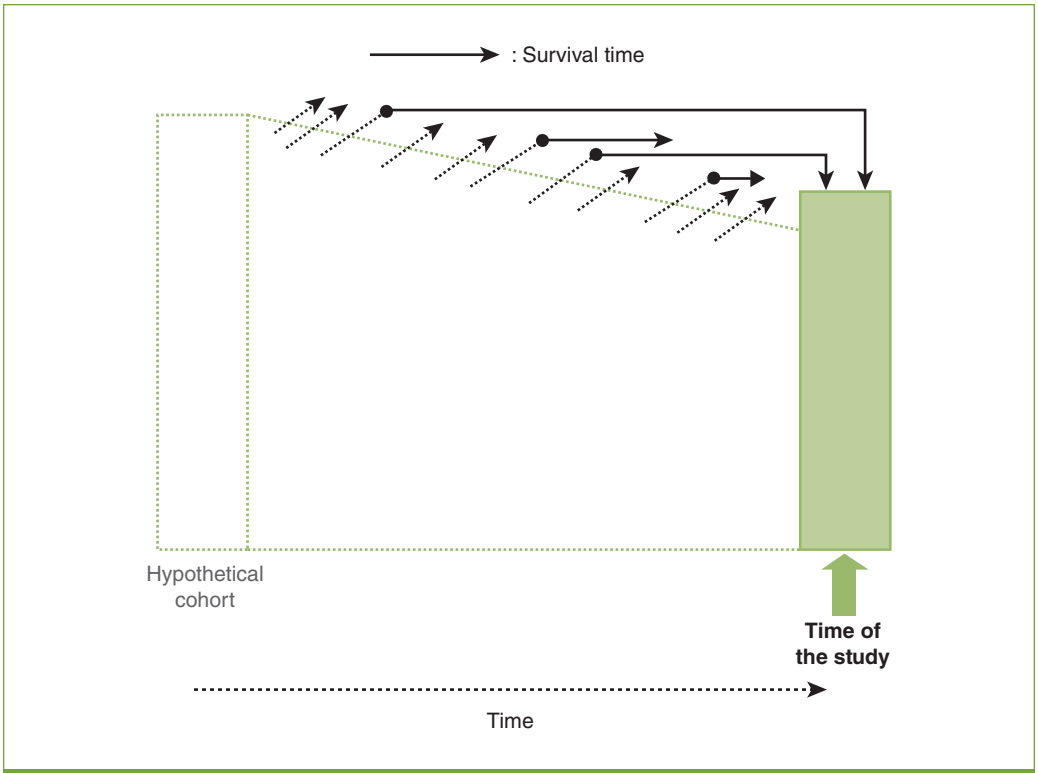


FIGURE 1-23 Schematic representation of a cross-sectional study. Notice that, as in the case-control design represented in Figure 1-19, the sample will be more likely to include cases with long survival times—to be discussed in Chapter 4. Broken diagonal lines with arrows represent losses to follow-up. Cases are represented by dots connected to horizontal arrows.

1.4.4 Case-Crossover Design

Initially proposed by Maclure,[53] the case-crossover study design consists of comparing the exposure status of a case immediately before its occurrence with that of the same case at some other prior time (e.g., the average level of exposure during the previous year). It is especially appropriate to study acute (brief) exposures that vary over time and that produce a transient change in risk of an acute condition after a short latency (incubation) period. For example, this design has been used to study acute triggers of intracranial aneurysms, such as vigorous physical exercise,[54] and asthma, such as traffic-related air pollution.[55]

The case-crossover design represents a special type of matching (see Section 1.4.5) in that individuals serve as their own controls. Thus, the analytical unit is time. The time just preceding the acute event ("case" time) is compared with some other time ("control" time). In this design, all fixed individual characteristics that might confound the association (e.g., gender and genetic susceptibility) are controlled for. This design, however, must assume that the disease does not have an undiagnosed stage that could inadvertently affect the exposure of interest ("reverse causation"). It also assumes that the exposure does not have a cumulative effect, as its strategy is to focus on its

TABLE 1-4 Odds ratios and 95% confidence intervals (CIs) for sleeping less than 10 hours/day in relation to unintentional injuries in children.

Study subjects	*n*	Ca+, Co+	Ca+, Co–	Ca–, Co+	Ca–, Co–	Odds ratio* (95% CI)
All children	292	62	26	14	190	1.86 (0.97, 3.55)
Boys	181	40	21	9	111	2.33 (1.07, 5.09)
Girls	111	22	5	5	79	1.00 (0.29, 3.45)

*Ratio of number of pairs of Ca+, Co– to the number of pairs of Ca–, Co+ (see Chapter 3, Section 3.4.1).
Data from Valent F, Brusaferro S, Barbone F. A case-crossover study of sleep and childhood injury. *Pediatrics.* 2001;107:E23.[57]

acute effect on the suspected outcome. Provided that data are available, other time-related differences that could *confound* the comparison between the case-control times (e.g., differences in the weather or in other varying environmental conditions) could be controlled for in the analyses (see Chapters 5 and 7).

Information on exposures in case-crossover studies is either obtained objectively—as, for example, in studies of environmental exposures, such as particulate matter[56]—or relies on participants' recall, thus being subject to recall bias (see Chapter 4, Section 4.3.1).

An example of a case-crossover design is given by a study conducted by Valent et al.[57] in which the association between sleep (and wakefulness) duration and childhood unintentional injury was examined in 292 children. The "case" and "control" periods were designated as the 24 and the 25 to 48 hours preceding the injury, respectively. **TABLE 1-4** presents results of the matched-paired analysis in which the association of the exposure (sleeping less than 10 hours/day) with unintentional injury was found to be present only in boys, thus suggesting the presence of qualitative interaction with gender (see Chapter 6, Section 6.7.1). In addition to analyzing data using the ratio of discrepant pairs to estimate the odds ratio (Table 1-4), analysis of case-crossover study data can also be done by means of conditional logistic regression (see Chapter 7, Section 7.4.6), as done by these authors, with additional adjustment for day of the week when injury occurred (weekend vs weekday) and the activity risk level of the child (higher vs lower level of energy).

1.4.5 Matching

In observational epidemiology, an important concern is that study groups may not be comparable with regard to characteristics that may distort ("confound") the associations of interest. The issue of *confounding* is key in epidemiologic inference and practice and is discussed in detail in Chapter 5. Briefly, this issue arises when spurious factors (*confounding variables*) influence the direction and magnitude of the association of interest. For example, if a case-control study shows an association between hypertension (exposure) and coronary disease (outcome), it can be argued that this association may (at least in part) be due to the fact that coronary disease cases tend to be older than controls. Because hypertension is more frequently seen in older people, the difference in age between cases and controls may produce the observed association (or exaggerate its magnitude). Thus, if the question of interest is to assess the net relationship between hypertension and coronary

disease (*independently* of age), it makes intuitive sense to select cases and controls with the same or similar ages (i.e., *matched* on age). Similarly, a putative association between serum markers of inflammation (e.g., C-reactive protein) and the risk of CHD may result from confounding by smoking (as smoking increases both the risk of CHD and the levels of inflammatory markers). Recognizing this possibility, researchers matched cases and controls according to smoking status (current, former, or never smoker) in a nested case-control study showing an association between C-reactive protein levels and CHD.[58]

Matching in Case-Control and in Cohort Studies

The practice of matching is particularly common and useful in the context of *case-control studies* when trying to make cases and controls as similar as possible with regard to potential confounding factors. In addition to the two examples just cited, another example is the previously mentioned study of risk factors for oral clefts, in which cases and controls were matched according to place of birth (by selecting controls from the same hospital where the case mother had given birth) and time (by selecting for each case the two preceding births as controls). A special example of matching is given by the nested case-control study design based on incidence density sampling (Section 1.4.2, Figure 1-20). As discussed previously, this strategy results in matching cases and controls on follow-up time. In addition to time in the study, controls may be matched to cases according to other variables that may confound the association of interest. For example, in the U.S. Air Force study of brain tumors mentioned previously,[43] controls sampled from the risk sets at the time of occurrence of each case were additionally matched on birth year and race.

In contrast, in cohort studies, matching on potentially confounding variables is not common. Cohort studies are often large and examine a multitude of exposures and outcomes. Thus, alternative means to control for confounding are usually preferred (e.g., adjustment—see Chapter 7). Among the relatively rare instances in which matching is used in cohort studies are studies of prognostic factors for survival after cancer diagnosis in certain settings. For example, in a study examining age (the "exposure" of interest) as a prognostic factor in multiple myeloma patients following an autologous transplant,[59] older individuals ($\geq$ 65 years old) were matched to younger individuals ($<$ 65 years old) with regard to other factors that affect prognosis and that could thus confound the age–survival association (levels of β_2-microglobulin, albumin, creatinine, and C-reactive protein and the presence/absence of chromosomal abnormalities); the results of this study suggested that, after controlling for these variables, age is not a "biologically adverse" prognostic parameter in these patients.

Types of Matching

Previous examples concerned studies in which cases and controls were *individually* matched; that is, for each case, one or more controls with the relevant characteristics matching those of the cases were selected from the pool of eligible individuals. Individual matching according to naturally categorical variables (e.g., gender) is straightforward. When matching is conducted according to continuous variables (e.g., age), a matching range is usually defined (e.g., the matched control's age should be equal to the index case's age plus or minus 5 years). In this situation, as well as when continuous or ordinal variables are arbitrarily categorized (e.g., hypertensive/normotensive or current/former/never smoker), differences between cases and controls may remain, resulting in residual confounding (see Chapter 5, Section 5.5.4, and Chapter 7, Section 7.5).

Individual matching may be logistically difficult in certain situations, particularly when there is a limited number of potentially eligible controls and/or if matching is based on multiple variables. An alternative strategy is to carry out *frequency matching*, which consists of selecting a control group

to balance the distributions of the matching variable (or variables) in cases and controls but without doing a case-by-case individual matching. To carry out frequency matching, advance knowledge of the distribution of the case group according to the matching variable(s) is usually needed so that the necessary sampling fractions within each stratum of the reference population for the selection of the control group can be estimated. For example, if matching is to be done according to gender and age (classified in two age groups, $<$ 45 years and $\geq$ 45 years), four strata would be defined: females younger than 45 years, females aged 45 years or older, males younger than 45 years, and males aged 45 years or older. After the proportion of cases in each of these four groups is obtained, the number of controls to be selected from each gender–age stratum is chosen so that it is proportional to the distribution in the case group. An example is given in **EXHIBIT 1-3**.

If the controls are to be selected from a large population frame from which information on the matching variables is available, frequency matching can easily be done by stratified random sampling with the desirable stratum-specific sampling fractions. On the other hand, if this information is not available in advance (e.g., when controls are chosen from among persons attending a certain outpatient clinic), control selection can be done by systematic sampling and successively adding the selected individuals to each stratum until the desired sample size is reached for that stratum. Another strategy, if the distribution of cases according to matching variables is not known in advance but the investigators wish to select and obtain information on cases and controls more or less concurrently, is to obtain (and periodically update) provisional distributions of cases, thus allowing control selection to be carried out before all cases are identified.

When matching for several variables, and particularly when matching for continuous variables is desired, the so-called *minimum Euclidean distance measure method* is a useful alternative.[60] For example, in the study of age as a prognostic factor after transplantation in multiple myeloma patients described previously, older and younger individuals were matched according to five prognostic factors (four of them continuous variables). Matching according to categorical definitions of all of those variables would have been rather cumbersome; furthermore, for some "exposed" individuals,

EXHIBIT 1-3 Hypothetical example of frequency matching.

The cases ($n = 254$) range between 45 and 64 years of age. We want to select a control group of size $n = 508$ (2 controls per case) with the same age distribution according to 5-year age groups.

- Step 1: Examine the distribution of cases by the variable to be matched (5-year age groups in this example).
- Step 2: Use the same sampling fractions used in the cases for the selection of controls, which will result in the same age distribution in the control group.

Age distribution (years)	Number of cases by age	Percentage of cases by age	Number of controls by age	Percentage of controls by age
45–49	57	22.4	$508 \times 0.224 = 114$	22.4
50–54	72	28.3	$508 \times 0.283 = 144$	28.3
55–59	83	32.7	$508 \times 0.327 = 166$	32.7
60–64	42	16.5	$508 \times 0.165 = 84$	16.5
Total	254	100	$254 \times 2 = 508$	100

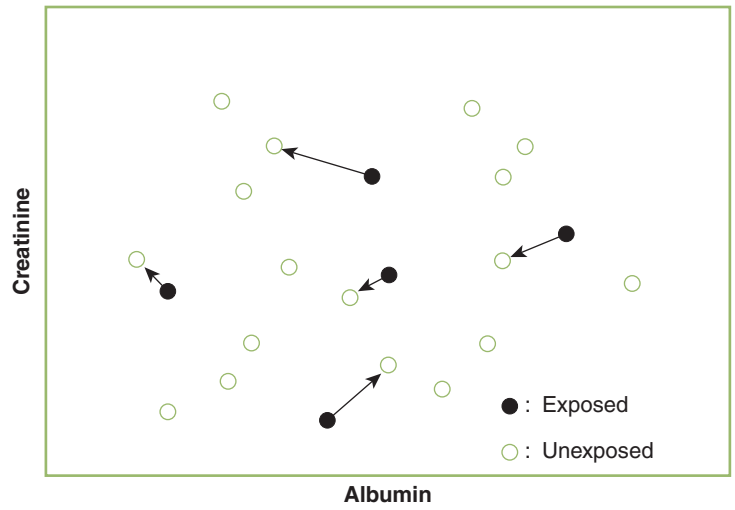


FIGURE 1-24 Matching according to minimal Euclidean distance measure method. Hypothetical example of a cohort study of survival after transplantation in multiple myeloma patients in which exposed individuals (e.g., older individuals) are matched to unexposed (younger) patients according to two prognostic factors: serum albumin and creatinine levels. For each case, the closest unexposed individual in the bidimensional space defined by the two matching variables is chosen as the control.

Data from Siegel DS, Desikan KR, Mehta J, et al. Age is not a prognostic variable with autotransplants for multiple myeloma. *Blood* 1999;93:51-54.[59]

it might have been difficult to find matches among "unexposed" persons. Thus, the authors of this study carried out matching using the minimum Euclidean distance measure method, as schematically illustrated in **FIGURE 1-24**. For the purpose of simplification, only two matching factors are considered in the figure. For each exposed case, the closest eligible person (e.g., unexposed patient) in this bidimensional space defined by albumin and creatinine levels is chosen as a control. In Siegel et al.'s study,[59] the authors used this method to match on more than two variables (levels of β_2-microglobulin, albumin, creatinine, and C-reactive protein and the presence/absence of chromosomal abnormalities), which would be impossible to represent in a diagram but still conceptually equivalent to the two-variable case illustrated in Figure 1-24. This method can also be used in the context of either case-based case-control studies or case-control studies within the cohort, as in the original application by Smith et al.,[60] representing a convenient and efficient alternative form of matching on multiple and/or continuous variables.

In situations in which there is a limited pool of cases, it might be desirable to select more than one matched control for each case to increase sample size and thus statistical power. For example, in the U.S. Air Force study of brain tumor risk factors cited previously,[43] each case was individually matched to four controls. In general, however, little statistical power is gained beyond four or five controls per case.[61]

Advantages and Disadvantages of Matching

Matching is a useful strategy to control for confounding, but it is far from being the only one. Chapter 7 is entirely devoted to describing alternative approaches that can be used at the analytical stage to address

the problem of confounding, namely, stratification and adjustment. Whether investigators choose to deal with confounding before data collection by matching during the recruitment phase of the study rather than by stratification or adjustment at the analysis stage depends on a number of considerations.

The *advantages* of matching include the following:

1. In addition to being easy to understand and describe, matching may be the only way to guarantee some degree of control for confounding in certain situations. This may be particularly important in studies in which a potentially strong confounding variable may produce such an imbalance in the composition of the study groups that adjustment is difficult or outright impossible. For example, in a case-control study of risk factors for prostate cancer, it would make sense to match controls to cases according to age; at the very least, given that most prostate cancer cases are in the older age brackets, the investigator should consider *restricting* the eligibility of the control group to a certain age range. (Restriction is a somewhat "loose" form of matching.) Otherwise, if controls were to be sampled from the general population, the age range could be so broad that there might not be enough overlap with the restricted age range of cases, particularly if the sample size were small (i.e., not enough older subjects in the control sample), to allow for adjustment.
2. If done according to strong confounders (variables that are related to both exposure and outcome; see Chapter 5), matching tends to increase the statistical power (efficiency) of the study.[62,63]
3. Matching (especially individual matching) is a logistically straightforward way to obtain a comparable control group when cases and controls are identified from a reference population for which there is no available sampling frame listing. For example, in a case-control study using cases of diabetes identified in an outpatient clinic, each case can be matched to the next nondiabetic person attending the clinic who has the same characteristics as the index case (e.g., similar age, gender).

Potential *disadvantages* of matching should also be considered. They include the following:

1. In certain situations, particularly when multiple variables are being matched for, it may be difficult or impossible to find a matched control or controls for a given case, particularly when sampling from a limited source population; when matching on multiple variables; or when the ratio of controls to cases is greater than 1:1. Furthermore, even if matched controls are available, the process of identifying them may be cumbersome and may add costs to the study's recruitment phase.
2. When matching is done, the association between the matching variable(s) and the outcome cannot be assessed because, after matching on a certain variable is carried out, the study groups (e.g., cases and controls) are set by design to be equal (or similar) with regard to this variable or set of variables.
3. It follows from number 2 that it is not possible to assess additive interaction in matched case-control studies between the matching variable(s) and the exposure(s) of interest. As discussed in Chapter 6, Section 6.4.2, the assessment of additive interaction in a case-control study relies on the formula for the joint expected relative risk (RR) of two variables A and Z, $RR_{A+Z+} = RR_{A+Z-} + RR_{A-Z+} - 1.0$ (using the odds ratios as estimates of the relative risks). Assuming that A is the matching variable, its independent effect (RR_{A+Z-}) cannot be estimated, as it has been set to 1.0 by design (see number 2). Therefore, this formula cannot be applied.
4. Matching implies some kind of tailoring of the selection of the study groups to make them as comparable as possible; this increased "internal validity" (comparability) may,

however, result in a reduced "external validity" (representativeness). For example, a control group that is made identical to the cases with regard to sociodemographic characteristics and other confounding variables may no longer constitute a representative sample of the reference population. In studies that examine the association between novel risk factors and disease, a secondary, yet important, objective may be to study the distribution or correlates of these factors in the reference population. If controls are matched to cases, it may be a complicated task to use the observed distributions in the control group to make inferences applicable to the population at large (complex weighted analyses taking into account the sampling fractions associated with the matching process would be required). On the other hand, if a random sample of the reference population is chosen (as is done in case-cohort studies), it would be appropriate to generalize the distributions of risk factors in the control group to the reference population. Obtaining these distributions (e.g., the prevalence of exposure in the population) is particularly important for the estimation of the population attributable risk (see Chapter 3, Section 3.4.2). In addition, as mentioned previously in this chapter, a control group that is selected as a random sample of the source population can be used as a comparison group for another case group selected from the same cohort or reference population.

5. Because when matching is done it cannot be "undone," it is important that the matching variables not be strongly correlated with the variable of interest; otherwise, the phenomenon of "overmatching" may ensue. For example, matching cases and controls on ethnic background may to a great extent make them very similar with regard to variables of interest related to socioeconomic status. For further discussion of this topic and additional examples, see Chapter 5, Section 5.5.3.
6. Finally, no statistical power is gained if the matching variables are weak confounders. If the matching variables are weakly related to exposure, even if they are related to the outcome, the gain in efficiency may be very small. Moreover, if the matching variables are weakly or not related to the outcome of interest, matching can result in a loss of power.[62,63]

When matching is conducted according to categorical definitions of continuous or ordinal variables, residual differences between cases and controls may remain (*residual confounding*) (see Chapter 5, Section 5.5.4, and Chapter 7, Section 7.6). In these situations, it may be necessary to adjust for the variable in question in the analyses to eliminate variation within the matching categories. For example, in a study on the relationship between cytomegalovirus antibodies in serum samples collected in 1974 (and retrospectively analyzed) and the presence of carotid atherosclerosis measured by B-mode ultrasound of the carotid arteries about 15 years later (1987–1989),[64] 150 controls (selected among individuals with normal carotid arteries) were frequency matched to 150 carotid atherosclerosis cases according to gender and two relatively broad age groups (45–54 years and 55–64 years). Thus, by design, both the case and control groups had an identical number of individuals in all four gender–age groups; however, cases in this study were 58.2 years of age on average, whereas the average age in controls was 56.2 years. Therefore, even though the study groups were matched on two age categories, the residual age difference prompted the authors to adjust for age *as a continuous variable* in the multivariate logistic regression analyses (see Chapter 7, Section 7.4.3).

The same residual differences may remain even in individually matched studies if the matching categories are broadly categorized. The reason for this phenomenon is illustrated in **FIGURE 1-25**. Even though the cases and controls are equally represented in both age groups, within each age group, cases tend to be older, thus resulting in an overall difference.

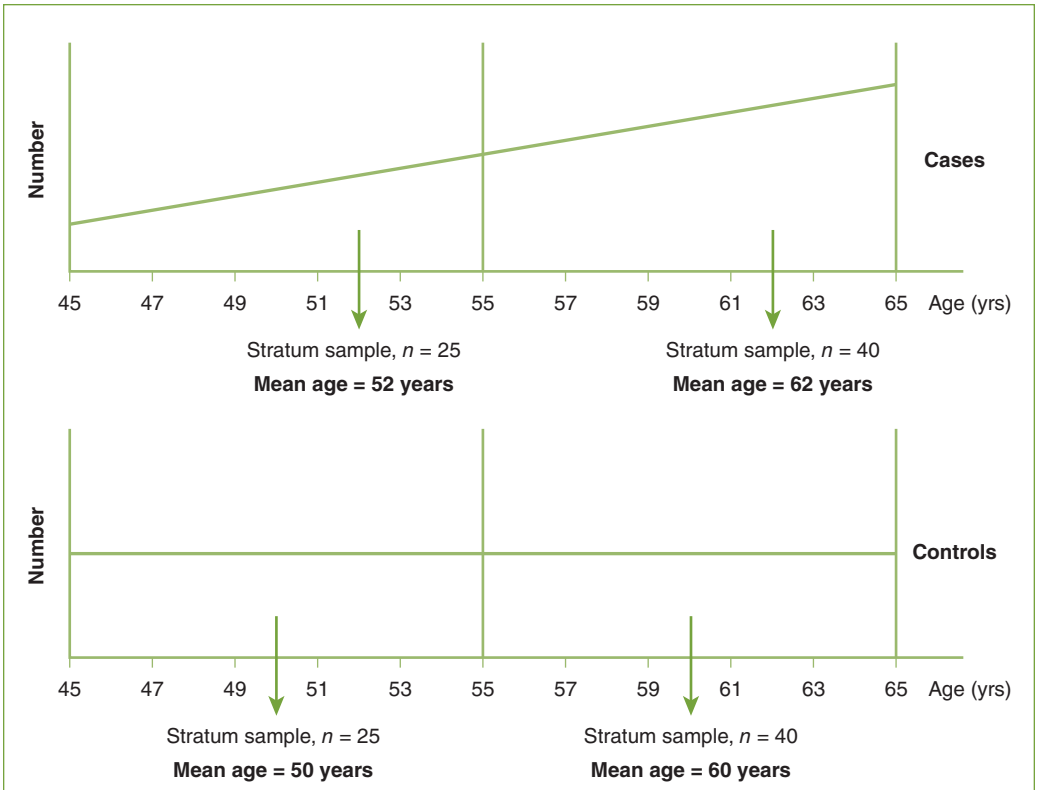


FIGURE 1-25 Schematic representation of a hypothetical situation where matching cases and controls according to broad age categories (45–54 years and 55–64 years) results in a residual mean age difference (residual confounding). Within age categories, cases are skewed toward older ages, but the age distribution of controls is flat. As a result, the mean age within each age group is higher in cases than in controls, resulting in an overall mean age of 58.2 years in cases and 56.2 years in controls.

In summary, investigators should always consider carefully whether to match. Unlike *post hoc* means to control for confounding (e.g., stratification and adjustment), matching is irreversible after implemented. Although it may be the strategy of choice in studies with limited sample size and a clear-cut set of objectives, it should be avoided in most situations in which a reasonable overlap on potential confounding variables is expected to exist (thus allowing adjustment). If matching is used, the investigators must keep in mind that ignoring the matching during the analysis of the data can lead to bias[65] and that special statistical techniques for analyses of matched data are available (see Chapter 7, Section 7.4.6).

References

1. Porta M. *A Dictionary of Epidemiology*. 6th ed. New York: Oxford University Press; 2014.
2. Koepsell TD, Weiss NS. *Epidemiologic Methods*. New York: Oxford University Press; 2003.
3. Gordis L. *Epidemiology*. 5th ed. Philadelphia: Elsevier Saunders; 2014.
4. Strong JP, McGill HC Jr. The pediatric aspects of atherosclerosis. *J Atheroscler Res*. 1969;9:251-265.

5. Frost WH. The age-selection of tuberculosis mortality in successive decades. *Am J Hyg.* 1939;30:91-96.
6. López-Abente G, Pollán M, Vergara A, et al. Age-period-cohort modeling of colorectal cancer incidence and mortality in Spain. *Cancer Epidemiol Biomar*. 1997;6:999-1005.
7. Dhillon PK, Yeole BB, Dikshit R, Kurkure AP, Bray F. Trends in breast, ovarian and cervical cancer incidence in Mumbai, India over a 30-year period, 1976-2005: an age-period-cohort analysis. *Br J Cancer.* 2011;105:723-730.
8. Newschaffer CJ, Bush TL, Hale WE. Aging and total cholesterol levels: cohort, period, and survivorship effects. *Am J Epidemiol*. 1992;136:23-34.
9. Keys A. *Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease*. Cambridge, MA: Harvard University Press; 1980.
10. Morgenstern H. Ecologic studies in epidemiology: concepts, principles, and methods. *Ann Rev Public Health*. 1995;16:61-81.
11. Diez-Roux AV, Nieto FJ, Muntaner C, et al. Neighborhood environments and coronary heart disease: a multilevel analysis. *Am J Epidemiol*. 1997;146:48-63.
12. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *Am J Public Health*. 1998;88:216-222.
13. Durkheim E. *Suicide: A Study in Sociology*. New York: Free Press; 1951.
14. MacMahon B, Pugh TF. *Epidemiology: Principles and Methods*. Boston: Little, Brown and Co.; 1970.
15. Piantadosi S, Byar DP, Green SB. The ecological fallacy. *Am J Epidemiol*. 1988;127:893-904.
16. Elliott P. Design and analysis of multicentre epidemiological studies: the INTERSALT Study. In: Marmot M, Elliott P, eds. *Coronary Heart Disease Epidemiology: From Aetiology to Public Health*. Oxford: Oxford University Press; 1992:166-178.
17. Wynder EL, Cohen LA, Muscat JE, Winters B, Dwyer JT, Blackburn G. Breast cancer: weighing the evidence for a promoting role of dietary fat. *J Natl Cancer Inst*. 1997;89:766-775.
18. Goldberger J, Wheeler GA, Sydenstricker E. A study of the relation of family income and other economic factors to pellagra incidence in seven cotton-mill villages of South Carolina in 1916. *Public Health Rep*. 1920;35:2673-2714.
19. Koopman JS, Longini IM Jr. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am J Public Health*. 1994;84:836-842.
20. Lobato JCP, Kale PL, Velarde LGC, Szklo M, Costa AJ. Correlation between mean body mass index in the population and prevalence of obesity in Brazilian capitals: empirical evidence for a population-based approach of obesity. *BMC Public Health.* 2015;15:322-327.
21. Rose G, Day S. The population mean predicts the number of deviant individuals. *Br Med J.* 1990;301:1031-1034.
22. Breslow NE, Day NE. Statistical methods in cancer research: Volume I: the analysis of case-control studies. *IARC Sci Publ*. 1980.
23. Breslow NE, Day NE. Statistical methods in cancer research: Volume II: the design and analysis of cohort studies. *IARC Sci Publ*. 1987.
24. Samet JM, Munoz A. Perspective: cohort studies. *Epidemiol Rev*. 1998;20:135-136.
25. Schlesselman J. *Case Control Studies: Design, Conduct, Analysis*. New York: Oxford University Press; 1982.
26. Armenian HK, Lilienfeld DE. Applications of the case-control method: overview and historical perspective. *Epidemiol Rev*. 1994;16:1-5.
27. Dawber TR. *The Framingham Study: The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press; 1980.
28. Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR Jr. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol*. 1987;126:310-318.
29. Stampfer MJ, Willett WC, Colditz GA, Rosner B, Speizer FE, Hennekens CH. A prospective study of postmenopausal estrogen therapy and coronary heart disease. *N Engl J Med*. 1985;313:1044-1049.
30. Ascherio A, Rimm EB, Stampfer MJ, Giovannucci EL, Willett WC. Dietary intake of marine n-3 fatty acids, fish intake, and the risk of coronary disease among men. *N Engl J Med*. 1995;332:977-982.
31. Garfinkel L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. *Natl Cancer Inst Monogr*. 1985;67:49-52.
32. Vandenbroucke JP. Prospective or retrospective: What's in a name? *Br Med J.* 1991;302:249-250.
33. Nieto FJ, Szklo M, Comstock GW. Childhood weight and growth rate as predictors of adult mortality. *Am J Epidemiol*. 1992;136:201-213.
34. Tynes T, Andersen A, Langmark F. Incidence of cancer in Norwegian workers potentially exposed to electromagnetic fields. *Am J Epidemiol*. 1992;136:81-88.
35. Cornfield J. A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*. 1951;11:1269-1275.
36. Fisher JL, Pettersson D, Palmisano S, et al. Loud noise exposure and acoustic neuroma. *Am J Epidemiol.* 2014;180:58-67.
37. Christensen K, Olsen J, Norgaard-Pedersen B, et al. Oral clefts, transforming growth factor alpha gene variants, and maternal smoking: a population-based case-control study in Denmark, 1991-1994. *Am J Epidemiol*. 1999;149:248-255.

38. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol*. 1990;131:169-176.
39. Thomas D. New techniques for the analysis of cohort studies. *Epidemiol Rev*. 1998;20:122-134.
40. Kleinbaum D, Kupper LL, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, CA: Lifetime Learning Publications; 1982.
41. Checkoway H, Pearce NE, Crawford-Brown, DJ. *Research Methods in Occupational Epidemiology.* New York: Oxford University Press; 1989.
42. Dekker JM, Crow RS, Folsom AR, et al. Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the ARIC Study: atherosclerosis risk in communities. *Circulation*. 2000;102:1239-1244.
43. Grayson JK. Radiation exposure, socioeconomic status, and brain tumor risk in the US Air Force: a nested case-control study. *Am J Epidemiol*. 1996;143:480-486.
44. Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet*. 2005;365:475-481.
45. Nieto FJ, Folsom AR, Sorlie PD, Grayston JT, Wang SP, Chambless LE. *Chlamydia pneumoniae* infection and incident coronary heart disease: the Atherosclerosis Risk in Communities Study. *Am J Epidemiol*. 1999;150:149-156.
46. Dallal CM, Tice JA, Buist DSM, et al. Estrogen metabolism and breast cancer risk among postmenopausal women: a case-cohort study with B-FIT. *Carcinogenesis*. 2014;35:346-355.
47. Black DM, Reiss TF, Nevitt MC, Cauley J, Karpf D, Cummings SR. Design of the Fracture Intervention Trial. *Osteoporos Int*. 1993;3(suppl. 3):S29-S39.
48. Hickman TB, Briefel RR, Carroll MD, et al. Distributions and trends of serum lipid levels among United States children and adolescents ages 4-19 years: data from the Third National Health and Nutrition Examination Survey. *Prev Med*. 1998;27:879-890.
49. Flegal KM, Carroll MD, Ogden CL, Curtin LR. Prevalence and trends in obesity among US adults, 1999-2008. *J Am Med Assoc*. 2010;303:235-241.
50. Lister SM, Jorm LR. Parental smoking and respiratory illnesses in Australian children aged 0-4 years: ABS 1989-90 National Health Survey results. *Aust N Z J Public Health*. 1998;22:781-786.
51. Howard G, Burke GL, Szklo M, et al. Active and passive smoking are associated with increased carotid wall thickness: the Atherosclerosis Risk in Communities Study. *Arch Intern Med*. 1994;154:1277-1282.
52. Howard G, Wagenknecht LE, Burke GL, et al. Cigarette smoking and progression of atherosclerosis: the Atherosclerosis Risk in Communities (ARIC) Study. *J Am Med Assoc*. 1998;279:119-124.
53. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133:144-153.
54. Vlak MH, Rinkel GJE, Greebe P, van der Bom JG, Algra A. Trigger factors and their attributable risk for rupture of intracranial aneurysms: a case-crossover study. *Stroke*. 2011;42:1878-1882.
55. Pereira G, Cook A, De Vos AJ, Holman CD. A case-crossover analysis of traffic-related air pollution and emergency department presentations for asthma in Perth, Western Australia. *Med J Aust*. 2010;193:511-514.
56. Rich KE, Petkau J, Vedal S, Brauer M. A case-crossover analysis of particulate air pollution and cardiac arrhythmia in patients with implantable cardioverter defibrillators. *Inhal Toxicol*. 2004;16:363-372.
57. Valent F, Brusaferro S, Barbone F. A case-crossover study of sleep and childhood injury. *Pediatrics*. 2001;107:E23.
58. Ridker PM, Cushman M, Stampfer MJ, Tracy RP, Hennekens CH. Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *N Engl J Med*. 1997;336:973-979.
59. Siegel DS, Desikan KR, Mehta J, et al. Age is not a prognostic variable with autotransplants for multiple myeloma. *Blood*. 1999;93:51-54.
60. Smith AH, Kark JD, Cassel JC, Spears GF. Analysis of prospective epidemiologic studies by minimum distance case-control matching. *Am J Epidemiol*. 1977;105:567-574.
61. Breslow N. Case-control studies. In: Ahrens W, Pigeot I, eds. *Handbook of Epidemiology*. Heidelberg, Germany: Springer-Verlag Berlin; 2005:287-319.
62. Samuels M. Matching and design efficiency in epidemiological studies. *Biometrika*. 1981;68:577-588.
63. Thompson WD, Kelsey JL, Walter SD. Cost and efficiency in the choice of matched and unmatched case-control study designs. *Am J Epidemiol*. 1982;116:840-851.
64. Nieto FJ, Adam E, Sorlie P, et al. Cohort study of cytomegalovirus infection as a risk factor for carotid intimal-medial thickening, a measure of subclinical atherosclerosis. *Circulation*. 1996;94:922-927.
65. Breslow N. Design and analysis of case-control studies. *Annu Rev Public Health*. 1982;3:29-54.

Exercises

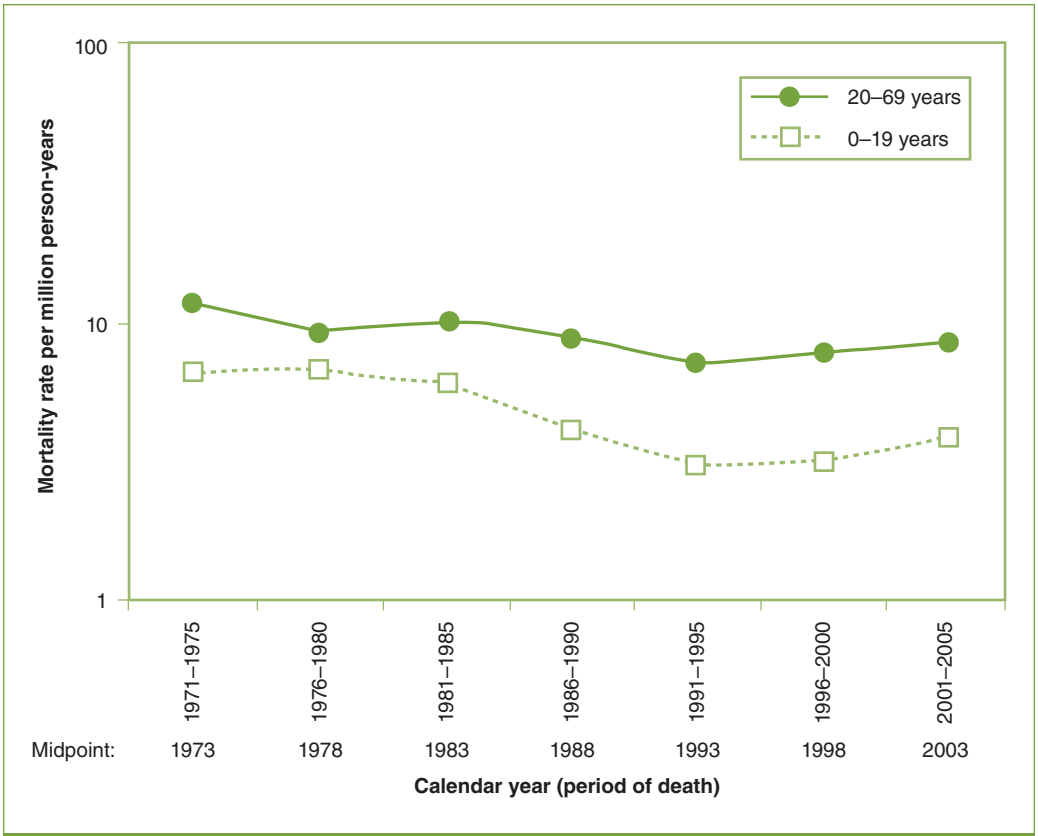
1. The table shows a series of cross-sectional incidence rates of cancer Y per 100,000 by age and calendar year.

Age	Calendar year							
	1950	1955	1960	1965	1970	1975	1980	1985
20–24	10	15	22	30	33	37	41	44
25–29	8	17	20	24	29	38	40	43
30–34	5	12	22	25	28	35	42	45
35–39	3	12	15	26	30	32	39	42
40–44	2	10	17	19	28	32	39	42
45–49	2	12	15	18	21	33	40	42
50–54	2	10	16	20	25	32	42	44
55–59	2	15	17	19	22	27	43	44

   a. After observing the incidence rates by age at any given year, it is concluded that aging is not related to an increase in the incidence of Y and may even be related to a decrease in the incidence. Do you agree with this observation? Justify your answer.
   b. What are the purposes of examining birth cohort vis-à-vis cross-sectional rates?

2. Chang et al. carried out a birth cohort analysis of epilepsy mortality in Taiwan from 1971 through 2005.[*] The following figure shows the epilepsy mortality rates per million person-years by calendar year for two of the three age groups examined by the authors.
   Assume that year of death for people dying in each calendar year category is the midpoint for that period (e.g., for the calendar year category of 1971–1975, the assumed year of death is 1973; for the category 1976–1980, it is 1978, and so on). The same should be assumed for the age groupings (e.g., for ages 0–19, assume that the age of death is 10 years, for the age group 20–69 years, it is 45 years, and so on).
   a. Is the use of the midpoint value a reasonable approach to analyzing birth cohorts?
   b. To which cohort do individuals dying in 1973 at ages 0–19 years belong?

[*]Chang Y, Li C, Tung T, Tsai J, Lu T. Age-period-cohort analysis of mortality from epilepsy in Taiwan, 1971-2005. *Seizure*. 2011;20:240-243.



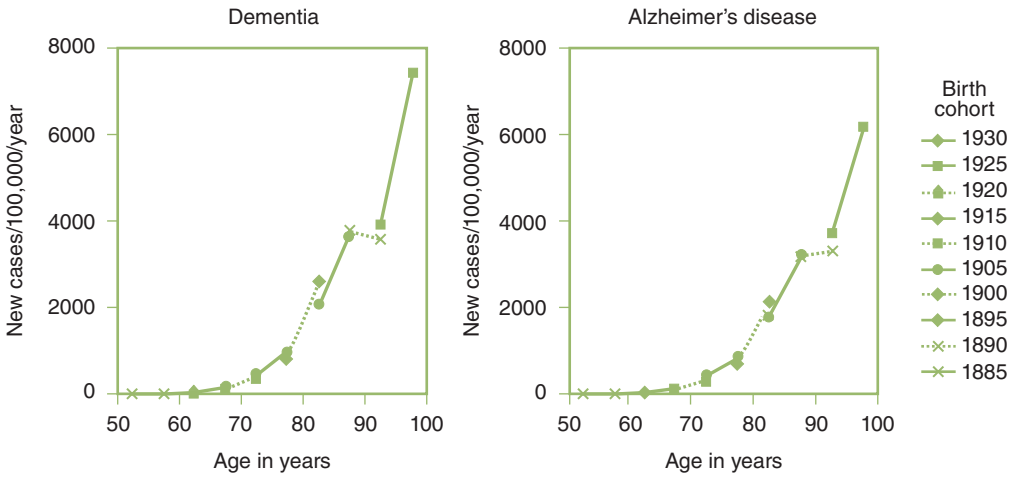
Secular trend in the age-specific epilepsy mortality rate in Taiwan, 1971–2005.

Reprinted from Chang Y, Li C, Tung T, Tsai J, Lu T. Age-period-cohort analysis of mortality from epilepsy in Taiwan, 1971-2005. *Seizure*. 2011;20:240-243. With permission from Elsevier.

c. Is there a birth cohort effect?
d. Is there an age effect?

3. The following figure shows the incidence of dementia and Alzheimer's disease per 100,000 person-years by age and birth cohort, both sexes combined, in residents of Rochester, Minnesota, from 1975 through 1984.[†]
   a. Are age effects apparent in the figure? Justify your answer.
   b. Are cohort effects apparent in the figure? Justify your answer.
   c. From your answer to Exercise 3b, would you expect age patterns to be similar in cross-sectional and cohort analyses? Justify your answer.

[†]Rocca WA, Cha RH, Waring SC, Kokmen E. Incidence of dementia and Alzheimer's disease: a reanalysis of data from Rochester, Minnesota, 1975-1984. *Am J Epidemiol*. 1998;148:51-62.



Five-year average incidence rates of dementia (new cases per 100,000 person-years) by age and birth cohort, both sexes combined, Rochester, Minnesota. For each birth cohort, two points are represented corresponding to the incidence in 1975 to 1979 and 1980 to 1984, respectively.

Reprinted with permission from Rocca WA, Cha RH, Waring SC, Kokmen E. Incidence of dementia and Alzheimer's disease: a reanalysis of data from Rochester, Minnesota, 1975-1984. *Am J Epidemiol.* 1998;148:51-62. By permission of Oxford University Press.

4. A case-control study is conducted within a well-defined cohort. The reason for this is that expensive additional data collection is needed and the budget is not sufficient to obtain these data from all cohort participants.
   a. What type of case-control study within this cohort would be ideal to study multiple outcomes, and why is the alternative case-control design not recommended?
   b. In this cohort study, prevalent cases were not excluded at baseline, and, thus, the investigators chose to use baseline data to examine associations between suspected risk factors and prevalent disease. What type of approach is this, and what are its main advantages and disadvantages?
5. In planning an individually matched case-based case-control study to test the hypothesis that air pollution (measured by individually placed monitors) is related to a certain type of respiratory cancer, the investigators decide to match cases and controls on age, gender, ethnic background, and smoking (yes or no).
   a. In addition to general logistical difficulties usually associated with matching, what is the main undesirable consequence that may result from matching cases and controls in this study?
   b. Because the disease of interest is rare, the investigators decide to individually match 10 controls for each case. Is this a reasonable strategy considering the additional costs involved and the tight budget to conduct this study?
6. The Multi-Ethnic Study of Atherosclerosis (MESA) has been conducted in six regions in the United States: Baltimore City and Baltimore County, Maryland; Chicago, Illinois; Forsyth

County, North Carolina; Los Angeles County, California; New York, New York; and St. Paul, Minnesota.[‡] The objectives of MESA are (1) to determine characteristics related to progression of subclinical cardiovascular disease (CVD) to clinical CVD; (2) to determine characteristics related to progression of subclinical CVD itself; (3) to assess ethnic, age, and sex differences in subclinical disease prevalence, risk of progression, and rates of clinical CVD; (4) to determine relations of newly identified risk factors with subclinical disease and their incremental predictive value over established risk factors; and (5) to develop methods, suitable for application in future screening and intervention studies, for characterizing risk among asymptomatic persons.

Study participants included 6500 men and women, in equal numbers, who were aged 45 to 84 years and free of clinical CVD at baseline. Four racial/ethnic groups from six U.S. communities were included. Approximately 38% of the cohort is white; 28%, African American; 23%, Hispanic; and 11%, Asian, predominantly of Chinese descent. The first examination, which began in July 2000 and was conducted over a 24-month period, was designed to be the most comprehensive. The second (from July 2002 to January 2004) and third (from January 2004 to July 2005) examinations, conducted over 18 months each, included repetitions of selected baseline measurements and new measures that could not be included at baseline. The fourth examination (from July 2005 to July 2007), conducted over a 24-month period, included repetition of selected measures to be studied for temporal trends.

a. What study design was used in the MESA?
b. If the investigators wished to analyze the associations of risk factors with a given outcome using only data from the first (baseline) exam, what type of study would they be conducting?

At the same time the study samples were chosen in each center, a random subsample of 1000 individuals was also selected. Serum samples of the whole cohort (including the subsample) were frozen/stored for future use. Some of the analyses done in the MESA were based on comparing cases of myocardial infarction that occurred during follow-up with the subsample. In these analyses, analytes measured in thawed serum samples were compared between cases and the subsample.

c. What type of study/analysis is this?
d. What are the main advantages of this study design?

7. A population-based case-control analysis was conducted to evaluate whether dietary patterns influence the risk of a rare disorder, classic Hodgkin lymphoma (cHL) in younger or older adults.[§] Cases of incident cHL were recruited from the greater Boston metropolitan area of Massachusetts and the state of Connecticut from August 1, 1997, to December 31, 2000. Eligible patients were aged 15 to 79 years, living within the target geographic area and without human immunodeficiency virus (HIV) infection at diagnosis. Cases were identified by using the rapid case ascertainment systems of Harvard and Yale universities with additional support from the Massachusetts and Connecticut state tumor registries. Six hundred seventy-seven eligible cases were invited to participate in the study, and 84% ($n = 567$) consented. Certain data used in this study were obtained from the Connecticut

[‡]Bild DE, Bluemke DA, Burke GL, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol.* 2002;151:871-881.

[§]Epstein MM, Chang ET, Zhang Y, et al. Dietary pattern and risk of Hodgkin lymphoma in a population-based case-control study. *Am J Epidemiol.* 2015;182:405-416.

Tumor Registry in the Connecticut Department of Public Health. Population-based controls without a history of cHL were frequency matched to cases by age (within 5 years), sex, and state of residence (Massachusetts or Connecticut). In greater Boston, controls were identified through the "Town Books," annual records documenting all citizens aged ≥ 17 years, which are 90% complete. Of 720 invited controls in Massachusetts, 51% ($n = 367$) consented. In Connecticut, 450 eligible controls aged 18 to 65 years were identified by random-digit dialing, and 61% ($n = 276$) consented. Of 69 eligible controls in Connecticut aged 66 to 79 years identified through the Health Care Financing Administration (Medicare), 52% ($n = 36$) consented to participate.

a. What type of study design is this?
b. What is a common indication for using this type of design?
c. What is the study base of this study?