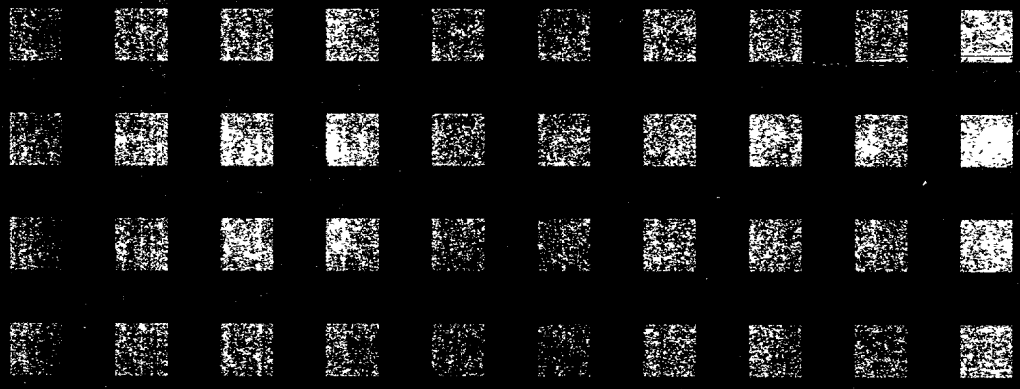
SECOND EDITION

MODERN EPIDEMIOLOGY



Kenneth J. Rothman

Sander Greenland

LIPPINCOTT WILLIAMS & WILKINS

Modern Epidemiology

Second Edition

Modern Epidemiology

Second Edition

Kenneth J. Rothman
Professor of Epidemiology
Boston University School of Public Health
Boston University Medical School
Boston, Massachusetts

Sander Greenland
Professor of Epidemiology
Department of Epidemiology
UCLA School of Public Health
Los Angeles, California

with 15 contributors

LIPPINCOTT WILLIAMS & WILKINS
A **Wolters Kluwer** Company
Philadelphia • Baltimore • New York • London
Buenos Aires • Hong Kong • Sydney • Tokyo

Acquisitions Editor: Richard Winters
Developmental Editor: Erin O'Connor
Manufacturing Manager: Dennis Teston
Cover Designer: Betty Sokol
Indexer: Maria Coughlin
Compositor: Lippincott Williams & Wilkins Desktop Division
Printer: Maple Press

Printed in the United States of America

9 8 7 6 5 4

Library of Congress Cataloging-in-Publication Data
Rothman, Kenneth J.
Modern epidemiology / by Kenneth J. Rothman and Sander Greenland : [additional contributions by James W. Buehler . et al.]. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-316-75780-2
1. Epidemiology. 2. Epidemiology—Statistical methods.
I. Greenland, Sander, 1951– II. Title.
[DNLM: 1. Epidemiology. 2. Epidemiologic Methods. WA 105 R846m 1998]
RA651.R63 1998
614.4–dc21
DNLM/DLC
For Library of Congress

Contents

PART III. DATA ANALYSIS

PART IV. SPECIAL TOPICS

While the hypotheses are often stated in qualitative terms, the testing of hypotheses is predicated on measurement. The role of measurement is central to all empirical sciences, not only epidemiology, no matter how qualitative the theories under evaluation. For example, qualitatively stated hypotheses about evolution, the formation of the earth, the effect of gravity on light, or the method by which birds find their way during migration are all tested by measurements of the phenomena that relate to the hypotheses.

The importance of measurement has been reflected in the evolution of epidemiologic understanding. Physicians throughout recorded history, from Hippocrates to Sydenham, have considered the causes of disease. Unfortunately, they seldom did more than consider. It was only when scientists began to measure the occurrence of disease rather than merely reflect on what may have caused disease that scientific knowledge about causation made impressive strides.

A central task in epidemiologic research is to quantify the occurrence of disease in populations. This chapter discusses four basic measures of disease occurrence. *Incidence times* are simply the times at which new disease occurs among population members. *Incidence rate* measures the occurrence of new disease per unit of person-time. *Incidence proportion* measures the proportion of people who develop new disease during a specified period of time. *Prevalence,* a measure of status rather than of newly occurring disease, measures the proportion of people who have disease at a specific time.

INCIDENCE TIME

In the attempt to measure the frequency of disease occurrence in a population, it is insufficient merely to record the number of people or even the proportion of the population that is affected. It is also necessary to take into account the time elapsed before disease occurs, as well as the period of time during which events are counted. Consider the frequency of death. Since all people are eventually affected, the time from birth to death becomes the determining factor in the rate of occurrence of death. If, on average, death comes earlier to the members of one population than to members of another population, it is natural to say that the first population has a higher death rate than the second. Time is the factor that differentiates between the two situations shown in Fig. 3-1.

In an epidemiologic study, we may measure the time of events in an individual's life relative to any one of several reference events. Using age, for example, the reference event is birth, but we might instead use the start of a treatment or the start of an exposure as the reference event. The reference event may be unique to each person, as it is with birth, or it may be identical for all persons, as with calendar time. The time of the reference event determines the time origin or *zero time* for measuring time of events.

Given an outcome event or "incident" of interest, a person's *incidence time* for this outcome is defined as the time span from zero time to the time at which the event occurs, if it occurs. A man who experienced his first myocardial infarction in 1990 at age 50 has an incidence time of 1990 in (Western) calendar time and an incidence time of 50 in age time. A person's incidence time is undefined if that person never experiences the event. There is a useful convention that classifies such a person as having an unspecified incidence time that is known to exceed the last time the person could have experienced the event. Under this convention, a woman who had a hysterectomy in 1990 without ever having had endometrial cancer is classified as having an endometrial cancer incidence time greater than 1990.

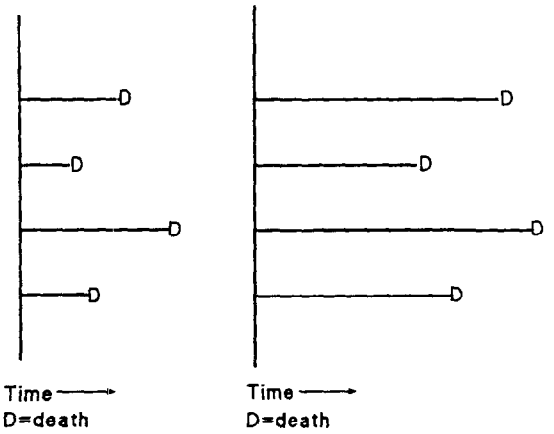


FIG. 3-1. Two different patterns of disease occurence.

INCIDENCE RATE

Epidemiologists often study events that are not inevitable or that may not occur during the period of observation. In such situations, the set of incidence times for a specific event in a population will not all be defined or observed, and another incidence measure must be sought. Ideally, such a measure would take into account the number of individuals in a population that become ill, as well as the length of time contributed by all persons during the period they were in the population and events were counted.

Person-Time

Consider any population at risk and a risk period over which we want to measure incidence in this population. Every member of the population experiences a specific amount of time in the population over the risk period; the sum of these times over all population members is called the total *person-time* at risk over the period. Person-time should be distinguished from clock time in that it is a summation of time that occurs simultaneously for many people, whereas clock time is not. Person-time represents the observational experience in which disease onsets can be observed. The number of new cases of disease (incident number) divided by the person-time is the incidence rate of the population over the period:

$$\text{Incidence rate} = \frac{\text{No. disease onsets}}{\sum_{\text{persons}} \text{time spent in population}}$$

When the risk period is of fixed length Δt, the total person-time at risk over the period is equal to the average size of the population over the period, $\bar{N}$, times the length of the period, Δt. If we denote the incident number by A, it follows that the person-time rate equals $A/(\bar{N}\cdot\Delta t)$. This formulation makes clear that the incidence rate has units of inverse time (per year, per month, per day, etc.). The units attached to an incidence rate can be written as year^{-1}, month^{-1}, or day^{-1}.

It is an important principle that the only events eligible to be counted in the numerator of an incidence rate are those that occur to persons who are contributing time to the denominator of the incidence rate at the time that the disease onset occurs. Likewise, only time contributed by persons eligible to be counted in the numerator if they suffer an event should be counted in the denominator. The time contributed by each person to the denominator is sometimes known as the "time at risk," that is, time at risk of an event's occurring. Analogously, the people who contribute time to the denominator of an incidence rate are referred to as the "population at risk."

Incidence rates often include only the first occurrence of disease onset as an eligible event for the numerator of the rate. For the many diseases that are irreversible states, such as diabetes, multiple sclerosis, cirrhosis, or death, there is at most only one onset that a person can experience. For some diseases that do recur, such as rhinitis, we may simply wish to measure the incidence of "first" occurrence, even though the disease can occur repeatedly. For other diseases, such as cancer or heart disease, the first occurrence is often of greater interest for study than subsequent occurrences in the same individual. Therefore, it is typical that the events in the numerator of an incidence rate correspond to the first occurrence of a particular disease, even in those instances in which it is possible for an individual to have more than one occurrence. In this book, we will assume we are dealing with first occurrences, except where stated otherwise.

When the events tallied in the numerator of an incidence rate are first occurrences of disease, then the time contributed by each individual in whom the disease develops should terminate with the onset of disease. The reason is that the individual is no longer eligible to experience the event (the first occurrence can only occur once per individual), so there is no more information to obtain from continued observation of that individual. Thus, each individual who experiences the event should contribute time to the denominator up until the occurrence of the event, but not afterward. Furthermore, for the study of first occurrences, the number of disease onsets in the numerator of the incidence rate is also a count of people experiencing the event, since only one event can occur per person.

An epidemiologist who wishes to study both first and subsequent occurrences of disease may decide not to distinguish between first and later occurrences and simply count all the events that occur among the population under observation. If so, then the time accumulated in the denominator of the rate would not cease with the occurrence of the event, since an additional event might occur in the same individual. Usually, however, there is enough of a biologic distinction between first and subsequent occurrences to warrant measuring them separately. One approach is to define the "population at risk" differently for each occurrence of the event: The population at risk for the first event would consist of individuals who have not experienced the disease before; the population at risk for the second event or first recurrence would be limited to those who have experienced the event once and only once, etc. A given individual should contribute time to the denominator of the incidence rate for first events only until the time that the disease first occurs. At that point, the individual should cease contributing time to the denominator of that rate and should now begin to contribute time to the denominator of the rate measuring the second occurrence. If and when there is a second event, the individual should stop contributing time to the rate measuring the second occurrence and begin contributing to the denominator of the rate measuring the third occurrence, and so forth.

Closed and Open Populations

Conceptually, we can imagine the person-time experience of two distinct types of populations, the *closed population* and the *open population*. A closed population adds no

new members over time and loses members only to death, whereas an open population may gain members over time, through immigration or birth, or lose members who are still alive through emigration. (Some demographers and ecologists use a broader definition of closed population in which births, but not immigration or emigration, are allowed.) Suppose we graph the survival experience of a closed population of 1000 people. Since death eventually claims everyone, after a period of sufficient time the original 1000 will have dwindled to zero. A graph of the size of the population with time might approximate that in Fig. 3-2.

The curve slopes downward because as the 1000 individuals in the population die, the population at risk of death is reduced. The population is closed in the sense that we consider the fate of only the 1000 individuals present at time zero. The person-time experience of these 1000 individuals is represented by the area under the curve in the diagram. As each individual dies, the curve notches downward; that individual no longer contributes to the person-time denominator of the death (mortality) rate. Each individual's contribution is exactly equal to the length of time that individual is followed from start to finish; in this example, since the entire population is followed until death, the finish is the individual's death. In other instances, the contribution to the person-time experience would continue until either the onset of disease or some arbitrary cutoff time for observation, whichever came sooner.

Suppose we added up the total person-time experience of this closed population of 1000 and obtained a total of 75,000 person-years. The death rate would be (1000/75,000) $\times$ year^{-1}, since the 75,000 person-years represent the experience of all 1000 people until their deaths. Furthermore, if time is measured from start of follow-up, the average death time in this closed population would be 75,000 person-years/1000 persons = 75 years, which is the inverse of the death rate.

A closed population facing a constant death rate would decline in size exponentially (which is what is meant by the term "exponential decay"). In practice, however, death rates for a closed population change with time, since the population is aging as time progresses. Consequently, the decay curve of a closed human population is never exponen-

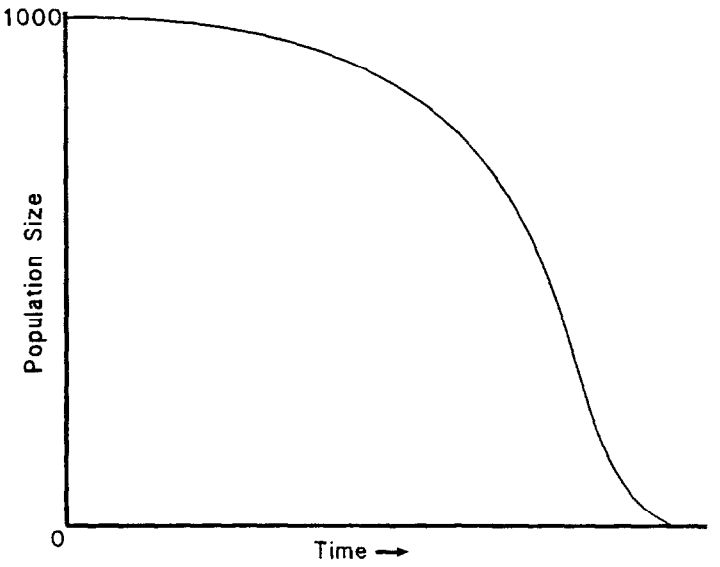


FIG. 3-2. Size of a closed population of 1000 people, by time.

tial. *Life-table* methodology is a procedure by which the death rate (o ... e) o ... closed population is evaluated within successive small age or time in ... that t ... age or time dependence of mortality can be elucidated. Even with such ... portant to distinguish an age-related effects from those related to oth ..., sir ... each individual's age increases directly with an increase along any o ... is. F ... example, a person's age increases with increasing duration of empl ... easi ... calendar time, and increasing time from start of follow-up.

An open population differs from a closed population in that indivi ... butic ... need not begin at the same time. Instead, the population at risk is ope ... ev ... embe ... who become eligible with passing time. People can enter a populatio ... alenc ... time through various mechanisms. Some are born into it; others migrat ... a pc ... ulation of people of a specific age, individuals can become eligible t ... popu ... tion by aging into it. Similarly, individuals can exit from the person- ... vation ... experience defining a given incidence rate by dying, aging out of a ... grou ..., emigrating, or becoming diseased (the latter method of exiting applies ... if ... st bou ... of a disease are being studied).

Steady State

If the number of people entering a population is balanced by the ... ting t ... population in any period of time within levels of age, sex, and other de ... of ris ..., the population is said to be *stationary,* or in a *steady state.* Steady stat ... pr ... ty th ... can occur only in open populations, not closed populations. It is, ho ... sible ... have a population in steady state in which no immigration or emigratio ... ng; th ... situation would require that births perfectly balance deaths in the pop ... e gra ... of the size of an open population in steady state is simply a horizont ... ople a ... continually entering and leaving the population in a way that might ... ned ... shown in Fig. 3-3.

In the diagram, the symbol > represents a person entering the popu ..., the se ... ment represents their person-time experience, and the termination of a ... nt re ... sents the end of their experience. A terminal D indicates that the ex ... ded b ... cause of disease onset, and a terminal C indicates that the experien ... r oth ... reasons. In theory, any time interval will provide a good estimate of th ... rate ... a stationary population. The value of incidence will be the ratio of th ... of cas ... of disease onset, indicated by D, to the area depicting the product of ... at ... × tim ...

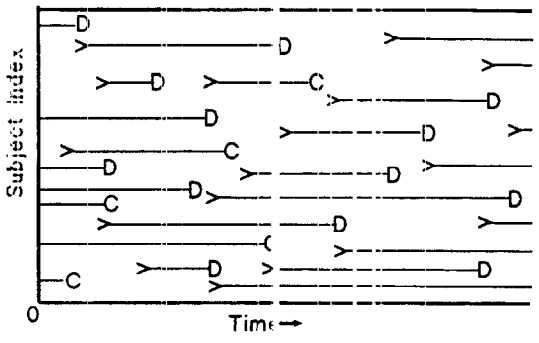


FIG. 3-3. Compos ... or ... n op ... population in appro ... ma ... s ... cy sta ..., by time. > indica ... s ... nt ... into ... e population, D indic ... tes ... s ... e ons ..., and C indicates exit ... or the ... opulat ... without disease.

Because this ratio is equivalent to the density of disease onsets in the observational area, the incidence rate has also been referred to as *incidence density* (Miettinen, 1976a). The measure has also been called the *person-time rate, force of morbidity* (or *force of mortality* in reference to deaths), *hazard rate,* and *disease intensity,* although the latter three terms are more commonly used to refer to the theoretical limit approached by an incidence rate as the time interval is narrowed toward zero.

Interpretation of an Incidence Rate

The numerical portion of an incidence rate has a lower bound of zero but has no upper bound; it has the mathematical range for the ratio of two non-negative quantities, in this case the number of events in the numerator and the person-time in the denominator. At first, it may seem surprising that an incidence rate can exceed the value of 1.0, which would seem to indicate that more than 100% of a population is affected. It is true that at most only 100% of persons in a population can get a disease, but the incidence rate does not measure the proportion of a population with illness and in fact is not a proportion at all. Recall that incidence rate is measured in units of the reciprocal of time. Among 100 people, no more than 100 deaths can occur, but those 100 deaths can occur in 10,000 person-years, in 1000 person-years, in 100 person-years, or even in 1 person-year (if the 100 deaths occur after an average of 3.65 days each). An incidence rate of 100 cases (or deaths) per 1 person-year might be expressed as

$$100 \frac{\text{cases}}{\text{person-year}}.$$

It might also be expressed as

$$10{,}000 \frac{\text{cases}}{\text{person-century}},$$

$$8.33 \frac{\text{cases}}{\text{person-month}},$$

$$1.92 \frac{\text{cases}}{\text{person-week}}, \text{ or}$$

$$0.27 \frac{\text{cases}}{\text{person-day}}.$$

The numerical value of an incidence rate in itself has no interpretability because it depends on the arbitrary selection of the time unit. It is thus essential in presenting incidence rates to give the appropriate time units, either as in the examples given above or as in 8.33 month^{-1} or 1.92 week^{-1}. Although the measure of time in the denominator of an incidence rate is often taken in terms of years, one can have units of years in the denominator regardless of whether the observations were collected over 1 year of time, over 1 week of time, or over 10 years of time.

The reciprocal of time is an awkward concept that does not provide an intuitive grasp of an incidence rate. The measure does, however, have a close connection to more interpretable measures of occurrence in closed populations. Referring to Fig. 3-2, one can see that the area under the curve is equal to $N \times T$, where N is the number of people starting out in the closed population and T is the average time until death. Equivalently, the area under the curve in Fig. 3-2 is equal to the area of a rectangle with height N and width T.

Since T is the average time until death for the N people, the total person-time experience is $N \times T$. The time-averaged death rate when the follow-up for the closed population is complete is $N/(N \times T) = 1/T$; that is, the death rate equals the reciprocal of the average time until death.

More generally, in a stationary population with no migration, the crude incidence rate of an inevitable outcome such as death will equal the reciprocal of the average time until the outcome. The time until the outcome is sometimes referred to as the "waiting time" until the event occurs (Morrison, 1979). Thus, in a stationary population with no migration, a death rate of 0.04 year^{-1} would translate to an average time until death of 25 years.

If the outcome of interest is not death but either disease onset or death from a specific cause, the waiting-time interpretation must be modified slightly: The waiting time is the average time until disease onset, assuming that a person is not at risk of other causes of death or other events that remove one from risk of the outcome of interest. That is, the waiting time must be redefined to account for *competing risks,* which are events that "compete" with the outcome of interest to remove persons from the population at risk.

Unfortunately, the interpretation of incidence rates as the inverse of the average waiting time will usually not be valid unless the incidence rate is calculated for a stationary population with no migration (no immigration or emigration) or a closed population with complete follow-up. For example, the death rate for the United States in 1977 was 0.0088 year^{-1}; in a steady state, this rate would correspond to a mean life-span, or expectation of life, of 114 years. Other analyses, however, indicate that the actual expectation of life in 1977 was 73 years (Alho, 1992). The discrepancy is due to immigration and to the lack of a steady state. Note that the no-migration assumption cannot hold within specific age groups, for people are always "migrating" in and out of age groups as they age.

While the notion of incidence is a central one in epidemiology, it cannot capture all aspects of disease occurrence. This much may be clear by considering that a rate of 1 case/(100 years) = 0.01 year^{-1} could be obtained by following 100 people for an average of 1 year and observing one case, but could also be obtained by following two people for 50 years and observing one case, a very different scenario. To distinguish these situations, concepts that directly incorporate the notion of follow-up time and risk are needed.

OTHER TYPES OF RATES

In addition to numbers of cases per unit of person-time, it is sometimes useful to examine numbers of events per other unit. In health services and infectious-disease epidemiology, epidemic curves are often depicted in terms of the number of cases per unit time, or *absolute rate,*

$$\frac{\text{No. of disease onsets}}{\text{Time span of observation}},$$

or $A/\Delta t$. Because the person-time rate is simply this absolute rate divided by the average size of the population over the time span, or $A/(\bar{N} \cdot \Delta t)$, the person-time rate has been called the *relative rate* (Elandt-Johnson, 1975); it is the absolute rate relative to or "adjusted for" the average population size.

Sometimes it is useful to express event rates in units not directly involving time. A common example is the expression of fatalities by travel modality in terms of passenger-

breast cancer, the rate of 652 per million person-years is a total for the rate of occurrence of cases caused by the radiation and the rate of occurrence of cases that are not related to radiation. By measuring the rate of disease among a population of Japanese women who had negligible radiation exposure, we might estimate what the rate would have been among those exposed to 100+ rad if their radiation exposure not occurred. By subtracting this value, we obtain an estimate of the excess rate due to the high dose of radiation. For this estimate to be valid, the rate among those with negligible radiation exposure must be equal to the rate that those with 100+ rad exposure would have had if they had not been exposed. This crucial (and unlikely) condition requires that there be no confounding.

Confounders

Consider again the fluoridation example. Suppose that within the year after fluoridation began, dental-hygiene education programs were implemented in some of the schools in the community. If these programs were effective, then (other things being equal) some reduction in caries incidence would have occurred as a consequence of the programs. Thus, even if fluoridation had not begun, the caries incidence would have declined in the postfluoridation time period. In other words, the programs alone would have caused the counterfactual rate in our effect measure to be lower than the prefluoridation rate that substitutes for it. As a result, the measure of association (which is the before–after rate difference) must be larger than the desired measure of effect (the causal rate difference). In this situation, we say the programs *confounded* the measure of association or that the program effects are confounded with the fluoridation effect in the measure of association. We also say that the programs are *confounders* of the association and that the association is confounded by the programs.

Confounders are factors (exposures, interventions, treatments, etc.) that explain or produce confounding. In the present example, the programs explain why the before–after association overstates the fluoridation effect: The before–after risk difference or ratio includes the effects of programs, as well as the effects of fluoridation. More generally, a confounder explains a discrepancy between the desired (but unobservable) counterfactual risk or rate (which the exposed would have had, had they been unexposed) and the unexposed risk or rate that was its substitute. In order for a factor to explain this discrepancy and thus confound, it must be capable of affecting or at least predicting the risk or rate in the unexposed (reference) group, and not be affected by the exposure or the disease. In the above example, we assumed that the presence of the dental-hygiene programs in the years after fluoridation entirely accounted for the discrepancy between the prefluoridation rate and the (counterfactual) rate that would have occurred 3 years after fluoridation if fluoridation had not been introduced.

A large portion of epidemiologic methods are concerned with avoiding or adjusting (controlling) for confounding. Such methods inevitably rely on the gathering and proper use of confounder measurements. We will repeatedly return to this topic. For now, we simply note that the most fundamental adjustment methods rely on the notion of *stratification* on confounders. If we make our comparisons within specific levels of a confounder, those comparisons cannot be confounded by that confounder. For example, we could limit our before–after fluoridation comparisons to schools in states in which no dental-hygiene program was introduced. In such schools, program introductions could not have had an effect (because no program was present), and so any decline following fluoridation could not be explained by effects of programs in those schools.

STANDARDIZED MEASURES

Consider again the concept of standardization as introduced at the end of Chapter 3. Given a standard distribution $T_1, \ldots, T_K$ of person-times across K categories or strata defined by one or more variables and a schedule $I_1, \ldots, I_K$ of incidence rates in those categories, we have the standardized rate

$$I_s = \frac{\sum_{k=1}^{K} T_k I_k}{\sum_{k=1}^{K} T_k},$$

which is the average of the I_k weighted by the T_k. If $I_1^*, \ldots, I_K^*$ represents another schedule of rates for the same categories, and

$$I_s^* = \frac{\sum_{k=1}^{K} T_k I_k^*}{\sum_{k=1}^{K} T_k},$$

is the standardized rate for this schedule, then

$$IR_s = \frac{I_s}{I_s^*}$$

is called a *standardized rate ratio*. The defining feature of this ratio is that the same standard distribution is used to weight the numerator and denominator rate.

Suppose $I_1, \ldots, I_K$ represent the rates observed or predicted for strata of a given target population if it is exposed to some cause or preventive of disease, $T_1, \ldots, T_K$ are the observed person-time in strata of that population, and $I_1^*, \ldots, I_K^*$ represent the rates predicted or observed for strata of the population if it is not exposed. The presumption is then that $IR_s = I_s/I_s^*$ is the effect of exposure on this population, comparing the overall (crude) rates that would occur under distinct exposure conditions. This interpretation assumes, however, that the relative distribution of person-times would be unaffected by exposure. As alluded to in Chapter 3, however, if $I_1^*, \ldots, I_K^*$ represent counterfactual rather than actual rates, say, because the population was actually exposed, then I_s^* need not represent the overall rate that would occur in the population if exposure were removed (Greenland, 1996a). For instance, the change in rates from the I_k to the I_k^* could shift the person-time distribution $T_1, \ldots, T_K$ to $T_1^*, \ldots, T_K^*$. In addition, the exposure could affect competing risks, and this effect could also shift the person-time distribution.

There are a few special conditions under which the effect of exposure on person-time will not affect the standardized rate ratio. If the stratum-specific ratios I_k/I_k^* are constant across categories, the standardized rate ratio will equal this constant stratum-specific ratio. If the exposure has only a small effect on person-time, then, regardless of the person-time distribution used as the standard, the difference between a standardized ratio and the actual effect will also be small. In general, however, one should be alert to the fact that a special assumption is needed to allow one to interpret a standardized rate ratio as an effect measure, even if there is no methodologic problem with the observations. Analogously, the standardized rate difference will not be an effect measure except when exposure does not affect the person-time distribution or when other special conditions, such as constant rate differences $I_k - I_k^*$ across categories, exist.

standardized risk ratios and differences, as opposed to standardized rate measures, have denominators that are not affected by changing rates or competing risks, and thus can be interpreted as effect measures without the need for special assumptions.

PREVALENCE RATIOS

In Chapter 3 we showed that the crude prevalence odds, PO, equals the crude incidence rate, I, times the average disease duration, $\overline{D}$, when both the population at risk and the prevalence pool are stationary and there is no migration in or out of the prevalence pool. Restating this relation separately for a single population under exposure and nonexposure, or for one exposed and one unexposed population, we have

$$PO_1 = I_1\overline{D}_1 \quad \text{and} \quad PO_0 = I_0\overline{D}_0\,, \tag{4-5}$$

where the subscripts 1 and 0 refer to exposed and unexposed, respectively. If the average durations $\overline{D}_1$ and $\overline{D}_0$ are equal, we find that the crude prevalence odds ratio POR, equals the crude incidence ratio IR:

$$POR = \frac{PO_1}{PO_0} = \frac{I_1}{I_0} = IR\,. \tag{4-6}$$

Unfortunately, if exposure affects mortality, it will also alter the age distribution of the population. This is because older people tend to die sooner, exposure will indirectly affect average duration, so that $\overline{D}_1$ will not equal $\overline{D}_0$. In that case equation 4-6 will not hold exactly although it may still hold approximately (Newman, 1988).

OTHER MEASURES

The measures that we have discussed are by no means exhaustive of all those that have been proposed. Not all proposed measures of effect meet our definition of effect measure—that is, a contrast of the outcome of a *single* population under two *different* conditions. Examples of measures that are *not* effect measures by our definition include correlation coefficients and related variance-reduction measures (Greenland et al., 1986, 1991). Examples of measures that are effect measures by our definition but not discussed here, include expected years of life lost, as well as risk and rate advancement periods. See Robins and Greenland (1991), Boshuizen and Greenland (1997), and Brenner et al. (1993) for overviews of issues in defining and estimating these measures.

Diagnostic Bias

Another type of selection bias occurring before subjects are identified for study is *diagnostic bias* (Sackett, 1979). When the relation between oral contraceptives and venous thromboembolism was first investigated with case-control studies of hospitalized patients, there was concern that some of the women had been hospitalized with a diagnosis of venous thromboembolism because their physicians suspected a relation between this disease and oral contraceptives and had known about oral contraceptive use in patients who presented with suggestive symptoms (Sartwell et al., 1969). A study of hospitalized patients with thromboembolism could lead to an exaggerated estimate of the effect of oral contraceptives on thromboembolism if the hospitalization and determination of the diagnosis were influenced by the history of oral-contraceptive use.

Confounding

The concept of confounding is a central one in modern epidemiology. Although confounding occurs in experimental research, it is a considerably more important issue in nonexperimental research. Consequently, the understanding of the concept has developed only recently in parallel with the growth of nonexperimental research. Therefore, we will here review the concepts of confounding and confounders and then discuss further issues in defining and identifying confounders.

Confounding as Mixing of Effects

On the simplest level, confounding may be considered a confusion of effects. Specifically, the apparent effect of the exposure of interest is distorted because the effect of an extraneous factor is mistaken for or mixed with the actual exposure effect (which may be null). The distortion introduced by a confounding factor can be large, and it can lead to overestimation or underestimation of an effect, depending on the direction of the associations that the confounding factor has with exposure and disease. Confounding can even change the apparent direction of an effect.

A more precise definition of confounding begins by considering the manner in which effects are estimated. As described in Chapter 4, we wish to estimate the degree to which exposure has changed the frequency of disease in an exposed cohort. To do so, we must estimate what the frequency of disease would have been in this cohort had exposure been absent. To accomplish this task, we observe the disease frequency in an unexposed cohort. But rarely could we take this unexposed frequency as fairly representing what the frequency would have been in the exposed cohort had exposure been absent, because the unexposed cohort would differ from the exposed cohort on many factors that affect disease frequency besides exposure. To express this problem, we say that the comparison of the exposed and unexposed is *confounded* because the difference in disease frequency between the exposed and unexposed results from a mixture of several effects, including (but not limited to) any exposure effect.

Confounders and Surrogate Confounders

The extraneous factors responsible for difference in disease frequency between the exposed and unexposed are called *confounders*. In addition, factors associated with

these extraneous causal factors that can serve as surrogates for these factors are also commonly called confounders. The most extreme example of such a surrogate is chronologic age. Increasing age is strongly associated with *aging*—the accumulation of cell mutations and tissue damage that leads to disease—but increasing age does not itself cause such pathogenic changes, for it is just a measure of how much time has passed since birth.

Regardless of whether a confounder is a cause of the study disease or merely a surrogate for such a cause, its chief characteristic is that it would be predictive of disease frequency within the unexposed (reference) cohort; otherwise, it could not explain why the unexposed cohort fails to represent properly what the exposed cohort would experience in the absence of exposure. For example, suppose all the exposed were men and all the unexposed were women. If unexposed men would have the same incidence as unexposed women, the fact that all the unexposed were women rather than men could not account for any confounding that is present.

Confounding of a Zero Effect

In the simple view, confounding occurs only if extraneous effects become mixed with the effect under study. Nevertheless, confounding can occur even if the factor under study has zero effect. Thus, "mixing of effects" should not be taken to imply that the exposure under study has a nonzero effect. The mixing of the effects comes about from an association between the exposure and extraneous factors.

As an example, consider a study to determine whether alcohol drinkers experience a greater incidence of oral cancer than nondrinkers. Smoking is an extraneous factor that is related to the disease among the unexposed (smoking has an effect on oral cancer incidence among alcohol abstainers); it is also associated with alcohol drinking, since there are many people who are general "abstainers," refraining from alcohol consumption, smoking, and perhaps other habits. Consequently, alcohol drinkers include among them a greater proportion of smokers than would be found among nondrinkers. Since smoking increases the incidence of oral cancer, alcohol drinkers will have a greater incidence than nondrinkers, quite apart from any influence of alcohol drinking itself, simply as a consequence of the greater amount of smoking among alcohol drinkers. Thus, the apparent effect of alcohol drinking is distorted by the effect of smoking; the effect of smoking becomes mixed with the estimated effect of alcohol in the comparison of alcohol drinkers with nondrinkers. The degree of bias or distortion depends on the magnitude of the smoking effect, as well as on the strength of association between alcohol and smoking. Either absence of a smoking effect on oral cancer incidence or absence of an association between smoking and alcohol would lead to no confounding. Smoking must be associated with both oral cancer and alcohol drinking for it to be a confounding factor.

Properties of a Confounder

In general, a confounder must be associated with both the exposure under study and the disease under study to be confounding. These associations do not, however, define a confounder, for a variable may possess these associations and yet not be a confounder. There are several ways this can happen. The most common way occurs when the exposure under study has an effect. In this situation, any correlate of that exposure will also be associated with the disease as a consequence of its association with a risk factor for

the disease. For example, suppose frequent beer consumption is associated with the consumption of pizza, and suppose that frequent beer consumption is a risk factor for rectal cancer. Would consumption of pizza be a confounding factor? At first, it might seem that the answer is yes, since consumption of pizza is associated both with beer drinking and with rectal cancer. But if pizza consumption is associated with rectal cancer only secondarily to its association with beer consumption, it would not be confounding. A confounding factor must be predictive of disease occurrence apart from its association with exposure; that is, as explained above, among unexposed (reference) individuals, the potentially confounding variate should be related to disease risk. If consumption of pizza were predictive of rectal cancer among nondrinkers of beer, then it could confound; otherwise, if it were associated with rectal cancer only from its association with beer drinking, it could not confound.

Analogous with this restriction on the association between a potential confounder and disease, the potential confounder should be associated with the exposure among the source population for cases, not merely among cases of the disease as a consequence of both variables being risk factors for disease.

Confounders as Extraneous Risk Factors

It is also important to clarify what is meant by the term *extraneous* in the phrase "extraneous risk factor." This term implies that the predictiveness for disease risk involves a mechanism other than the one under study. Specifically, consider a causal mechanism where

$$\text{smoking} \xrightarrow{\text{causes}} \text{elevated blood pressure} \xrightarrow{\text{causes}} \text{heart disease}$$

Is elevated blood pressure a confounding factor? It is certainly a risk factor for disease, and it is also correlated with exposure, since it can result from smoking. It is even a risk factor for disease among nonexposed individuals, since elevated blood pressure can result from causes other than smoking. Nevertheless, it cannot be considered a purely confounding factor, since the effect of smoking is mediated through the effect of blood pressure. In this example, there may be no mixing of confounder with exposure effects, but the factor (elevated blood pressure) does mediate the exposure (smoking) effects. Any factor that represents a step in the causal chain between exposure and disease should not be treated as an extraneous confounding factor, but instead requires special treatment as an *intermediate* factor (Greenland and Neutra, 1980; Robins, 1989).

Judging the Causal Role of a Potential Confounder

Usually, an explicit mechanism for the causal action of the exposure is not postulated. How then can an investigator decide if a factor is extraneous or not? Such decisions must be made on the basis of the best available information, including nonepidemiologic (i.e., clinical) data. Uncertainties about the mechanism can justify the handling of a potential confounding factor as both confounding and not confounding in different analyses. For example, in evaluating the effect of coffee on heart disease, it is unclear how to treat serum cholesterol levels. Elevated levels are a risk factor for heart disease and may be associated with coffee use, but serum cholesterol may mediate the action of coffee use on heart disease risk; that is, elevated cholesterol may be an intermediate factor in the etiologic sequence under study. In the face of uncertainty, one might conduct two analyses, one in which serum cholesterol is controlled (which would be appropriate if coffee does

not affect serum cholesterol) and one in which it is not controlled (which would be more appropriate if coffee affects serum cholesterol and is not associated with uncontrolled determinants of serum cholesterol). The interpretation of the results would depend on which of the theories about serum cholesterol were correct.

Criteria for a Confounding Factor

We can summarize our observations thus far with three criteria for a variable to be a confounder. To be a confounder, the extraneous variable must have three necessary (but not sufficient or defining) characteristics, which we will discuss in detail. We will then point out some limitations of these characteristics in defining and identifying confounding.

1. A confounding factor must be a risk factor for the disease.

As mentioned earlier, the potential confounding factor need not be an actual cause of the disease, but if it is not, it must be a marker for an actual cause of the disease. The association between the potential confounder and the disease should not derive only secondarily from an association with the exposure, which may be a cause of the disease. Therefore, a confounding factor must be a risk factor within the reference level of the exposure under study. Furthermore, the data may serve as a guide to the relation between the potential confounder and the disease, but it is the actual relation between the potentially confounding factor and disease, not the apparent relation observed in the data, that determines whether confounding can occur (Miettinen and Cook, 1981). In large studies, which are subject to less sampling error, we expect the data to reflect more closely the underlying relation, but in small studies the data may be a less reliable guide.

The following example illustrates the role that prior knowledge can play in evaluating confounding. Suppose that in a cohort study of airborne glass fibers and lung cancer, the data show more smoking and more cancers among the heavily exposed but no relation between smoking and lung cancer within exposure levels. The latter absence of a relation does not mean that some smoking effect was not confounding (mixed into) the estimated effect of glass fibers: It may be that some or all of the excess cancers in the heavily exposed were produced solely by smoking and that the lack of smoking–cancer association was produced by unmeasured confounding of this association in this cohort. The latter confounding might arise from nothing more than an unfortunate confluence of several unmeasured risk factors among the nonsmokers.

As a converse example, suppose we conduct a cohort study of sunlight exposure and melanoma. Our best current information indicates that after control for age, there is no relation between social security number and melanoma occurrence. Thus, we would not consider social security number a confounder, regardless of its association with melanoma in the reference exposure cohort, because we think it cannot be used to predict the rate in this cohort (i.e., we think the rate in this cohort would not have been different had the subjects received different social security numbers). Even if control of social security number would change the effect estimate, the resulting estimate of effect would be less valid than one that ignores social security number, given our prior information about the lack of a real effect of social security number.

Nevertheless, because external information is usually limited, investigators rely heavily on their data to infer the predictive ability of a potential confounder. For example, a cause of disease in one population will be causally unrelated to disease in another popu-

lation that lacks complementary component causes (i.e., susceptibility factors). A discordance between the data and external information about a suspected or known risk factor may therefore signal inadequacy in the detail of information about interacting factors rather than an error in the data. Similarly, external information about the absence of an effect for a possible risk factor may be inadequate, if based on studies with considerable bias toward the null. On the other hand, it is also conceivable that external information about the absence of an effect could override any evidence to the contrary in the data, as in the melanoma example above.

2. A confounding factor must be associated with the exposure under study in the source population (the population at risk from which the cases are derived).

The association between a potential confounding factor and the exposure must not derive secondarily from the association between the exposure and disease. In a cohort study, this proviso implies only that the association between the potential confounding factor and the exposure must be present among subjects at the start of follow-up. Thus, in cohort studies, the exposure–confounder association can be evaluated from the data in hand and does not even theoretically depend on prior knowledge.

When the exposure under study has been randomly assigned, it is sometimes mistakenly thought that confounding cannot occur because randomization somehow guarantees there will be no association between the exposure and other factors. In reality, randomization is only a probabilistic procedure that can leave some association of the exposure and extraneous risk factors, especially if the total number of subjects is small. Thus, confounding can occur in randomized trials, even though it tends to be more minor in extent than in nonrandomized studies (Rothman, 1977) and to be negligible in well-conducted very large trials.

In a case-control study, the proviso implies that the association must be present in the source population that gave rise to the cases. If the control series is large and there is no selection bias, it should provide a reasonable estimate of the association between the potential confounding variable and the exposure in the source population. Nevertheless, the ultimate concern focuses on the degree of association between the potential confounder and the exposure in the source population that produced the study cases, of which the controls are only a substitute (Miettinen and Cook, 1981). If available, information on this population association can be used to adjust findings from the control series. Unfortunately, reliable external information about the associations among risk factors in the source population is seldom available. Thus, in case-control studies, the data in hand will usually have to provide an estimate of the association between the exposure and the potentially confounding factor.

Consider a case-control study of occupational exposure to airborne glass fibers and the occurrence of lung cancer that randomly sampled cases and controls from cases and noncases in an occupational cohort. Suppose we knew the association of exposure and smoking in the full cohort. We could then use the discrepancy between the true association and the exposure–smoking association observed in the controls as a measure of the extent to which random sampling had failed to produce representative controls. If this discrepancy were known, it could be used to adjust the control numbers to make them appear representative of the cohort. Regardless of the size of this discrepancy, if there was no association of smoking and exposure in the source cohort, the unadjusted estimate would be the best available estimate, and so smoking would not be a confounder in the case-control study (Robins and Morgenstern, 1987).

In contrast, consider a randomized trial of a treatment. Although the average association between any risk factor and treatment is zero over repeated randomizations, it can easily happen that a risk factor (despite the randomization) is associated with the treatment in the one randomized cohort that is observed. In this situation, adjustment for the risk factor would produce the best available estimate, and so the factor would be a confounder in the trial.

3. A confounding factor must not be affected by the exposure or the disease. In particular, it cannot be an intermediate step in the causal path between the exposure and the disease.

This criterion is obviously satisfied if the factor precedes exposure and disease. Otherwise, the criterion requires information outside the data. The investigator must decide whether a causal mechanism exists that might lead from exposure or disease to the potentially confounding factor. If the factor is an intermediate step between exposure and disease, it should not be treated as simply a confounding factor; instead, a more careful analysis that takes account of its intermediate nature is required (Robins, 1989; Robins and Greenland, 1992).

It is important to remember that confounding is a bias and therefore must be considered and dealt with as a quantitative problem. It is the amount of confounding rather than mere presence or absence that is important to evaluate. In one study, a rate ratio of 5 may become 4.6 after control of age, whereas in another study a rate ratio of 5 may change to 1.6 after control of age. Although age is confounding in both studies, in the former the amount of confounding is comparatively unimportant, whereas in the latter confounding accounts for nearly all of the strong effect. Methods to evaluate confounding quantitatively are described in Chapter 15.

Although the above three characteristics of confounders are sometimes taken to define a confounder, it is a mistake to do so for both conceptual and technical reasons. Conceptually, the essence of confounding is the confusion or mixing of extraneous effects with the effect of interest. The first two properties are simply logical consequences of the basic definition, properties that a factor must satisfy in order to confound; the third property excludes situations in which the effects cannot be disentangled in a straightforward manner (except in special cases). Technically, it is possible for a factor to possess all three characteristics and yet not have its effects mixed with the exposure, in the sense that a factor may produce no spurious excess or deficit of disease among the exposed, despite its association with exposure and its effect on disease. This result can occur, for example, when the factor is but one of several potential confounders and the excess of incidence produced by the factor among the exposed is perfectly balanced by the excess incidence produced by another factor in the unexposed.*

Information Bias

Once the subjects to be compared have been identified, the information to be compared must be obtained. Bias in evaluating an effect can occur from errors in obtaining the

*This discussion omits a number of subtleties that arise in determining which variables should or should not be controlled in a given analysis. For discussions of these issues and their relation to standard criteria for confounder control, see Pearl (1995), Pearl and Robins (1995), and Greenland et al. (1999).

= 0.52 among the tolbutamide treated, but was 120/205 = 0.59 among the placebo treated. Finally, we know with certainty that tolbutamide does *not* alter a person's age.

Although it is possible to obtain a general appreciation for the presence or absence of confounding in data by examining whether a potentially confounding factor is associated with disease conditional on exposure and with exposure in the source population, the magnitude of the confounding is difficult to assess in this way because it is a function of both of these component associations. Further, when several factors are simultaneously confounding, the component associations should ideally be examined conditional on the other confounding factors, thereby complicating the problem.

More direct methods for confounder assessment compare the estimates of effect obtained with and without control of each potential confounder (assuming that the potential confounder is not affected by exposure). The magnitude of confounding is estimated by the degree of discrepancy between the two estimates. For example, the unadjusted risk difference in Table 15-1 is 0.147 − 0.102 = 0.045. If we adjust for age confounding by standardizing (averaging) the age-specific risks in Table 15-1 using the total cohort as the standard (see Chapter 4), we obtain a standardized risk-difference of

$$\frac{226(0.076) + 183(0.224)}{226 + 183} - \frac{226(0.042) + 183(0.188)}{226 + 183} = 0.142 - 0.107 = 0.035.$$

Thus, the relatively crude age adjustments obtained by treating age as a dichotomy has reduced the estimated risk difference produced by tolbutamide from 4.5% to 3.5%. Similarly, the unadjusted risk ratio in Table 15-1 is 0.147/0.102 = 1.44, where the age-standardized risk ratio is 0.142/0.107 = 1.33.

Selecting Confounders for Control

Having computed estimates both with and without adjustment for the age dichotomy (under 55 versus 55+), the analyst must now decide whether it is important to adjust for this variable when presenting results. It may be important to do so simply because many readers would not trust results that are not adjusted for age. This distrust stems from knowledge that age is strongly related to disease and mortality rates (similar comments would apply to sex). Suppose, however, we wish to apply a quantitative criterion to see whether we must control for age and other variables. To do so, the analyst must choose a cut-off for what constitutes an important change in the estimate. In Table 15-1 the unadjusted risk ratio is (1.44 − 1.33)/1.33 = 8% larger than the adjusted. If only changes of greater than 10% are considered important, then this change is not important; but if changes of greater than 5% are considered important, then this change is important and indicates that age should not be ignored in further analyses.

The exact cutoff for importance is somewhat arbitrary but limited in range by the subject matter. For example, a 5% change in the risk ratio would be considered ignorable in most contexts, but rarely if ever would a 50% change. Similar observations would apply when considering confidence limits. The most important point is that one should report the criterion used to select confounders for adjustment.

Although many have argued against the practice (Miettinen, 1976b; Breslow and Day, 1980; Greenland and Neutra, 1980; Greenland, 1989), one often sees statistical tests used to select confounders (as in stepwise regression), rather than the change-in-estimate criterion just discussed. Usually, the tests are of the confounder-disease association, although sometimes the difference between the unadjusted and adjusted estimates are tested (the latter approach is often termed collapsibility testing). It has been argued that

these testing approaches will perform adequately if the tests have high enough power to detect any important confounder effects. One way to insure adequate power is to raise the alpha-level for rejecting the null (of no confounding) to 0.20 or even more, instead of using the traditional 0.05 level (Dales and Ury, 1978). Limited simulation studies indicate that this approach is reasonable, in that use of a 0.20 or higher alpha level instead of a 0.05 level for confounder selection can make the difference between acceptable and poor performance of statistical testing for confounder selection (Mickey and Greenland, 1989; Maldonado and Greenland, 1993a).

Several important subtleties must be considered when more than one potential confounder must be examined. First, it can make a big difference in the observed change in estimate whether one evaluates the change with or without adjustment for other confounders. For example, suppose we have to consider adjustment for age and sex. To evaluate age, we could compare the estimates without and with age adjustment, ignoring sex in both instances. Or we could compare the estimate with age and sex adjustment to that with only sex adjustment. In other words, we could evaluate age confounding without or with background adjustment for sex. Furthermore, we could evaluate sex confounding with or without background adjustment for age. Our decision about importance could be strongly influenced by the strategy we choose.

To cope with this complexity, several authors have suggested the following "backward deletion" strategy (Miettinen, 1976b; Kleinbaum et al., 1984): First, one adjusts for all the potential confounders one can. Then, if one would like to use fewer confounders in further analyses, one deletes the confounders from adjustment one-by-one in a stepwise fashion, at each step deleting that confounder that makes the smallest change in the exposure effect estimate upon deletion. One stops deleting confounders when the *total* change in the estimate and confidence limits accrued from the start of the process (with all confounders controlled) would exceed the chosen limit of importance. One often sees analogous stepwise confounder-selection strategies based on testing the confounder coefficients and deleting in sequence the least statistically significant coefficient; again, such strategies can produce extremely confounded results unless the alpha-levels for deletion and retention are set much higher than 0.05 (Dales and Ury, 1978; Maldonado and Greenland, 1993).

Sometimes not a single confounder can be deleted without producing important changes, but more often at least a few will appear to be ignorable if others are controlled. Sometimes, however, it is impossible to control all the confounders (at least by stratification) because the data become too thinly spread across strata to yield any estimate at all (this occurs when no stratum contains both a case and a noncase, as well as in other situations). When this problem occurs, the pure "backwards deletion" strategy just described cannot be implemented. One approach proposed for this situation is to use a "forward selection" strategy, in which one starts with the exposure effect estimate from the simplest acceptable stratification (e.g., one involving only age and sex), then stratifies on the confounder that makes the most difference in the estimate, then adds confounders one-by-one to the stratification, at each step adding the confounder that makes the most difference among those not yet added. The process stops when addition of variables ceases to make an "important" difference.

Statistical Biases in Variable Selection

If the data become very thin when all or most confounders are used for stratification, all confounder-selection strategies based on approximate statistics can suffer from certain

statistical artifacts that lead to very biased final results. No conventional approach to confounding (based on change-in-estimate or more traditional significance testing) can wholly address this problem (Robins and Greenland, 1986). There are certain modeling methods (which are briefly discussed in Chapter 21 under the topic of hierarchical regression) that can cope with these situations, but these methods are unavailable in most software packages. For this reason, epidemiologists often resort to some sort of forward-selection strategy when data are sparse.

There is a hallmark symptom of the bias that arises when stratification has exceeded the limits of the data: The exposure effect estimates begin to get further and further from the null as more variables are added to the stratification or regression model. For example, one might observe only modest effect estimates as one moves from adjustment for the strongest confounder alone to adjustment for the two or three strongest confounders. Then, with further adjustment, the exposure effect estimate becomes enormous (e.g., odds ratios of greater than 10 or less than 0.10) as more confounders are controlled. This inflation is sometimes mistakenly interpreted as evidence of confounding, but in our experience is more often bias due to applying large-sample methods to excessively sparse data.

Another problem with all variable-selection approaches (again, whether based on change-in-estimate or statistical testing) is their potential to distort *P*-values and confidence intervals for exposure effect away from their nominal behavior. For example, conventional 95% confidence intervals computed after using the data to select variables can have true coverage less than 95% because the computation of such intervals assumes no selection of variables has been done (Greenland, 1989a, 1993a; Hurvich and Tsai, 1990). The limited studies performed thus far suggest that the distortion produced by typical confounder-selection strategies need not be large in practice (Mickey and Greenland, 1989; Maldonado and Greenland, 1993), but further study is needed.

One way to reduce distortion due to confounder selection is to insist that the confidence limits do not change to an important degree if a confounder is to be deleted from control. If one uses confidence limits rather than the point estimate to monitor the change produced by adding or deleting control of a confounder, one can use exact confidence limits rather than the usual large-sample approximate limits produced by Mantel-Haenszel or maximum-likelihood methods. With exact limits, the sparse-data bias discussed earlier will not occur. Unfortunately, exact intervals can become very conservative (and very wide) if computed by the traditional Fisher-*P* method, which is the default method in most software (see Chapter 13).

Selection of confounders can lead to complex problems, especially if there are many confounders to choose from. Strategies based on examining changes in the exact confidence limits for exposure effect seem to be the best that can be carried out with standard software, although if enough data are available one may instead use approximate limits to monitor the changes. Most importantly, if selection is done, one should report the strategy used to select potential confounders for control in the methods section of the research report. In addition, one may have to include certain potential confounders on subject-matter grounds, even if they do not meet the quantitative criteria for inclusion. For example, a study of lung cancer might be well advised to adjust for smoking whenever possible, as well as age and sex, because of the known strong relations of these variables to lung-cancer rates.

Selecting Confounder Categories

An issue closely related to that of selecting confounders is that of selecting confounder categories. Some aspects of this issue are discussed in Chapter 13. In particular, we