4

Measuring Disease Occurrence and Causal Effects

As with most sciences, measurement is a central feature of epidemiology, which has been defined as the study of the occurrence of illness.[1] The broad scope of epidemiology demands a correspondingly broad interpretation of illness, to include injuries, birth defects, health outcomes, and other health-related events and conditions. The fundamental observations in epidemiology are measures of the occurrence of illness. In this chapter, I discuss several measures of disease frequency, including *risk*, *incidence rate*, and *prevalence*. I also examine how these fundamental measures can be used to obtain derivative measures that aid in quantifying potentially causal relations between exposure and disease.

MEASURES OF DISEASE OCCURRENCE

Risk and Incidence Proportion

The concept of *risk* for disease is widely used and reasonably well understood by many people. It is measured on the same scale and interpreted in the same way as a *probability*. Epidemiologists sometimes speak about risk applying to an individual, in which case they are describing the probability that a person will develop a given disease. It is usually pointless, however, to measure risk for a single person, because for most diseases, the person simply either does or does not contract the disease. For a larger group of people, we can describe the proportion who developed the disease. If a population has N people and A people of the N develop disease during a period of time, the proportion A/N represents the average risk of disease in the population during that period:

$$\text{Risk} = \frac{A}{N} = \frac{\text{Number of subjects developing disease during a time period}}{\text{Number of subjects followed for the time period}}$$

The measure of risk requires that all of the N people are followed for the entire time during which the risk is being measured. The average risk for a group is also referred to as the *incidence proportion*. The word *risk* often is used in reference to a single person, and *incidence proportion* is used in reference to a group of people (Table 4–1). Because averages taken from populations are used to estimate the risk for individuals, the two terms often are used synonymously. We can use risk or incidence proportion to assess the onset of disease, death from a given disease, or any event that marks a health outcome.

One of the primary advantages of using risk as a measure of disease frequency is the extent to which it is readily understood by many people, including those who have little familiarity with epidemiology. To make risk useful as a technical or scientific measure, however, we need to clarify the concept. Suppose you read in the newspaper that women who are 60 years old have a 2% risk of dying of cardiovascular disease. What does this statement mean? If you consider the possibilities, you may soon realize that the statement as written cannot be interpreted. It is certainly not true that a typical 60-year-old woman has a 2% chance of dying of cardiovascular disease within the next 24 hours or in the next week or month. A 2% risk would be high even for 1 year, unless the women in question have one or more characteristics that put them at unusually high risk compared with most 60-year-old women. The risk of developing fatal cardiovascular disease over the remaining lifetime of 60-year-old women, however, would likely be well above 2%. There might be some period over which the 2% figure would be correct, but any other period of time would imply a different value for the risk.

The only way to interpret a risk is to know the length of time over which the risk applies. This period may be short or long, but without identifying it, risk values are not meaningful. Over a very short time period, the risk of any particular disease is usually extremely low. What is the probability that a given person will develop a given disease in the next 5 minutes? It is close to zero. The total risk over a period of time may climb from zero at the start of the period to a maximum theoretical limit of 100%, but it cannot decrease with time. Figure 4–1 illustrates two different possible patterns of risk during a 20-year interval. In pattern A, the risk climbs rapidly early during the period and then plateaus, whereas in pattern B, the risk climbs at a steadily increasing rate during the period.

How might these different risk patterns occur? As an example, a pattern similar to A could occur if a person who is susceptible to an infectious disease becomes immunized, in which case the leveling off of risk is sudden, not gradual. Pattern A also could occur if those who come into contact with a susceptible person become immunized, reducing the susceptible person's risk of acquiring the disease. A pattern similar to B could occur if a person who has been exposed to a cause

Table 4–1 Comparison of Incidence Proportion (Risk) and Incidence Rate

Property	Incidence Proportion	Incidence Rate
Smallest value	0	0
Greatest value	1	Infinity
Units (dimensionality)	None	1/time
Interpretation	Probability	Inverse of waiting time

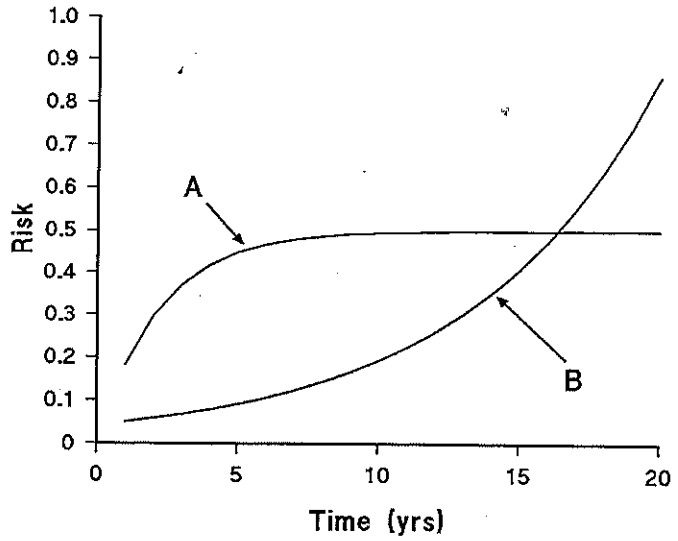

Figure 4–1 Two possible patterns of disease risk with time.

is nearing the end of the typical induction time for the causal action, such as risk of adenocarcinoma of the vagina among young women who were exposed to diethylstilbestrol (DES) while they were fetuses, as discussed in Chapter 3. In that example, the shape of the curve is similar to that of B in Figure 4–1, but the actual risks are much lower than those in Figure 4–1. Another phenomenon that can give rise to pattern B is the aging process, which often leads to sharply increasing risks as people progress beyond middle age.

Risk is a cumulative measure. For a given person, risk increases with the length of the risk period. For a given risk period, however, risks for a person can rise or fall with time. Consider the 1-year risk of dying in an automobile crash for a driver. For any one person during a period of 1 year, the risk cumulates steadily from zero at the beginning of the year to a final value at the end of that year. Nevertheless, the 1-year risk is greater for most drivers in their teenage years than for the same drivers when they reach their 50s.

Risk carries an important drawback as a tool for assessing the occurrence of illness; over any appreciable time interval, it is usually technically impossible to measure risk. The reason is a practical one: For almost any population followed for a sufficient time, some people in the population will die from causes other than the outcome under study.

Suppose that you are interested in measuring the occurrence of domestic violence in a population of 10,000 married women over a 30-year period. Unfortunately, not all 10,000 women will survive the 30-year period. Some may die from extreme instances of domestic violence, but many more are likely to die from cardiovascular disease, cancer, infection, vehicular injury, or other causes. What if a woman died after 5 years of being followed without having been a victim of domestic violence? We could not say that she would not have been

a victim of domestic violence during the subsequent 25 years. If we count her as part of the denominator, N, we will obtain an underestimate of the risk of domestic violence for a population of women who do survive 30 years. To understand why, imagine that there are many women who do not survive the 30-year follow-up period. It is likely that among them there are some women who would have experienced domestic violence if they had instead survived. If we count the women who die during the follow-up period in the denominator, N, of a risk measure, then the numerator, A, which gives the number of cases of domestic violence, will be underestimated because A is supposed to represent the number of victims of domestic violence among a population of women who were followed for a full 30 years.

This phenomenon of people being removed from a study through death from other causes is sometimes referred to as *competing risks*. There is one outcome for which there can be no competing risk: the outcome of death from any cause. If we study all deaths, there is no possibility of someone dying of a cause that we are not measuring. For any other outcome, it will always be possible for someone to die before the end of the follow-up period without experiencing the event that we are measuring. Therefore, unless we are studying all deaths, competing risks become a consideration.

Over a short period of time, the influence of competing risks usually is small. It is not unusual for studies to ignore competing risks if the follow-up period is short. For example, in the experiment in 1954 in which the Salk vaccine was tested, hundreds of thousands of schoolchildren were given either the Salk vaccine or a placebo. All of the children were followed for 1 year to assess the vaccine's efficacy. Because only a small proportion of school-age children died of competing causes during the year of the study, it was reasonable to report the results of the Salk vaccine trial in terms of the observed risks. When study participants are older or are followed for longer periods, competing risks are greater and may need to be taken into account. One way to remove competing risks is to measure incidence rates instead, and convert these to risk measures, and another is to use a life-table analysis. Both approaches are described later in this chapter.

A related issue that affects long-term follow-up is *loss to follow-up*. Some people may be hard to track to assess whether they have developed disease. They may move away or choose not to participate further in a research study. The difficulty in interpreting studies in which there have been considerable losses to follow-up is sometimes similar to the challenge of interpreting studies in which there are strong competing risks. In both situations, the researcher lacks complete follow-up of a study group for the intended period of follow-up.

Because of competing risks, it is often useful to think of risk or incidence proportion as hypothetical measures in the sense that they usually cannot be directly observed in a population. If competing risks did not occur and all losses to follow-up could be avoided, we could measure incidence proportion directly in a population by dividing the number of observed cases by the number of people in the population followed. As mentioned earlier, if the outcome of interest is death from any cause, there will be no competing risk; any death that occurs represents an outcome that will count in the numerator of the risk measure. Most attempts to measure disease risk are focused on outcomes more specific than death from any cause, such as death from a specific cause (eg, cancer, multiple

sclerosis, infection) or the occurrence of a disease rather than death. For these outcomes, there is always the possibility of competing risks. In reporting the fraction A/N, which is the observed number of cases divided by the number of people who were initially being followed, the incidence proportion that would have been observed had there been no competing risk will be underestimated, because competing risks will have removed some people from the at-risk population before their disease developed.

Attack Rate and Case-Fatality Rate

A term for risk or incidence proportion that is sometimes used in connection with infectious outbreaks is *attack rate*. An attack rate is the incidence proportion, or risk, of contracting a condition during an epidemic period. For example, if an influenza epidemic has a 10% attack rate, 10% of the population will develop the disease during the epidemic period. The time reference for an attack rate is usually not stated but is implied by the biology of the disease being described. It is usually short, typically no more than a few months, and sometimes much less. A *secondary attack rate* is the attack rate among susceptible people who come into direct contact with *primary cases*, the cases infected in the initial wave of an epidemic (see Chapter 6).

Another version of the incidence proportion that is encountered frequently in clinical medicine is the *case-fatality rate*, which is described in greater detail in Chapter 13. The case-fatality rate is the proportion of people dying of the disease (fatalities) among those who develop the disease (cases). Thus, the population at risk when a case-fatality rate is used is the population of people who have already developed the disease. The event being measured is not development of the disease but rather death from the disease (sometimes all deaths among patients, rather than only deaths from the disease, are counted). Like an attack rate, the case-fatality rate is seldom accompanied by a specific time referent, and this lack of time specificity can make it difficult to interpret. It is typically used and easiest to interpret as a description of the proportion of people who succumb to an infectious disease, such as measles. The case-fatality rate for measles in the United States is about 1.5 deaths per 1000 cases. The period for this risk of death is the comparatively short time frame during which measles infects an individual, ending in recovery, death, or some other complication. For diseases that continue to affect a person over long periods, such as multiple sclerosis, it is more difficult to interpret a measure such as case-fatality rate, and other types of mortality or survival measures are used instead.

Incidence Rate

To address the problem of competing risks, epidemiologists often resort to a different measure of disease occurrence, the *incidence rate*. This measure is similar to incidence proportion in that the numerator is the same. It is the number of cases, A, that occur in a population. The denominator is different. Instead of dividing the number of cases by the number of people who were initially being followed, the incidence rate divides the number of cases by a measure of time. This time measure is the summation across all individuals of the time experienced by the population being followed.

$$\text{Incidence rate} = \frac{A}{\text{Time}} = \frac{\text{Number of subjects developing disease}}{\text{Total time experienced for the subjects followed}}$$

One way to obtain this measure is to sum the time that each person is followed for every member of the group being followed. If a population was followed for 30 years and a given person died after 5 years of follow-up, that person would have contributed only 5 years to the sum for the group. Others might have contributed more or fewer years, up to a maximum of the full 30 years of follow-up.

For people who do not die during follow-up, there are two methods of counting the time during follow-up. These methods depend on whether the disease or event can recur. Suppose that the disease is an upper respiratory tract infection, which can occur more than once in the same person. Because the numerator of an incidence rate could contain more than one occurrence of an upper respiratory tract infection from a single person, the denominator should include all the time during which each person was at risk for getting any of these bouts of infection. In this situation, the time of follow-up for each person continues after that person recovers from an upper respiratory tract infection. On the other hand, if the event were death from leukemia, a person would be counted as a case only once. For someone who dies of leukemia, the time that would count in the denominator of an incidence rate would be the interval that begins at the start of follow-up and ends at death from leukemia. If a person can experience an event only once, the person ceases to contribute follow-up time after the event occurs.

In many situations, epidemiologists study events that can occur more than once in an individual, but they count only the first occurrence of the event. For example, researchers may count the occurrence of the first heart attack in an individual and ignore (or study separately) second or later heart attacks. If only the first occurrence of a disease is of interest, the time contribution of a person to the denominator of an incidence rate will end when the disease occurs. The unifying concept in regard to tallying the time for the denominator of an incidence rate is simple: The time that goes into the denominator corresponds to the time experienced by the people being followed during which the disease or event being studied could have occurred. For this reason, the time tallied in the denominator of an incidence rate is often referred to as the *time at risk for disease*. The time in the denominator of an incidence rate should include every moment during which a person being followed is at risk for an event that would get tallied in the numerator of the rate. For events that cannot recur, after a person experiences the event, he or she will have no more time at risk for the disease, and therefore the follow-up for that person ends with the disease occurrence. The same is true of a person who dies from a competing risk.

Figure 4–2 illustrates the time at risk for five hypothetical people being followed to measure the mortality rate of leukemia. A *mortality rate* is an incidence

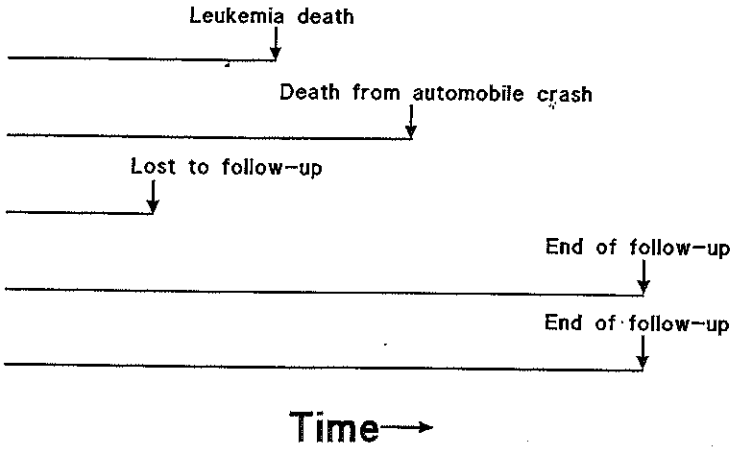

Figure 4–2 Time at risk for leukemia death for five people.

rate in which the event being measured is death. Only the first of the five people died of leukemia during the follow-up period. This person's time at risk ended with his or her death from leukemia. The second person died in an automobile crash, after which he or she was no longer at risk for dying of leukemia. The third person was lost to follow-up early during the follow-up period. After a person is lost, even if that person dies of leukemia, the death will not be counted in the numerator of the rate because the researcher would not know about it. Therefore the time at risk to be counted as a case in the numerator of the rate ends when a person becomes lost to follow-up. The last two people were followed for the complete follow-up period. The total time tallied in the denominator of the mortality rate for leukemia for these five people corresponds to the sum of the lengths of the five line segments in Figure 4–2.

Incidence rates treat one unit of time as equivalent to another, regardless of whether these time units come from the same person or from different people. The incidence rate is the ratio of cases to the total time at risk for disease. This ratio does not have the same simple interpretability as the risk measure.

A comparison of the risk and incidence rate measures (Table 4–1) shows that, whereas the incidence proportion, or risk, can be interpreted as a probability, the incidence rate cannot. Unlike a probability, the incidence rate does not have the range of [0,1]. Instead, it can theoretically become extremely large without numeric limit. It may at first seem puzzling that a measure of disease occurrence can exceed 1; how can more than 100% of a population be affected? The answer is that the incidence rate does not measure the proportion of the population that is affected. It measures the ratio of the number of cases to the time at risk for disease. Because the denominator is measured in time units, we can always imagine that the denominator of an incidence rate could be smaller, making the rate larger. The numeric value of the incidence rate depends on what time unit is chosen.

Suppose that we measure an incidence rate in a population as 47 cases occurring in 158 months. To make it clear that the time tallied in the denominator of an incidence rate is the sum of the time contribution from various people,

we often refer to these time values as *person-time*. We can express the incidence rate as

$$\frac{47 \text{ cases}}{158 \text{ person-months}} = \frac{0.30 \text{ cases}}{\text{person-month}}$$

We could also restate this same incidence rate using person-years instead of person-months:

$$\frac{47 \text{ cases}}{13.17 \text{ person-years}} = \frac{3.57 \text{ cases}}{\text{person-year}}$$

These two expressions measure the same incidence rate; the only difference is the time unit chosen to express the denominator. The different time units affect the numeric values. The situation is much the same as expressing speed in different units of time or distance. For example, 60 miles/hr is the same as 88 ft/sec or 26.84 m/sec. The change in units results in a change in the numeric value.

The analogy between incidence rate and speed is helpful in understanding other aspects of incidence rate as well. One important insight is that the incidence rate, like speed, is an instantaneous concept. Imagine driving along a highway. At any instant, you and your vehicle have a certain speed. The speed can change from moment to moment. The speedometer gives you a continuous measure of the current speed. Suppose that the speed is expressed in terms of kilometers per hour. Although the time unit for the denominator is 1 hour, it does not require an hour to measure the speed of the vehicle. You can observe the speed for a given instant from the speedometer, which continuously calculates the ratio of distance to time over a recent short interval of time. Similarly, an incidence rate is a momentary rate at which cases are occurring within a group of people. Measuring an incidence rate takes a nonzero amount of time, as does measuring speed, but the concepts of speed and incidence rate can be thought of as applying at a given instant. If an incidence rate is measured, as is often the case, with person-years in the denominator, the rate nevertheless may characterize only a short interval, rather than a year. Similarly, speed expressed in kilometers per hour does not necessarily apply to an hour but perhaps to an instant. It may seem impossible to get an instantaneous measure of incidence rate, but in a situation analogous to use of the speedometer, current incidence or mortality for a sufficiently large population can be measured by counting, for example, the cases occurring in 1 day and dividing that number by the person-time at risk during that day. Time units can be measured in days or hours but may be expressed in years by dividing by the number of days or hours in a year. The unit of time in the denominator of an incidence rate is arbitrary and has no implication for the period of time over which the rate is actually measured, nor does it communicate anything about the actual time to which it applies.

Incidence rates commonly are described as *annual incidence* and expressed in the form of "50 cases per 100,000." This is a clumsy description of an incidence rate, equivalent to describing an instantaneous speed as an "hourly distance." Nevertheless, we can translate this phrasing to correspond with what we have

already described for incidence rates. We can express this rate as 50 cases per 100,000 person-years, or 50/100,000 yr^{-1}. The negative 1 in the exponent means inverse, implying that the denominator of the fraction is measured in units of years.

Whereas the risk measure typically transmits a clear message to epidemiologists and nonepidemiologists alike (provided that a time period for the risk is specified), the incidence rate may not. It is more difficult to conceptualize a measure of occurrence that uses the ratio of events to the total time in which the events occur. Nevertheless, under certain conditions, there is an interpretation that we can give to an incidence rate. The dimensionality of an incidence rate is that of the reciprocal of time, which is another way of saying that in an incidence rate, the only units involved are time units, which appear in the denominator. Suppose we invert the incidence rate. Its reciprocal is measured in units of time. To what time does the reciprocal of an incidence rate correspond?

Under steady-state conditions—a situation in which the rates do not change with time—the reciprocal of the incidence rate equals the average time until an event occurs. This time is referred to as the *waiting time*. Take as an example the incidence rate described earlier, 3.57 cases per person-year. This rate can be written as 3.57 yr^{-1}; the cases in the numerator of an incidence rate do not have units. The reciprocal of this rate is 1/3.57 years = 0.28 years. This value can be interpreted as an average waiting time of 0.28 years until the occurrence of an event.

As another example, consider a mortality rate of 11 deaths per 1000 person-years, which could also be written as 11/1000 yr^{-1}. If this is the total mortality rate for an entire population, the waiting time that corresponds to it will represent the average time until death. The average time until death is also referred to as the *expectation of life* or *expected survival time*. Using the reciprocal of 11/1000 yr^{-1}, we obtain 90.9 years, which can be interpreted as the expectation of life for a population in a steady state that has a mortality rate of 11/1000 yr^{-1}. Because mortality rates typically change with time over the time scales that apply to this example, taking the reciprocal of the mortality rate for a population is not a practical method for estimating the expectation of life. Nevertheless, it is helpful to understand what kind of interpretation we may assign to an incidence rate or a mortality rate, even if the conditions that justify the interpretation are often not applicable.

Chicken and Egg

An old riddle asks, "If a chicken and one half lay an egg and one half in a day and one half, how many eggs does one chicken lay in 1 day?" This riddle is a rate problem. The question amounts to asking, "What is the rate of egg laying expressed in eggs per chicken-day?" To get the answer, we express the rate as the number of eggs in the numerator and the number of chicken-days in the denominator: 1.5 eggs/[(1.5 chickens) • (1.5 days)] = 1.5 eggs/2.25 chicken-days. This calculation gives a rate of 2/3 egg per chicken day.

The Relation Between Risk and Incidence Rate

Because the interpretation of risk is so much more straightforward than the interpretation of incidence rate, it is often convenient to convert incidence rate measures into risk measures. Fortunately, this conversion usually is not difficult. The simplest formula to convert an incidence rate to a risk is as follows:

$$\text{Risk} \approx \text{Incidence rate} \times \text{Time} \qquad [4\text{–}1]$$

For Equation 4–1 and other such formulas, it is a good habit to confirm that the dimensionality on both sides of the equation is equivalent. In this case, risk is a proportion, and therefore has no dimensions. Although risk applies for a specific period of time, the time period is a descriptor for the risk but not part of the measure itself. Risk has no units of time or any other quantity built in; it is interpreted as a probability. The right side of Equation 4–1 is the product of two quantities, one of which is measured in units of the reciprocal of time and the other of which is time itself. Because this product has no dimensionality, the equation holds as far as dimensionality is concerned.

In addition to checking the dimensionality, it is useful to check the range of the measures in an equation such as Equation 4–1. The risk is a pure number in the range [0,1]; values outside this range are not permitted. In contrast, incidence rate has a range of $[0,\infty]$, and time also has a range of $[0,\infty]$. The product of incidence rate and time does not have a range that is the same as risk, because the product can exceed 1. This analysis shows that Equation 4–1 is not applicable throughout the entire range of values for incidence rate and time. In general terms, Equation 4–1 is an approximation that works well as long as the risk calculated on the left is less than about 20%. Above that value, the approximation deteriorates.

For example, suppose that a population of 10,000 people experiences an incidence rate of lung cancer of 8 cases per 10,000 person-years. If we followed the population for 1 year, Equation 4–1 suggests that the risk of lung cancer is 8 in 10,000 for the 1-year period (ie, 8/10,000 person-years × 1 year), or 0.0008. If the same rate applied for only 0.5 year, the risk would be one half of 0.0008, or 0.0004. Equation 4–1 calculates risk as directly proportional to both the incidence rate and the time period, so as the time period is extended, the risk becomes proportionately greater.

Now suppose that we have a population of 1000 people who experience a mortality rate of 11 deaths per 1000 person-years for a 20-year period. Equation 4–1 predicts that the risk of death over 20 years will be 11/1000 yr^{-1} × 20 yr = 0.22, or 22%. In other words, Equation 4–1 predicts that among the 1000 people at the start of the follow-up period, there will be 220 deaths during the 20 years. The 220 deaths are the sum of 11 deaths that occur among 1000 people every year for 20 years. This calculation neglects the fact that the size of the population at risk shrinks gradually as deaths occur. If the shrinkage is taken into account, fewer than 220 deaths will have occurred at the end of 20 years.

Table 4–2 describes the number of deaths expected to occur during each year of the 20 years of follow-up if the mortality rate of 11/1000 yr^{-1} is applied to a population of 1000 people for 20 years. The table shows that at the end of

Table 4–2 NUMBER OF EXPECTED DEATHS OVER 20 YEARS AMONG 1000 PEOPLE WITH A MORTALITY RATE OF 11 DEATHS PER 1000 PERSON-YEARS

Year	Expected Number Alive at Start of Year	Expected Deaths	Cumulative Deaths
1	1000.000	10.940	10.940
2	989.060	10.820	21.760
3	978.240	10.702	32.461
4	967.539	10.585	43.046
5	956.954	10.469	53.515
6	946.485	10.354	63.869
7	936.131	10.241	74.110
8	925.890	10.129	84.239
9	915.761	10.018	94.257
10	905.743	9.909	104.166
11	895.834	9.800	113.966
12	886.034	9.693	123.659
13	876.341	9.587	133.246
14	866.754	9.482	142.728
15	857.272	9.378	152.106
16	847.894	9.276	161.382
17	838.618	9.174	170.556
18	829.444	9.074	179.630
19	820.370	8.975	188.605
20	811.395	8.876	197.481

20 years, about 197 deaths have occurred, rather than 220, because a steadily smaller population is at risk of death each year. The table also shows that the prediction of 11 deaths per year from Equation 4–1 is a good estimate for the early part of the follow-up but the number of deaths expected each year gradually becomes considerably lower than 11. Why is the number of expected deaths not quite 11 even for the first year, in which there are 1000 people being followed at the start of the year? As soon as the first death occurs, the number of people being followed is less than 1000, which influences the number of expected deaths in the first year. As is seen in Table 4–2, the expected deaths decline gradually throughout the period of follow-up.

If we extended the calculations in the table further, the discrepancy between the risk calculated from Equation 4–1 and the actual risk would grow. Figure 4–3 graphs the cumulative total of deaths that would be expected and the number projected from Equation 4–1 over 50 years of follow-up. Initially, the two curves are close, but as the cumulative risk of death rises, they diverge. The bottom curve in the figure is an exponential curve, related to the curve that describes *exponential decay*. If a population experiences a constant rate of death, the proportion remaining alive follows an exponential curve with time. This exponential decay is the same curve that describes radioactive decay. If a population of radioactive atoms converts from one atomic state to another at a constant rate, the proportion of atoms left in the initial state follows the curve of exponential decay. The lower

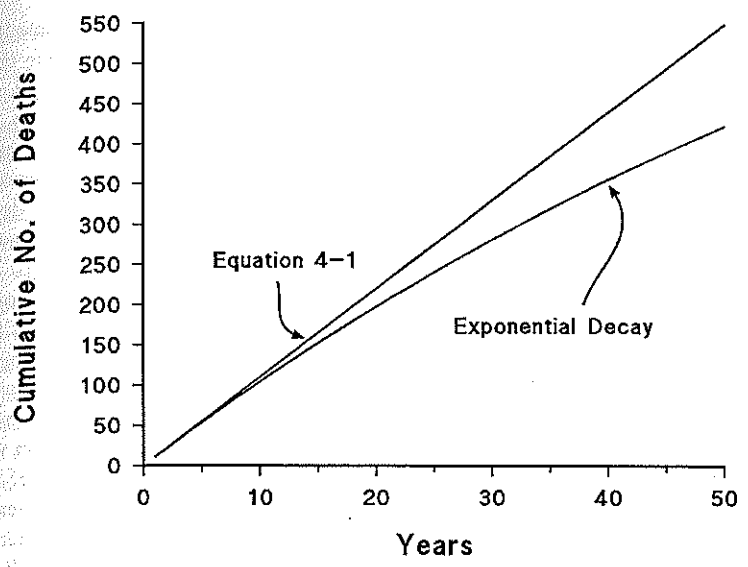

Figure 4–3 Cumulative number of deaths among 1000 people with a mortality rate of 11 deaths per 1000 person-years, presuming no population shrinkage (see Equation 4–1) and taking the population shrinkage into account (ie, exponential decay).

curve in Figure 4–3 is actually the complement of an exponential decay curve. Instead of showing the decreasing number remaining alive (ie, the curve of exponential decay), it shows the increasing number who have died, which is the total number in the population minus the number remaining alive. Given enough time, this curve gradually flattens, and the total number of deaths approaches the total number of people in the population. In contrast, the curve based on Equation 4–1 continues to predict 11 deaths each year regardless of how many people remain alive, and it eventually would predict a cumulative number of deaths that exceeds the original size of the population.

Clearly, Equation 4–1 cannot be used to calculate risks that are large, because it provides a poor approximation in such situations. For many epidemiologic applications, however, the calculated risks are reasonably small, and Equation 4–1 is quite adequate for converting incidence rates to risks.

Equation 4–1 calculates risk for a time period over which a single incidence rate applies. The calculation assumes that the incidence rate, an instantaneous concept, remains constant over the time period. What if the incidence rate changes with time, as is often the case? In that event, risk can still be calculated, but it should be calculated first for separate subintervals of the time period. Each of the time intervals should be short enough so that the incidence rate that applies to it could be considered approximately constant. The shorter the intervals, the better the overall accuracy of the risk calculation, although the intervals should not be so short that there are inadequate data to obtain meaningful incidence rates for each interval.

The method of calculating risks over a time period with changing incidence rates is known as *survival analysis*. It can also be applied to nonfatal risks, but the

approach originated from data related to deaths. The method is implemented by creating a table similar to Table 4–2, called a *life table*. The purpose of a life table is to calculate the probability of surviving through each successive time interval that constitutes the period of interest. The overall survival probability is equal to the cumulative product of the probabilities of surviving through each successive interval, and the overall risk of death is equal to 1 minus the overall probability of survival.

Table 4–3 is a simplified life table that enables calculation of the risk of dying of a motor vehicle injury in a hypothetical cohort of 100,000 people followed from birth through age 85.[2] In this example, the time periods correspond to age intervals. The number initially at risk has been arbitrarily set to 100,000 people. The life-table calculation is strictly hypothetical, because the number at risk at the start of each age group is reduced only by deaths from motor vehicle injury in the previous age group, ignoring all other causes of death. With this assumption that there are no competing risks, the results are interpretable as risks or survival probabilities that would result if the only risk faced by a population was the one under study. The risk of dying of a motor vehicle injury for each of the age intervals is calculated by taking the number of deaths in each age interval (column 3) and dividing it by the number who are at risk during that age interval (column 2). The survival probability in column 5 is equal to 1 minus the risk for that age category. The cumulative survival probability (column 6) is the product of the age-specific survival probabilities up to that age. The bottom number in column 6 is the probability of surviving to age 85 without dying of a motor vehicle injury, assuming that there are no competing risks (ie, assuming that without a motor vehicle injury, the person would survive to age 85).

Subtracting the final cumulative survival probability from 1 gives the total risk, from birth until the 85th birthday, of dying of a motor vehicle injury. This risk is $1 - 0.98378 = 1.6\%$. Because this calculation is based on the assumption that everyone will live to their 85th birthday except those who die of motor vehicle accidents, it overstates the actual proportion of people who will die in a motor vehicle accident before they reach age 85. Another assumption in the calculation is that these mortality rates, which have been gathered from a cross section of the population at a given time, can be applied to a group of people over the course of 85 years of life. If the mortality rates changed with time, the risk estimated from the life table would be inaccurate.

Table 4–3 LIFE TABLE FOR DEATH FROM MOTOR VEHICLE INJURY FROM BIRTH THROUGH AGE 85[a]

Age	Number at Risk	Deaths in Interval	Risk of Dying	Survival Probability	Cumulative Survival Probability
0–14	100,000	70	0.00070	0.99930	0.99930
15–24	99,930	358	0.00358	0.99642	0.99572
25–44	99,572	400	0.00402	0.99598	0.99172
45–64	99,172	365	0.00368	0.99632	0.98807
65–84	98,807	429	0.00434	0.99566	0.98378

[a]Mortality rates are deaths per 100,000 person-years.
Adapted from Iskrant and Joliet, Table 24.[2]

Table 4–3 shows a hypothetical cohort being followed for 85 years. If this had been an actual cohort, there would have been some people lost to follow-up and some who died of other causes. When follow-up is incomplete for either of these reasons, the usual approach is to use the information that is available for those with incomplete follow-up; their follow-up is described as *censored* at the time that they are lost or die of another cause.

Table 4–4 shows what the same cohort experience would look like under the more realistic situation in which many people have incomplete follow-up. Two new columns have been added with hypothetical data on the number that are censored because they were lost to follow-up or died of other causes (column 4) and the effective number at risk (column 5). The effective number at risk is calculated by taking the number at risk in column 2 and subtracting one half of the number who are censored (column 4). Subtracting one half of those who are censored is based on the assumption that the censoring occurred uniformly throughout each age interval. If there is reason to believe that the censoring tended to occur nonuniformly within the interval, the calculation of the effective number at risk should be adjusted to reflect that belief.

Point-Source and Propagated Epidemics

An *epidemic* is an unusually high occurrence of disease. The definition of *unusually high* depends on the circumstances, and there is no clear demarcation between an epidemic and a smaller fluctuation. The high occurrence may represent an increase in the occurrence of a disease that still occurs in the population in the absence of an epidemic, although less frequently than during the epidemic, or it may represent an *outbreak*, which is a sudden increase in the occurrence of a disease that is usually absent or nearly absent (Fig. 4–4).

If an epidemic stems from a single source of exposure to a causal agent, it is considered a *point-source epidemic*. Examples of point-source epidemics are food poisoning of restaurant patrons who have been served contaminated food and cancer occurrence among survivors of the atomic bomb blasts in Hiroshima

Table 4–4 LIFE TABLE FOR DEATH FROM MOTOR VEHICLE INJURY FROM BIRTH THROUGH AGE 85[a]

Age	At Risk	Motor Vehicle Injury Deaths in Interval	Lost to Follow-up or Died of Other Causes	Effective Number at Risk	Risk of Dying	Survival Probability	Cumulative Survival Probability
0–14	100,000	67	9,500	95,250	0.00070	0.99930	0.99930
15–24	90,433	301	12,500	84,183	0.00358	0.99642	0.99572
25–44	77,632	272	20,000	67,632	0.00402	0.99598	0.99172
45–64	57,360	156	30,000	42,360	0.00368	0.99632	0.98807
65–84	27,204	64	25,000	14,704	0.00435	0.99565	0.98377

[a]Mortality rates are deaths per 100,000 person-years.

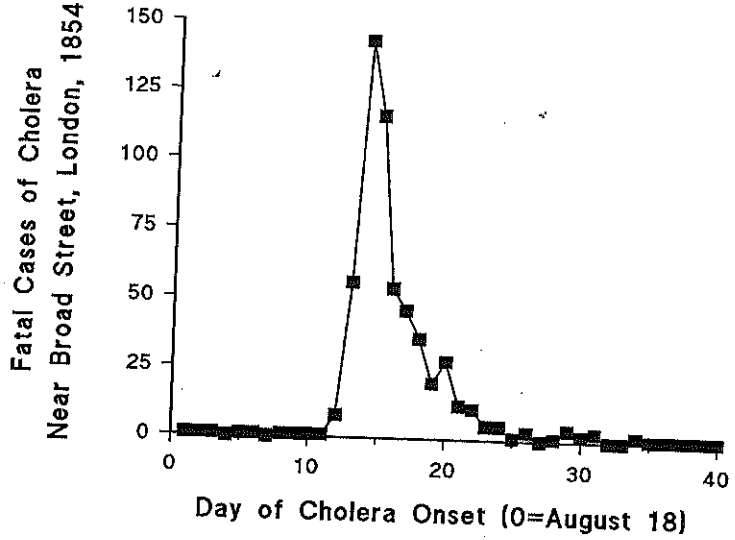

Figure 4–4 Epidemic curve for fatal cholera cases during the Broad Street outbreak in London in 1854.

and Nagasaki. Although the time scales of these epidemics differ dramatically, along with the nature of the diseases and their causes, all people in both cases were exposed to the same causal component that produced the epidemic—contaminated food in the restaurant or ionizing radiation from the bomb blast. The exposure in a point-source epidemic is typically newly introduced into the environment, thus accounting for the epidemic.

Typically, the shape of the epidemic curve for a point-source epidemic shows an initial steep increase in the incidence rate followed by a more gradual decline in the incidence rate; this pattern is often described as a log-normal distribution. The asymmetry of the curve stems partly from the fact that biologic curves with a meaningful zero point tend to be asymmetric because there is less variability in the direction of the zero point than in the other direction. For example, the distribution of recovery times for healing of a wound is log-normal. Similarly, the distribution of induction times until the occurrence of illness after a common exposure is log-normal. If the zero point is sufficiently far from the modal value, the asymmetry may not be apparent. For example, birth weight has a meaningful zero point, but the zero point is far from the center of the distribution, and the distribution is almost symmetric.

An example of an asymmetric epidemic curve is that of the 1854 cholera epidemic described by John Snow.[3] In that outbreak, exposure to contaminated water in the neighborhood of the water pump at Broad Street in London produced a log-normal epidemic curve (see Fig. 4–4). Snow is renowned for having convinced local authorities to remove the handle from the pump, but they only did so on September 8 (day 21), when the epidemic was well past its peak and the number of cases was almost back to zero.

The shape of an epidemic curve also may be affected by the way in which the curve is calculated. It is common, as in Figure 4–4, to plot the number of new cases instead of the incidence rate among susceptible people. People who have

already succumbed to an infectious disease may no longer be susceptible to it for some period of time. If a substantial proportion of a population is affected by the outbreak, the number of susceptible people will decline gradually as the epidemic progresses and the attack rate increases. This change in the susceptible population leads to a more rapid decline over time in the number of new cases compared with the incidence rate in the susceptible population. The incidence rate declines more slowly than the number of new cases because in the incidence rate, the declining number of new cases is divided by a dwindling amount of susceptible person-time.

A *propagated epidemic* is one in which the causal agent is transmitted through a population. Influenza epidemics are propagated by person-to-person transmission of the virus. The epidemic of lung cancer during the 20th century was a propagated epidemic attributable to the spread of tobacco smoking through many cultures and societies. The curve for a propagated epidemic tends to show a more gradual initial rise and a more symmetric shape than for a point-source epidemic because the causes spread gradually through the population. Transmission of infectious disease within a population is discussed further in Chapter 6, which also presents the Reed-Frost model, a simple model that describes transmission of an infectious disease in a closed population.

Although we may think of point-source epidemics as occurring over a short time span, they are not always briefer than propagated epidemics. The epidemic of cancer attributable to exposure to the atomic bombs detonated in Hiroshima and Nagasaki was a point-source epidemic that began a few years after the explosions and continues into the present. Another possible point-source epidemic that occurred over decades was an apparent outbreak of multiple sclerosis in the Faroe Islands that followed the occupation of those islands by British troops during the Second World War[4] (although this interpretation of the data has been questioned[5]). Propagated epidemics can occur over extremely short time spans. An example is epidemic hysteria, a disease often propagated from person to person in minutes. An example of an epidemic curve for a hysteria outbreak is depicted in Figure 4–5. In this epidemic, 210 elementary school children developed symptoms of headache, abdominal pain, and nausea. These symptoms were attributed by the investigators to hysteric anxiety.[6]

Prevalence Proportion

Incidence proportion and incidence rate are measures that assess the frequency of disease onsets. The numerator of either measure is the frequency of events that are defined as the occurrence of disease. In contrast, *prevalence proportion*, often referred to simply as *prevalence*, does not measure disease onset. Instead, it is a measure of disease status.

The simplest way of considering disease status is to consider disease as being either present or absent. The prevalence proportion is the proportion of people in a population who have disease. Consider a population of size N, and suppose that P individuals in the population have disease at a given time. The prevalence proportion is P/N. For example, suppose that among 10,000 women residents of a town on July 1, 2001, 1200 have hypertension. The prevalence proportion of hypertension among women in that town on that date is 1200/10,000 = 0.12,

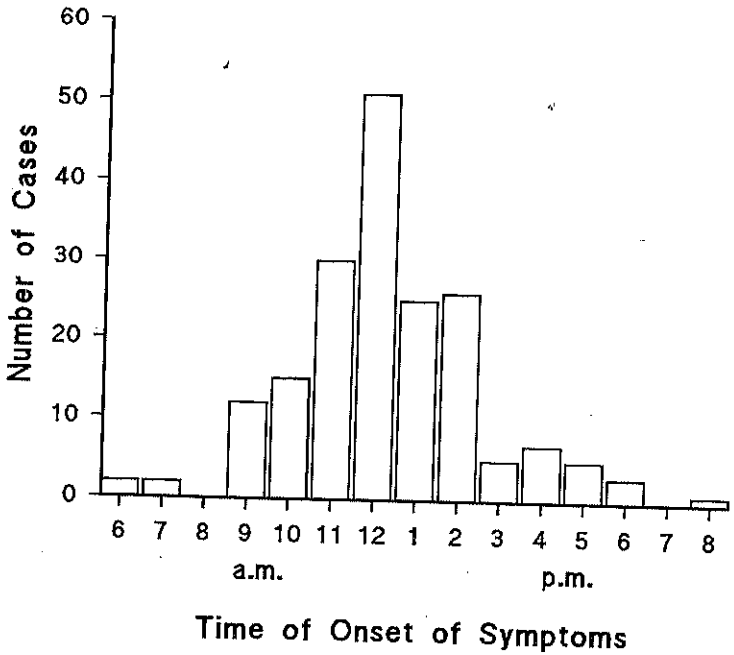

Figure 4–5 Epidemic curve for an outbreak of hysteria among elementary school children on November 6, 1985.

or 12%. This prevalence applies only to a single point in time, July 1, 2001. Prevalence can change with time as the factors that affect prevalence change.

What factors affect prevalence? Clearly, disease occurrence affects prevalence. The greater the incidence of disease, the more people there are who have it. Prevalence is also related to the length of time that a person has disease. The longer the duration of disease, the higher the prevalence. Diseases with short duration may have a low prevalence even if the incidence rate is high. One reason is that if the disease is benign, there may be a rapid recovery. For example, the prevalence of upper respiratory infection may be low despite a high incidence, because after a brief period, most people recover from the infection and are no longer in the disease state. Duration may also be short for a grave disease that leads to rapid death. The prevalence of aortic hemorrhage would be low even with a high incidence because it usually leads to death within minutes. The low prevalence means that, at any given moment, only an extremely small proportion of people are suffering from an aortic hemorrhage. Some diseases have a short duration because either recovery or death ensues promptly; appendicitis is an example. Other diseases have a long duration because, although a person cannot recover from them, they are compatible with a long survival time (although survival is often shorter than it would be without the disease). Diabetes, Crohn's disease, multiple sclerosis, parkinsonism, and glaucoma are examples.

Because prevalence reflects both incidence rate and disease duration, it is not as useful as incidence alone for studying the causes of disease. It is extremely

useful, however, for measuring the disease burden on a population, especially if those who have disease require specific medical attention. For example, the prevalent number of people in a population with end-stage renal disease predicts the need in that population for dialysis facilities.

In a *steady state*, which is the situation in which incidence rates and disease duration are stable over time, the prevalence proportion, P, has the following relation to the incidence rate:

$$\frac{P}{1-P} = I\bar{D} \qquad [4\text{–}2]$$

In Equation 4–2, I is the incidence rate and $\bar{D}$ is the average duration of disease. The quantity $P/(1-P)$ is known as the *prevalence odds*. In general, when a proportion, such as prevalence proportion, is divided by 1 minus the proportion, the resulting ratio is referred to as the *odds* for that proportion. If a horse is a 3-to-1 favorite at a racetrack, it means that the horse is thought to have a probability of winning of 0.75. The odds of the horse winning is $0.75/(1-0.75) = 3$, usually described as 3 to 1. Similarly, if a prevalence proportion is 0.75, the prevalence odds would be 3, and a prevalence of 0.20 would correspond to a prevalence odds of $0.20/(1-0.20) = 0.25$. For small prevalences, the value of the prevalence proportion and that of the prevalence odds are close because the denominator of the odds expression is close to 1. For small prevalences (eg, <0.1), we can rewrite Equation 4–2 as follows:

$$P \approx I\bar{D} \qquad [4\text{–}3]$$

Equation 4–3 indicates that, given a steady state and a low prevalence, prevalence is approximately equal to the product of the incidence rate and the mean duration of disease. Note that this relation does not hold for age-specific prevalences. In that case, $\bar{D}$ corresponds to the duration of time spent within that age category rather than the total duration of time with disease.

As we did earlier for risk and incidence rate, we should check this equation to make certain that the dimensionality and ranges of both sides of the equation are satisfied. For dimensionality, the right-hand sides of Equations 4–2 and 4–3 involve the product of a time measure, disease duration, and an incidence rate, which has units of reciprocal of time. The product is dimensionless, a pure number. Prevalence proportion, like risk or incidence proportion, is also dimensionless, which satisfies the dimensionality requirement for the two equations, 4–2 and 4–3. The range of incidence rate and that of mean duration of illness is $[0,\infty]$, because there is no upper limit to an incidence rate or the duration of disease. Therefore Equation 4–3 does not satisfy the range requirement, because the prevalence proportion on the left side of the equation, like any proportion, has a range of $[0,1]$. For this reason, Equation 4–3 is applicable only for small values of prevalence. The measure of prevalence odds in Equation 4–2, however, has a range of $[0,\infty]$, and it is applicable for all values, rather than just for small values of the prevalence proportion. We can rewrite Equation 4–2 to solve for the prevalence proportion as follows:

$$P = \frac{I\bar{D}}{1+I\bar{D}} \qquad [4\text{–}4]$$

Prevalence measures the disease burden in a population. This type of epidemiologic application relates more to administrative areas of public health than to causal research. Nevertheless, there are research areas in which prevalence measures are used more commonly than incidence measures, even to investigate causes. Examples are birth defects and birth-related phenomena such as birth weight or preterm birth. We use a prevalence measure when describing the occurrence of congenital malformations among liveborn infants in terms of the proportion of these infants who have a malformation. For example, the proportion of infants who are born alive with a defect of the ventricular septum of the heart is a prevalence. It measures the status of liveborn infants with respect to the presence or absence of a ventricular septal defect. Measuring the incidence rate or incidence proportion of ventricular septal defects would require ascertainment of a population of embryos who were at risk for developing the defect and measurement of the defect's occurrence among these embryos. Such data are usually not obtainable, because many pregnancies end before the pregnancy is detected, and the population of embryos is not readily identified. Even when a woman knows she is pregnant, if the pregnancy ends early, information about the pregnancy may never come to the attention of researchers. For these reasons, incidence measures for birth defects are uncommon. Prevalence at birth is easier to assess and often is used as a substitute for incidence measures. Although prevalence measures are easier to obtain, they have a drawback when used for causal research: Factors that increase prevalence may do so not by increasing the occurrence of the condition but by increasing the duration of the condition. For example, a factor associated with the prevalence of ventricular septal defect at birth could be a cause of ventricular septal defect, but it could also be a factor that does not cause the defect but instead enables embryos that develop the defect to survive until birth. On the other hand, there may be practical interest in understanding the factors that are related to being born alive with the defect.

Prevalence is sometimes used in research to measure diseases that have insidious onset, such as diabetes or multiple sclerosis. These are conditions for which it may be difficult to define onset, and it therefore may be necessary in some settings to describe the condition in terms of prevalence rather than incidence.

Prevalence of Characteristics

Because prevalence measures status, it is often used to describe the status of characteristics or conditions other than disease in a population. For example, the proportion of a population that engages in cigarette smoking often is described as the prevalence of smoking. The proportion of a population exposed to a given agent is often referred to as the exposure prevalence. Prevalence can be used to describe the proportion of people in a population who have brown eyes, type O blood, or an active driver's license. Because epidemiology relates many individual and population characteristics to disease occurrence, it often employs prevalence measures to describe the frequency of these characteristics.

MEASURES OF CAUSAL EFFECTS

A central objective of epidemiologic research is to study the causes of disease. How should we measure the effect of exposure to determine whether exposure causes disease? In a courtroom, experts are asked to opine whether the disease of a given patient has been caused by a specific exposure. This approach of assigning causation in a single person is radically different from the epidemiologic approach, which does not attempt to attribute causation in any individual instance. The epidemiologic approach is to evaluate the proposition that the exposure is a cause of the disease in a theoretical sense, rather than in a specific person.

An elementary but essential principle to keep in mind is that a person may be exposed to an agent and then develop disease without there being any causal connection between the exposure and the disease. For this reason, we cannot consider the incidence proportion or the incidence rate among exposed people to measure a causal effect. For example, if a vaccine does not confer perfect immunity, some vaccinated people will get the disease that the vaccine is intended to prevent. The occurrence of disease among vaccinated people is not a sign that the vaccine is causing the disease, because the disease will occur even more frequently among unvaccinated people. It is merely a sign that the vaccine is not a perfect preventive. To measure a causal effect, we have to contrast the experience of exposed people with what would have happened in the absence of exposure.

The Counterfactual Ideal

It is useful to consider how to measure causal effects in an ideal way. People differ from one another in myriad ways. If we compare risks or incidence rates between exposed and unexposed people, we cannot be certain that the differences in risks or rates are attributable to the exposure. They could be attributable to other factors that differ between exposed and unexposed people. We may be able to measure and to take into account some of these factors, but others may elude us, hindering any definite inference. Even if we matched people who were exposed with similar people who were not exposed, they could still differ in inapparent ways. The ideal comparison would be the result of a thought experiment: the comparison of people with themselves, followed through time simultaneously in both an exposed and an unexposed state. Such a comparison envisions the impossible, because it requires each person to exist in two incarnations: one exposed and the other unexposed. If such an impossible goal were achievable, it would allow us to know the effect of exposure, because the only difference between the two settings would be the exposure. Because this situation is impossible, it is called *counterfactual.*

The counterfactual goal posits not only a comparison of a person with himself or herself but also a repetition of the experience during the same time period. Some studies do pair the experiences of a person under both exposed and unexposed conditions. The experimental version of such a study is called a *crossover study,* because the study subject crosses over from one study group to the other after a period of time. Although crossover studies come close to the ideal of a counterfactual comparison, they do not achieve it because a person can be in only