3

Measuring Disease Occurrence and Causal Effects

Rothman 2002

As in most sciences, measurement is a central feature of epidemiology. Epidemiology has been defined as the study of the occurrence of illness.[1] The broad scope of epidemiology today 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, we discuss several measures of disease frequency: *risk*, *incidence rate*, and *prevalence*. We 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 readily understood by many people. It is measured on the same scale and interpreted in the same way as a probability. In epidemiology, we often speak about risk applying to an individual, in which case we are describing the probability that a person will develop a given disease. It is usually pointless, however, to measure risk in a single person, since for most diseases we would say that the person either did or did not get the disease. Among a larger group of people, we could describe the proportion who developed the disease. If a population has N people and A people out 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 period during which the risk is being measured. The average

risk in a group is also referred to as the *incidence proportion*. Often the word *risk* is used in reference to a single person and *incidence proportion* is used in reference to a group of people. Because we use averages taken from populations to estimate the risk experienced by individuals, we often use the two terms 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 of age have a 2% risk of dying from 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 from 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 of time 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 the time period over which the risk applies. This time 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 3–1 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 might occur if a person who is susceptible to an infectious disease becomes immunized, in which case the leveling off of risk would not be gradual but sudden. Another way that a pattern like A might occur is if those who come into contact with a susceptible person become immunized, reducing the person's risk of acquiring the disease. A pattern similar to B might occur if a person has been exposed to a cause and is nearing the end of the typical induction time for the causal

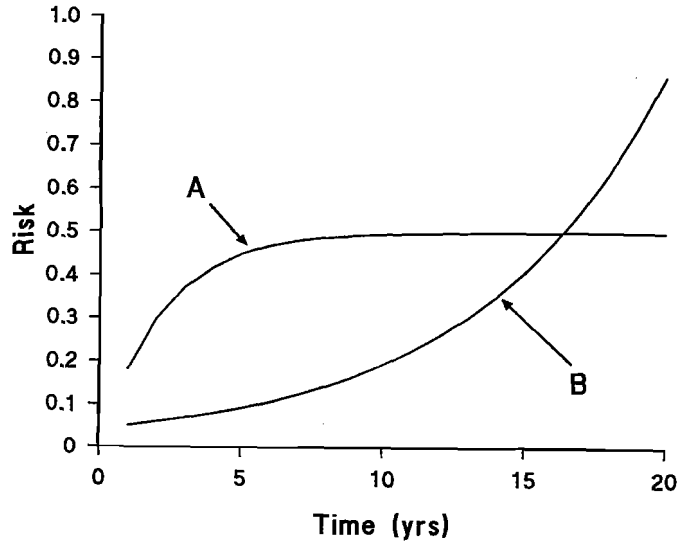

Figure 3–1. *Two possible patterns of disease risk with time.*

action, such as the risk of adenocarcinoma of the vagina among young women who were exposed to diethylstilbestrol while they were fetuses, discussed in Chapter 2. Another situation that can give rise to pattern B is simply the aging process, which often leads to sharply increasing risks as people progress beyond middle age.

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: if a population is followed over a period of 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 of the 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, and 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 end up with an underestimate of the risk of domestic violence in a population of women who do survive 30 years. To see why, imagine that there are many women who do not survive the 30-year follow-up period. It is likely that among them there would be some who would have experienced domestic violence if they

had instead survived. Thus, if we count these 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 followed for a full 30 years. In contrast, if we happen to be studying the risk of death from any cause, there would be no possibility of anyone dying from a cause that we were not measuring. Nevertheless, outside of studying death from any cause, 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.

This phenomenon of people being removed from a study through death from other causes is sometimes referred to as *competing risks*. Over a short period of time, the influence of competing risks is generally small, and it is not unusual for studies to ignore competing risks if the follow-up 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 efficacy. Because only a small proportion of school-age children died from 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 of time, competing risks are greater and may need to be taken into account.

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 losses to follow-up is sometimes similar to that 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 in the sense that it usually cannot be directly observed in a population. If competing risks did not occur and we could avoid all losses to follow-up, we could measure incidence proportion directly by dividing the number of observed cases by the number of people in the population followed. As mentioned above, if we study death from any cause as our outcome, there will be no competing risk; any death that occurs will count in the numerator of the risk measure. Most attempts to measure disease risk, however, are aimed at more specific outcomes or at disease onset rather than death. For such outcomes, there will always be competing risks. If one chooses to report the fraction A/N, which is the observed number of cases divided by the number of people who were initially being followed, it will underesti-

mate the incidence proportion that would have been observed if there had been no competing risk.

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 simply the incidence proportion, or risk, of becoming afflicted with a condition during an epidemic period. For example, we might speak of an influenza epidemic with an attack rate of 10%, which means that 10% of the population developed the disease during the epidemic period. The time reference for an attack rate is usually not stated but implied by the biology of the disease being described. It is seldom measured in periods of more than a few months. 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.

Another version of the incidence proportion that is encountered frequently in clinical medicine is the *case fatality rate*. The case fatality rate is the proportion of people, among those who develop a disease, who then proceed to die from the disease. 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 just deaths from the disease, are counted). The case fatality rate is seldom accompanied by a specific time referent, which sometimes makes it difficult to interpret. It is typically used, and easiest to interpret, as a description of the proportion of people who succumb from an infectious disease, such as measles. The case fatality rate for measles in the United States is about 1.5 per 1000 cases. The time period for this risk of death is the comparatively short time frame in which measles infects an individual, ending in either recovery, death, or some other complication. For diseases that continue to affect a person over long periods of time, 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, however, is different. Instead of dividing the number of cases by the number of people who were initially being followed, we divide 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 is being followed for 30 years and a given person dies after 5 years of follow-up, then that person would contribute only 5 years to the sum for the group. Others might contribute more or fewer years, up to a maximum of the full 30 years of follow-up.

There are two methods of counting the time of an individual who develops the disease being measured. 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. As a result, the numerator of an incidence rate could contain more than one occurrence of an upper respiratory tract infection from a single person. The denominator, then, should include all of the time that each person was at risk of 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 is death from leukemia, a person can be counted as a case only once. For someone who dies from 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 could occur more than once in an individual but count only the first occurrence of the event. For example, researchers might count the occurrence of the first heart attack in an individual and ignore (or study separately) second or later heart attacks. Whenever 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 how to tally 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 of disease*. The time in the denominator of an incidence rate should include every moment in 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, once a person experiences the event, that per-

son will have no more time at risk for disease, so the follow-up ends with the disease occurrence. The same is true of a person who dies from a competing risk.

The following diagram illustrates the time at risk for five hypothetical people being followed to measure the mortality rate of leukemia. (A *mortality rate* is an incidence rate in which the event being measured is death.) Only the first of the five people died from leukemia during the follow-up period. This person's time at risk ended with his or her death from leukemia. The second person died from another cause, an automobile crash, after which he or she was no longer at risk of dying from leukemia. The third person was lost to follow-up early during the follow-up period. Once a person is lost, if that person dies from leukemia, the death cannot be counted in the numerator of the rate because the researcher will 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 period of follow-up. The total time that would be tallied in the denominator of the mortality rate for leukemia for these five people would correspond to the sum of the lengths of the five line segments in Figure 3–2.

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 measure is the ratio of cases to the total time at risk of disease. This ratio does not have the same simple interpretability as the risk measure. Let us compare the risk and incidence rate measures to see how they differ.

Whereas the incidence proportion, or risk, measure can be interpreted as a probability, the incidence rate cannot. First of all, unlike a proba-

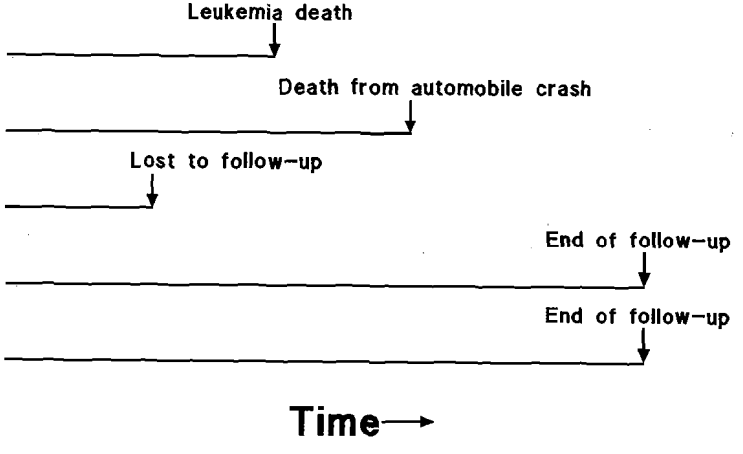

Figure 3–2. Time at risk for leukemia death for five people.

Table 3–1. 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

bility, the incidence rate does not even have the range of [0,1]. Instead, it can theoretically become as great as infinity. At first, it may seem puzzling that a measure of disease occurrence can exceed 1; after all, how can more than 100% of a population be affected? The answer is simply 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. In fact, the numerical 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*. Accordingly, we might restate the preceding incidence rate as follows.

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

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

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

The above 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 numerical values. The situation is much the same as expressing speed in different units of time or distance. For example, 60 miles/hour is the same as 88 feet/second or 26.84 meters/second. The change in units results in a change in the numerical value. The analogy between incidence rate and speed is helpful in understanding other aspects of incidence rate. One important concept is that 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 speedome-

ter gives you a continuous measure of the current speed. Suppose that the speed is expressed in terms of kilometers/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 note the speed for a given instant from the speedometer (which continuously calculates the ratio of distance to time over a recent finite short interval of time). Similarly, an incidence rate is the momentary rate at which cases are occurring within a group of people. To measure an incidence rate takes a finite amount of time, just as it does to measure speed; but the concepts of speed and incidence rate can be thought of as applying at a given instant. Thus, if an incidence rate is measured, as is often the case, with person-years in the denominator, the rate nevertheless might apply to an instant rather than to a year. Similarly, speed expressed in kilometers/hour does not necessarily apply to an hour but perhaps to an instant. The point is that for both measures, the unit of time in the denominator is arbitrary and has no implication for any period of time over which the rate is measured or applies.

One commonly finds incidence rates expressed in the form of 50 cases per 100,000 and described as "annual incidence." This is a clumsy description of an incidence rate, equivalent to describing an instantaneous speed in terms of an "hourly distance." Nevertheless, we can translate this phrasing to correspond with what we have already described for incidence rates. We could express this rate as 50 cases per 100,000 person-years or, equivalently, $50/100{,}000\ \text{yr}^{-1}$. (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 has a clear interpretation for epidemiologists and non-epidemiologists alike (provided that a time period for the risk is specified), incidence rate does not appear to have a clear interpretation. It is difficult to conceptualize a measure of occurrence that takes 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 just 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 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 above of 3.57 cases per person-year. Let us write this rate as $3.57\ \text{yr}^{-1}$. (The cases in the numerator of an incidence rate do not have any units.) If we take the reciprocal of this rate, we obtain 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 the first event that the rate measures. As another example, consider a mortality rate of 11 deaths per 1000 person-years, which we could also write as $11/1000\ \text{yr}^{-1}$. If this is the total mortality rate for an entire population, then the waiting time that corresponds to it would represent the average time until death. The average time until death is also referred to as the "expectation of life," or expected survival time. If we take the reciprocal of $11/1000\ \text{yr}^{-1}$, we obtain 90.9 years, which would be interpretable as the expectation of life for a population in a steady state that had a mortality rate of $11/1000\ \text{yr}^{-1}$. Unfortunately, mortality rates typically change with time over the time scales that apply to this example. Consequently, 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 might 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 lays an egg and one-half in a day and one-half, then how many eggs does one chicken lay in one 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, so we have 1.5 eggs/(1.5 chickens·1.5 days) = 1.5 eggs/2.25 chicken-days. This calculation gives a rate of ⅔ egg per chicken-day, so the answer to the riddle is ⅔.

Relation Between Risk and Incidence Rate

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

$$\text{Risk} \doteq \text{Incidence rate} \times \text{Time} \qquad (3\text{–}1)$$

It is a good habit when applying an equation such as 3–1 to check the dimensionality of each expression and make certain that both sides of the equation are equivalent. In this case, risk is measured as a proportion and 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, but is interpreted as a probability. The right side of equation 3–1 is the

product of two quantities, one of which is measured in units of the reciprocal of time and the other of which is simply time itself. This product has no dimensionality either, so 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 3–1. Note that 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,∞], and time has a range of [0,∞] as well. Therefore, the product of incidence rate and time will not have a range that is the same as risk; the product can easily exceed 1. This analysis tells us that equation 3–1 is not applicable throughout the entire range of values for incidence rate and time. In more general terms, equation 3–1 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 worsens.

Let us consider an example of how this equation works. Suppose that we have a population of 10,000 people who experience an incidence rate of lung cancer of 8 cases per 10,000 person-years. If we followed the population for 1 year, equation 3–1 tells us that the risk of lung cancer would be 8 in 10,000 for the 1-year period (the product of 8/10,000 person-years and 1 year), or 0.0008. If the same rate were experienced for only half a year, then the risk would be half of 0.0008, or 0.0004. Equation 3–1 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 3–1 predicts that the risk of death over 20 years would be $11/1000\ \text{yr}^{-1} \times 20\ \text{yr} = 0.22$, or 22%. In other words, equation 3–1 predicts that among the 1000 people at the start of the follow-up, 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 of death shrinks gradually as deaths occur. If we took the shrinkage into account, we would not end up with 220 deaths at the end of 20 years, but fewer.

Table 3–2 describes how many deaths would be expected to occur during each year of the 20 years of follow-up if the mortality rate of $11/1000\ \text{yr}^{-1}$ were applied to a population of 1000 people for 20 years. The table shows that at the end of 20 years we would actually expect about 197 deaths 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 3–1 is a good estimate for the early part of the follow-up, but that gradually the number of deaths expected becomes considerably lower than the estimate. Why is the number of expected deaths not quite 11 even for the first year, in which there are

Table 3–2. Number of expected deaths over 20 years among 1000 people experiencing 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

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, and the number of expected deaths is consequently influenced. As seen in Table 3–2, 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 3–1 and the expected risk would grow. Figure 3–3 graphs the cumulative total of deaths that would be expected and the number projected from equation 3–1 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. Strictly speaking, the lower curve in Figure 3–3 is the complement of an exponential decay curve. Instead of show-

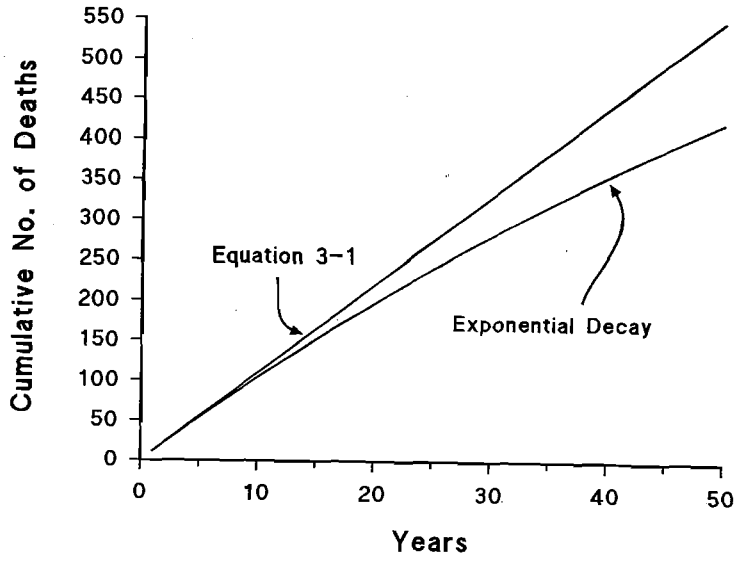

Figure 3–3. Cumulative number of deaths in 1000 people experiencing a mortality rate of 11 deaths per 1000 person-years, presuming no population shrinkage (*equation 3–1*) and taking the population shrinkage into account (*exponential decay*).

ing the decreasing number remaining alive (which would be 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 out so that the total number of deaths approaches the total number of people in the population. The curve based on equation 3–1, in contrast, continues to predict 11 more deaths each year regardless of how many people remain alive, and eventually it would predict a cumulative number of deaths that exceeds the original size of the population.

Clearly, we cannot use equation 3–1 to calculate risks that are large because it is a poor approximation in such situations. For many epidemiologic applications, however, the calculated risks are reasonably small and equation 3–1 is perfectly adequate for calculating risks from incidence rates.

Equation 3–1 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 would often be the case? In that event, one can still calculate risk, but separately for subintervals of the time period. Each of the time intervals should be short enough so that the incidence rate applied to it can be considered approximately constant. The shorter the intervals, the better the overall accuracy of the risk calculation. On the other hand, it is impractical to use many short

intervals unless there are adequate 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 be applied to nonfatal risks as well as to death but the approach originated from data that related to deaths. To implement the method, one creates a table similar to Table 3–2, 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 is equal to 1 minus the overall probability of survival.

Table 3–3 is a simplified life-table that enables us to calculate the risk of dying from a motor-vehicle injury,[2] based on applying the mortality rates to a hypothetical group of 100,000 people followed from birth through age 85. In this example, the time periods correspond to age intervals. As is often true of life-table calculations, it is assumed that there is no competing risk from other causes. The number initially at risk has been arbitrarily set at 100,000 people. Mortality rates are then used to calculate how many deaths occur among those remaining at risk in each age interval. This calculation is strictly hypothetical because the number at risk is reduced only by deaths from motor-vehicle injury. All other causes of death are ignored. The risk for each age interval can be calculated by applying the mortality rate to the time interval. The number of deaths and the number remaining at risk are not needed for this calculation but are included to show how the initial group would shrink slightly as some people are lost to fatal motor-vehicle accidents. The complement for the risk in each age category is the survival probability, calculated as 1 minus the risk. The cumulative product of the survival probabilities for each successive age category is the overall probability of surviving from birth through that age interval without dying from a

Table 3–3. Life-table for death from motor-vehicle injury from birth through age 85 (mortality rates are deaths per 100,000 person-years)

Age (years)	Mortality Rate	At Risk	Deaths in Interval	Risk	Survival Probability	Cumulative Survival Probability
0–14	4.7	100,000	70.5	0.000705	0.999295	0.999295
15–24	35.9	99,930	358.1	0.003584	0.996416	0.995714
25–44	20.1	99,571	399.5	0.004012	0.995988	0.991719
45–64	18.4	99,172	364.3	0.003673	0.996327	0.988077
65–84	21.7	98,808	427.9	0.004331	0.995669	0.983798

Adapted from Iskrant and Joliet, table 24[2]

motor-vehicle accident. Because all other causes of death have been ignored, this survival probability is conditional on the absence of competing risks. If we subtract the final cumulative survival probability from 1, we obtain the total risk, from birth until the 85th birthday, of dying from a motor-vehicle accident. This risk is 1 − 0.9838 = 1.6%. It assumes that everyone will live to the 85th birthday if not for the occurrence of motor-vehicle accidents, so it overstates the actual proportion of people who will die in a motor-vehicle accident before they reach age 85. Another assumption is that these mortality rates, which have been gathered from a cross-section of the population at a given time, would apply 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.

Because the overall risk of motor-vehicle death calculated from the rates in Table 3–3 is low, a simpler approach would have worked nearly as well. The simpler method applies equation 3–1 repeatedly to each age group, without subtracting the deaths from the total population at risk.

Risk from birth until age 85 of dying from a motor-vehicle injury =

$$\frac{4.7}{100{,}000\ \text{yr}}(15\ \text{yr}) + \frac{35.9}{100{,}000\ \text{yr}}(10\ \text{yr}) + \frac{20.1}{100{,}000\ \text{yr}}(20\ \text{yr}) + \frac{18.4}{100{,}000\ \text{yr}}(20\ \text{yr}) + \frac{21.7}{100{,}000\ \text{yr}}(20\ \text{yr})$$

$$= \frac{4.7(15) + 35.9(10) + 20.1(20) + 18.4(20) + 21.7(20)}{100{,}000} = 1.6\%$$

This result is same as the one obtained using a life-table approach. This method is often used to estimate lifetime risks for many diseases, such as suicide, cancer, or heart disease.

Point-Source and Propagated Epidemics

An *epidemic* is an unusually high occurrence of disease. The definition of "unusually high" may differ depending on the circumstances, so there is no clear demarcation between an epidemic and a smaller fluctuation. Furthermore, the high occurrence could 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. 3–4).

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 would be food poisoning of restaurant patrons who had been served contaminated food, or cancer among survivors of the atomic

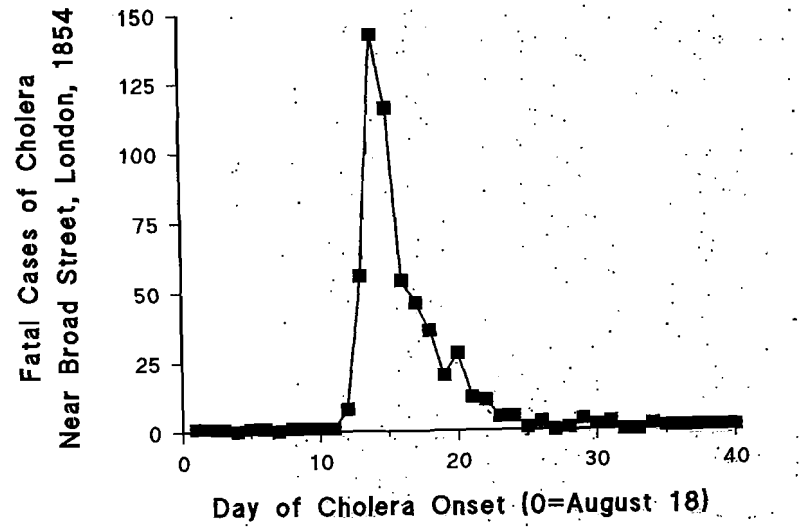

Figure 3–4. Epidemic curve of fatal cholera cases during the Broad Street outbreak, London 1854.[3]

bomb blasts in Hiroshima and Nagasaki. Although the time scales of these epidemics differ dramatically along with the nature of the diseases and their causes, both have in common that all people would have been exposed to the same causal component that produced the epidemic, either the contaminated food in the restaurant or the 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 of a point-source epidemic shows an initial steep increase in the incidence rate followed by a more gradual decline (often described as log-normal). The asymmetry of the curve stems partly from the fact that biologic curves with a meaningful zero point tend to be asymmetrical because there is less variability in the direction of the zero point than in the other direction. (If the zero point is sufficiently far from the modal value, the asymmetry may not be apparent, as in the distribution of birth weights.) For example, the distribution of recovery times for a wound to heal will be log-normal. Similarly, the distribution of induction times until the occurrence of illness after a common exposure will be log-normal.

An example of an asymmetrical 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 (Fig. 3–4). Snow is renowned for having convinced local authorities to remove the handle from the pump, but they did so only on September 8, when the

epidemic was well past its peak and the number of cases was already declining.

Another factor that may affect the shape of an epidemic curve is the way in which the curve is calculated. It is common, as in Figure 3–4, 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 will lead to a more rapid decline over time in the number of new cases than in the incidence rate. The incidence rate will decline 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 itself transmitted through a population. Influenza epidemics are propagated by person-to-person transmission of the virus. The epidemic of lung cancer during the twentieth 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 symmetrical shape than that for a point-source epidemic because the causes spread gradually through the population.

Although we may think of point-source epidemics as occurring over a short time span, they do not always occur over shorter time spans than propagated epidemics. The epidemic of cancer attributable to exposure to the atomic bombs 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, which followed the occupation of those islands by British troops during the Second World War.[4] Furthermore, propagated epidemics can occur over extremely short time spans. One 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 3–5. 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.[5]

Prevalence Proportion

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

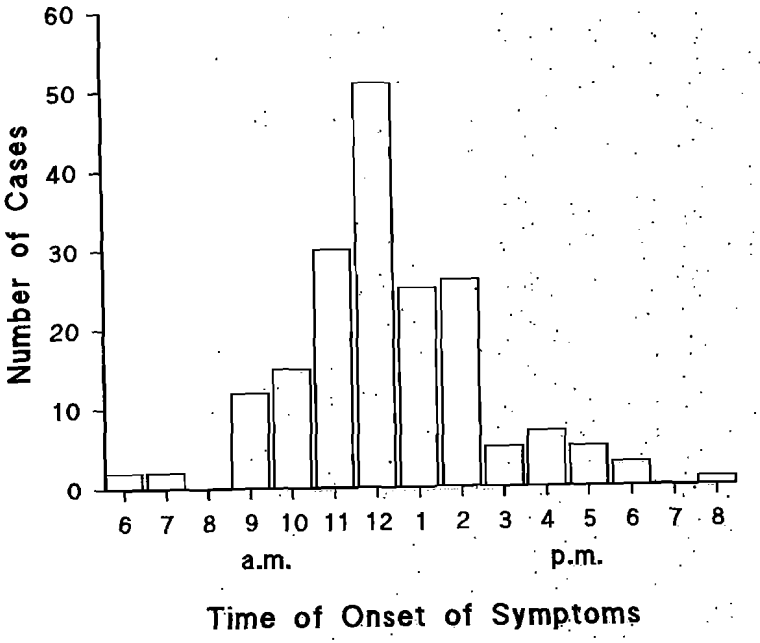

Figure 3–5. Epidemic curve of an outbreak of hysteria in elementary school children, November 6, 1985.

The simplest way of considering disease status is to consider disease either present or absent. The prevalence proportion is the proportion of people in a population that has disease. Consider a population of size N, and suppose that P individuals in the population have disease at a given time. The prevalence proportion will be P/N. For example, suppose that among 10,000 female residents of a town on July 1, 2001, 1200 have hypertension. The prevalence proportion of hypertension among women in the town on that date is 1200/10,000 = 0.12, or 12%. This prevalence applies only to the 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 will have it. But prevalence is also related to the length of time that a person has disease. The longer the duration of disease once it occurs, 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. Thus, 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. Thus, the prevalence of aortic hemorrhage would

be low even if it had a high incidence, because it generally leads to death within minutes. What the low prevalence means is that at any given moment, there will be only an extremely small proportion of people who are at that moment 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 one cannot recover from them, but they are compatible with a long survival (although the 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 for studying the causes of disease. It is extremely useful, however, for measuring the disease burden on a population, especially if those who have the 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 a situation in which incidence rates and disease duration are stable over time, the prevalence proportion, P, will have the following relation to the incidence rate.

$$\frac{P}{1-P} = I\bar{D} \qquad (3\text{–}2)$$

In equation 3–2, 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, whenever we take a proportion, such as prevalence proportion, and divide it 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 race track, 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 the prevalence odds will be close because the denominator of the odds expression will be close to 1. Therefore, for small prevalences, say less than 0.1, we could rewrite equation 3–2 as follows.

$$P \doteq I\bar{D} \qquad (3\text{–}3)$$

Equation 3–3 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.

As we did earlier for risk and incidence rate, it is useful to check the equation to make certain that the dimensionality and ranges of both sides are satisfied. For dimensionality, we find that the right-hand side of equations 3–2 and 3–3 involves the product of a time measure, disease duration, with incidence rate, which has units that are the reciprocal of time. The product is dimensionless, a pure number. Prevalence proportion, like risk or incidence proportion, is also dimensionless, which satisfies the dimensionality requirements for both equations 3–2 and 3–3. The range of incidence rates and mean durations of illness, however, is $[0,\infty]$, because there is no upper limit to either. Equation 3–3 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]. That is the reason that equation 3–3 is applicable only to small values of prevalence. The prevalence odds in equation 3–2, however, has a range of $[0,\infty]$, and is applicable for all values rather than just for small values of the prevalence proportion. We can rewrite equation 3–2 to solve for the prevalence proportion as follows.

$$P = \frac{I\bar{D}}{1 + I\bar{D}} \qquad (3\text{–}4)$$

As mentioned above, prevalence is used to measure 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. One of these is the area of birth defects. When we describe the occurrence of congenital malformations among live-born infants in terms of the proportion of these infants who have a malformation, we use a prevalence measure. 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 live-born infants with respect to the presence or absence of a ventricular septal defect. To measure the incidence rate or incidence proportion of ventricular septal defects would require the ascertainment of a population of embryos who were at risk to develop 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, so 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 used as a substitute for incidence measures. Although easier to obtain, prevalence measures 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. Thus, 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.

Prevalence is also 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 would often be 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 could be used to describe the proportion of people in a population with 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. Rather, 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 that epidemiologists must keep in mind is that a person may be exposed to an agent and then develop disease without there being any causal connection between exposure and disease. For this reason, we cannot consider the incidence proportion or the incidence rate among exposed people to measure a causal effect. Indeed, there might be no effect or even a preventive effect of exposure. For example, if a vaccine does not confer perfect immunity, then 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 we might 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 risk or rate are attributable to the exposure. Instead, they could be attributable to other factors that differ between exposed and unexposed people. We may be able to measure and take into account some of these other 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 might still differ in unapparent ways. The ideal comparison would be of people with themselves in both an exposed and an unexposed state. Such a comparison envisions the impossible goal of matching each person with himself or herself, being exposed in one incarnation and unexposed in the other. 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 not realistic, 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. That is, some studies actually do pair the experiences of a person under both exposed and unexposed conditions. The experimental version of such studies 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 one study group at a given time. The time sequence may affect the interpretation, and the passage of time means that the two experiences may differ by factors other than the exposure. Thus, the counterfactual setting is truly impossible, as it implies that a person relives the same experience twice, once with exposure and once without.

In the theoretical ideal of a counterfactual study, each exposed person would be compared with his or her unexposed counterfactual experience. The incidence proportion among exposed people could be compared with the incidence proportion among the counterfactual unexposed. Any difference in these proportions would have to be an effect of exposure. Suppose we observed 100 exposed people and found that in 1 year 25 developed disease, for an incidence proportion of 0.25. We

would theoretically like to compare this experience with the counterfactual, unobservable experience of the same 100 people going through the same year under the same conditions, except for their being unexposed. Suppose that in those conditions 10 developed disease. Then the incidence proportion for comparison would be 0.10. The difference, 15 cases in 100 during the year, or 0.15, would be a measure of the causal effect of the exposure.

Effect Measures

Because we can never achieve the counterfactual ideal, we strive to come as close as possible in the design of epidemiologic studies. Instead of comparing the experience of an exposed group with its counterfactual ideal, we must compare that experience with that of a real unexposed population. The goal is to find an unexposed population that would give a result close, if not identical, to that from a counterfactual comparison.

Suppose we consider the same 100 exposed people mentioned above, among whom 25 get the disease in 1 year. As a substitute for their missing counterfactual experience, we seek the experience of 100 unexposed persons who can provide an estimate of what would have occurred among the exposed had they not been exposed. This substitution is the crucial concern in many epidemiologic studies: does the experience of the unexposed group actually simulate what would have happened to the exposed group had they been unexposed? If we observe 10 cases of disease in the unexposed group, how can we know that the difference between the 25 cases in the exposed group and the 10 in the unexposed group is attributable to the exposure? Perhaps the exposure has no effect, but the unexposed group is at a lower risk for disease than the exposed group. What if we had observed 25 cases in both the exposed and the unexposed groups? The exposure might have no effect, but it might also have a strong effect that is balanced by the fact that the unexposed group has a higher risk for disease.

To achieve a valid substitution for the counterfactual experience, we resort to various design methods that promote comparability. The crossover study is one example, which promotes comparability by comparing the experience of each exposed person to himself or herself at a different time. This approach will be feasible only if the exposure can be studied in an experimental setting and if it has a brief effect. Another approach is a randomized experiment. In these studies, all participants are randomly assigned to the exposure groups. Given enough randomized participants, we can expect the distributions of other characteristics in the exposed and unexposed groups to be similar. Other approaches might involve choosing unexposed study subjects who have the same or similar risk-factor profiles for disease as the exposed subjects. However the comparability is achieved, its success is the overriding concern for any epidemiologic study that aims at evaluating a causal effect.

If we can assume that the exposed and unexposed groups are otherwise comparable with regard to risk for disease, we can compare measures of disease occurrence to assess the effect of the exposure. The two most commonly compared measures are the incidence proportion, or risk, and the incidence rate. The *risk difference* would be the difference in incidence proportion or risk between the exposed and unexposed groups. If the incidence proportion is 0.25 for the exposed and 0.10 for the unexposed, then the risk difference would be 0.15. With an incidence rate instead of a risk to measure disease occurrence, we can likewise calculate the *incidence rate difference* for the two measures.

Difference measures such as risk difference and incidence rate difference measure the absolute effect of an exposure. It is also possible to measure the relative effect. As an analogy, consider how one might assess the performance of an investment over a period of time. Suppose that an initial investment of $100 became $120 after 1 year. One might take the difference in the value of the investment at the end of the year and at the beginning as a measure of how well the investment did. This difference, $20, measures the absolute performance of the investment. The relative performance is obtained by dividing the absolute increase by the initial amount, which gives $20/$100, or 20%. Contrast this investment experience with that of another investment, in which an initial sum of $1000 grew to $1150 after 1 year. For the latter investment, the absolute increment is $150, far greater than the $20 from the first investment. On the other hand, the relative performance of the second investment is $150/$1000, or 15%, which is worse than the first investment.

We can obtain relative measures of effect in the same manner that we figure the relative success of an investment. We first obtain an absolute measure of effect, which would be either the risk difference or the incidence rate difference, and then we divide that by the measure of occurrence of disease among the unexposed. For risks, the relative effect is

$$\text{Relative effect} = \frac{\text{Risk difference}}{\text{Risk in unexposed}} = \frac{RD}{R_0}$$

[RD is the risk difference, and R_0 is the risk among the unexposed. Because $RD = R_1 - R_0$ (R_1 is the risk among exposed), this expression can be rewritten as follows.

$$\text{Relative effect} = \frac{RD}{R_0} = \frac{R_1 - R_0}{R_0} = RR - 1 \qquad (3\text{–}5)$$

where the *risk ratio (RR)* is defined as R_1/R_0. Thus, the relative effect is the risk ratio minus 1. This result is exactly parallel to the investment analogy, in which the relative success of the investment was the ratio of the value after investing divided by the value before investing, minus 1.

For the smaller of the two investments, this computation would give ($120/$100) − 1 = 1.2 − 1 = 20%. If we have a risk in exposed of 0.25 and a risk in unexposed of 0.10, then the relative effect is (0.25/0.10) − 1, or 1.5 (sometimes expressed as 150%). The *RR* is 2.5, and the relative effect is the part of the *RR* in excess of 1.0. The value of 1.0 is the value of *RR* when there is no effect. By defining the relative effect in this way, we ensure that we have a relative effect of 0 when the absolute effect is also 0.

Because the relative effect is simply *RR* − 1, it is common for epidemiologists to refer to the *RR* itself as a measure of relative effect, without subtracting the 1. When the *RR* is used in this way, it is important to keep in mind that a value of 1 corresponds to the absence of an effect. For example, *RR* of 3 represents twice as great an effect as *RR* of 2. Sometimes epidemiologists refer to the percentage increase in risk to convey the magnitude of relative effect. For example, one might describe an effect that represents a 120% increase in risk. Obviously, this increase is meant to describe a relative, not an absolute, effect because we cannot have an absolute effect of 120%. Describing an effect in terms of a percentage increase in risk is precisely the same as the relative effect defined above. An increase of 120% corresponds to *RR* of 2.2, which is 2.2 − 1.0 = 120% greater than 1. Thus, the 120% is a description of the relative effect that subtracts the 1 from the *RR*. Usually, it is straightforward to determine from the context whether a description of relative effect is *RR* or *RR* − 1. If the effect is described as a fivefold increase in risk, it means that the *RR* is 5. If the effect is described as a 10% increase in risk, it will correspond to *RR* of 1.1, which is 1.1 − 1.0.

Effect measures that involve the incidence rate difference and the incidence rate ratio are defined analogously to those involving the risk difference and the risk ratio. Table 3–4 compares absolute and relative measures constructed from risks and rates.

The range of the risk difference measure derives from the range of risk itself, which is [0,1]. The lowest possible risk difference would result from an exposed group with zero risk and an unexposed group at 100% risk, giving −1 for the difference. Analogously, the greatest possible risk difference, 1, comes from an exposed group with 100% risk and an unexposed group with zero risk. Risk difference has no dimensionality (that is, it has no units and is measured as a pure number) because the underlying measure, risk, is also dimensionless and the dimensionality of a difference is the same as that of the underlying measure.

The risk ratio has a range that is never negative because a risk cannot be negative. The smallest risk ratio occurs when the risk in the exposed group, the numerator of the risk ratio, is zero. The largest risk ratio occurs when the risk among the unexposed is zero, giving a ratio of infinity. Any ratio measure will be dimensionless if the numerator and denominator quantities have the same dimensionality because the dimensions divide out. In the case of risk ratio, both the numerator and the denominator, as well as their ratio, are dimensionless.

Table 3–4. Comparison of absolute and relative effect measures

Measure	Numerical Range	Dimensionality
Risk difference	[−1, +1]	None
Risk ratio	[0, ∞]	None
Incidence rate difference	[−∞, +∞]	1/time
Incidence rate ratio	[0, ∞]	None

Incidence rates range from zero to infinity and have the dimensionality of 1/time. From these characteristics, it is straightforward to deduce the range and dimensionality of the incidence rate difference and the incidence rate ratio.

Examples

Table 3–5 presents data on the risk of diarrhea among breast-fed infants during a 10-day period following their infection with *Vibrio cholerae 01*, according to the level of antipolysaccharide antibody titers in their mother's breast milk.[6] The data show a substantial difference in the risk of developing diarrhea according to whether the mother's breast milk contains a low or a high level of antipolysaccharide antibody. The risk difference for infants exposed to milk with low compared with high levels of antibody is 0.86 − 0.44 = 0.42. This risk difference reflects the additional risk of diarrhea among infants whose mother's breast milk has low antibody titers compared with the risk among infants whose mother's milk has high titers, under the assumption that the infants exposed to low titers would experience a risk equal to that of those exposed to high titers except for the lower antibody levels.

Table 3–5. Diarrhea during a 10-day follow-up period in breast-fed infants colonized with *Vibrio cholera 01* by the level of antipolysaccharide antibody titer in their mother's breast milk*

	Antibody Level		
	Low	**High**	**Total**
Diarrhea	12	7	19
No diarrhea	2	9	11
Total	14	16	30
Risk	0.86	0.44	0.63

*Data from Glass et al.[6]

We can also measure the effect of breast-feeding on diarrhea risk in relative terms. The *RR* is 0.86/0.44 = 1.96. The relative effect is 1.96 − 1, or 0.96, which would be expressed as a 96% greater risk of diarrhea among infants exposed to low antibody titers in the mother's breast milk. Commonly, we would simply describe the risk among infants exposed to low titers as being 1.96 times the risk among infants exposed to high titers.

The calculation of effects from incidence rate data is analogous to the calculation of effects from risk data. Table 3–6 gives data for the incidence rate of breast cancer among women who were treated for tuberculosis early in the twentieth century.[7] Some women received a treatment that involved repeated fluoroscopy of the lungs, with a resulting high dose of ionizing radiation to the chest.

The incidence rate among those exposed to radiation is 14.6/10,000 yr^{-1} compared with 7.9/10,000 yr^{-1} among those unexposed. The incidence rate difference is (14.6 − 7.9)/10,000 yr^{-1} = 6.7/10,000 yr^{-1}. This difference reflects the rate of breast cancer among exposed women that can be attributed to radiation exposure, under the assumption that exposed women would have had a rate equal to that among unexposed women if not for the exposure. As before, we can also measure the effect in relative terms. The incidence rate ratio is 14.6/7.9, or 1.86. The relative effect is 1.86 − 1, or 0.86, which would be expressed as an 86% greater rate of breast cancer among women exposed to the radiation. Alternatively, one might simply describe the incidence rate ratio as indicating a rate of breast cancer among exposed women that is 1.86 times that of the rate among unexposed women.

Relation Between Risk Ratios and Rate Ratios

Risk data produce estimates of effect that are either risk differences or risk ratios, and rate data produce estimates of effect that are rate differences or rate ratios. Risks cannot be compared directly with rates (they have different units), and for the same reason risk differences cannot be compared with rate differences. Under certain conditions, however, a risk ratio will be equivalent to a rate ratio. Suppose that we have incidence rates that are constant over time, with the rate among exposed people equal to I_1 and the rate among unexposed people equal to I_0. From equation 3–1, we know that a constant incidence rate will result in a risk approximately equal to the product of the rate times the time period, provided that the time period is short enough so that the risk remains under about 0.20. Above that value, the approximation does not

Table 3–6. Breast cancer cases and person-years of observation for women with tuberculosis repeatedly exposed to multiple x-ray fluoroscopies and unexposed women with tuberculosis*

	Radiation Exposure		
	Yes	**No**	**Total**
Breast cancer cases	41	15	56
Person-years	28,010	19,017	47,027
Rate (cases/10,000 person-yr)	14.6	7.9	11.9

*Data from Boice and Monson.[7]

Rounding: How many digits should be reported?

A frequent question that arises in the reporting of results is how many digits of accuracy should be reported. In some published papers, a risk ratio might be reported as 4.1; in others, the same number might be reported as 4.0846. It is clear that the number of digits should reflect the amount of precision in the data. The number 4.0846 implies that one is fairly sure that the data warrant a reported value that lies between 4.084 and 4.085. Only a truly large study would produce that level of precision. Nevertheless, it is surprisingly hard to offer a general rule for the number of digits that should be reported. For example, suppose one believes that for a given study reporting should carry into the first decimal, say, 4.1. If the study reported risk ratios, however, and these took on values below 1.0, the ratios would be rounded to values such as 0.7 or 0.8. This amount of rounding error is greater, in proportion to the size of the effect, than the rounding error in a reported value such as 4.1. So a simple rule such as one decimal place (or two, or whatever) will not suffice. How about the rule that suggests using a constant number of meaningful digits? With this rule, 4.1 would have the same reporting accuracy as 0.83. This rule may appear to be an improvement, but it breaks down near the value of 1.0 for ratio measures: it suggests that we should distinguish 0.98 from 0.99, but that we should not distinguish 1.00 from 1.01. Both of the latter numbers would be rounded to 1.0, and the next reportable value would be 1.1. If all of the risk ratios to be reported ranged from 0.9 to 1.1, this rule would make little sense.

No rule is needed as long as the writer uses good judgment and thinks about the number of digits to report. One should remember never to round values used in intermediate calculations; round only in the final step before reporting. Also, consider that rounding 1.41 to 1.4 is not a large error, but rounding 1.25 to 1.2 or to 1.3 is a rounding error that amounts to 20% of the effect for a rate ratio (keeping in mind that 1.0 equals no effect). Finally, when rounding a number ending in 5, it is customary to round upward, but it is preferable to use an unbiased strategy, such as rounding to the nearest even number. Thus, under this strategy, both 1.75 and 1.85 would be rounded to 1.8.

work very well. Suppose that we are dealing with short time periods. Then the ratio of the risk among the exposed to the risk among the unexposed, R_1/R_0, will be expressed as follows.

$$\text{Risk ratio} = \frac{R_1}{R_0} = \frac{I_1 \cdot \text{Time}}{I_0 \cdot \text{Time}} = \frac{I_1}{I_0}$$

This relation shows that the risk ratio will be the same as the rate ratio, provided that the time period over which the risks apply is sufficiently short or the rates are sufficiently low for equation 3–1 to apply. The shorter the time period or the lower the rates, the better the approximation represented by equation 3–1 and the closer the value of the risk ratio to the rate ratio. Over longer time periods (the length depending on the value of the rates involved), risks may become sufficiently great that the risk ratio will begin to diverge from the rate ratio. Because risks cannot exceed 1.0, the maximum value of a risk ratio cannot be greater than 1 divided by the risk among the unexposed. Consider the data in Table 3–5, for example. The risk in the high antibody group (which we consider to be the unexposed group) is 0.44. With this risk for the unexposed group, the risk ratio cannot exceed 1/0.44, or 2.3. In fact, the observed risk ratio of 1.96 is not far below the maximum possible risk ratio. Incidence rate ratios are not constrained by this type of ceiling, so when the unexposed risk is high, we can expect there to be a divergence between the incidence rate ratio and the risk ratio. We do not know the incidence rates that gave rise to the risks illustrated in Table 3–5, but it is reasonable to infer that the ratio of the incidence rates, were they available, would be much greater than 1.96.

If the time period over which a risk is calculated approaches 0, the risk itself also approaches 0: thus, the risk of a given person having a myocardial infarction may be 10% in a decade, but in the next 10 seconds it will be extremely small, its value shrinking along with the length of the time interval. Nevertheless, the ratio of two quantities that both approach 0 does not necessarily approach 0; in the case of the risk ratio calculated for risks that apply to shorter and shorter time intervals, as the risks approach 0, the risk ratio approaches the value of the incidence rate ratio. The incidence rate ratio is thus the limiting value for the risk ratio as the time interval over which the risks are taken approaches 0. Therefore, we can describe the incidence rate ratio as an *instantaneous risk ratio*. This equivalence of the two types of ratio for short time intervals has resulted in some confusion of terminology: often, the phrase *relative risk* is used to refer to either an incidence rate ratio or a risk ratio. Either of the latter terms is preferable to *relative risk*, since they describe the nature of the data from which the ratio derives. Nevertheless, because the risk ratio and the rate ratio are equivalent for small risks, the more general term *relative risk* has some justification. Thus, the often-used notation *RR* is sometimes read to mean relative risk, which might equally be read as risk ratio or rate ratio, all of which are equivalent if the risks are sufficiently small.

When risk does not mean risk

In referring to effects, some speakers or writers inaccurately use the word *risk* in place of the word *effect*. For example, suppose that a study reports two risk ratios for lung cancer from asbestos exposure, 5.0 for young adults and 2.5 for older adults. One might occasionally see these effect values described as follows: "The risk of lung cancer from asbestos exposure is not as great among older people as among younger people." This statement is incorrect. In fact, the risk difference between those exposed and those unexposed to asbestos is sure to be greater among older adults than younger adults, and thus the risk attributable to the effect of asbestos is greater in older adults. The risk ratio is smaller among older adults because the risk of lung cancer increases steeply with age, so the ratio for older adults is based on a larger denominator. The statement is wrong because the term *risk* has been used in place of the term *risk ratio*, or the more general term *effect*. It is perfectly correct to describe the data as follows: "The risk ratio of lung cancer from asbestos exposure is not as great among older people as among younger people."

Attributable Fraction

If we take the risk difference between exposed and unexposed people, $R_1 - R_0$, and divide it by the risk in the unexposed group, we obtain the relative measure of effect (see equation 3–5 above). We can also divide the risk difference by the risk in exposed people to get an expression that we refer to as the attributable fraction.

$$\text{Attributable fraction} = \frac{RD}{R_1} = \frac{R_1 - R_0}{R_1} = 1 - \frac{1}{RR} = \frac{RR - 1}{RR} \qquad (3\text{–}6)$$

If the risk difference reflects a causal effect that is not distorted by any bias, then the attributable fraction is a measure that quantifies the proportion of the disease burden among exposed people that is caused by the exposure. To illustrate, consider the hypothetical data in Table 3–7. The risk of disease during a 1-year period is 0.05 among the exposed and 0.01 among the unexposed. Let us suppose that this difference can be reasonably attributed to the effect of the exposure (because we believe that we have accounted for all substantial biases). The risk difference is 0.04, which is 80% of the risk among the exposed. We would then say that the exposure accounts for 80% of the disease that occurs among

Table 3–7. Hypothetical data giving 1-year disease risks for exposed and unexposed people

	Unexposed	Exposed	Total
Disease	900	500	1400
No disease	89,100	9500	98,600
Total	90,000	10,000	100,000
Risk	0.01	0.05	0.014

exposed people during the 1-year period. Another way to calculate the attributable fraction is from the risk ratio: $(5 - 1)/5 = 80\%$.

If we wish to calculate the attributable fraction for the entire population of 100,000 people in Table 3–7, we would first calculate the attributable fraction for exposed people. To obtain the overall attributable fraction for the total population, the fraction among the exposed should be multiplied by the proportion of all cases in the total population that is exposed. There are 1400 cases in the entire population, of whom 500 are exposed. Thus, the proportion of exposed cases is $500/1400 = 0.357$. The overall attributable fraction for the population is the product of the attributable fraction among the exposed and the proportion of exposed cases: $0.8 \times 0.357 = 0.286$. That is, 28.6% of all cases in the population are attributable to the exposure. This calculation is based on a straightforward idea: no case can be caused by exposure unless the person is exposed, so among all of the cases, only some of the exposed cases can be attributable to the exposure. There were 500 exposed cases, of whom we calculated that 400 represent excess cases caused by the exposure. None of the 900 cases among the unexposed is attributable to the exposure, so among the total of 1400 cases in the population, only 400 of the exposed cases are attributable to the exposure: the proportion $400/1400 = 0.286$, which is the same value that we calculated.

If the exposure is categorized into more than two levels, we can use formula 3–7, which takes into account each of the exposure levels.

$$\text{Total attributable fraction} = \sum_i (AF_i \times P_i) \qquad (3\text{–}7)$$

AF_i is the attributable fraction for exposure level i, P_i represents the proportion of all cases that falls in exposure category i, and Σ indicates the sum of each of the exposure-specific attributable fractions. For the unexposed group, the attributable fraction would be 0.

Let us apply formula 3–7 to the hypothetical data in Table 3–8, which give risks for a population with three levels of exposure. The attributable fraction for the group with no exposure is 0. For the low-exposure group, the attributable fraction is 0.50 because the risk ratio is 2. For the high-exposure group, the attributable fraction is 0.75 because the risk ratio is 4. The total attributable fraction is as follows:

$$0 + 0.50(0.48) + 0.75(0.48) = 0.24 + 0.36 = 0.60$$

The same result can also be calculated directly from the number of attributable cases at each exposure level.

$$(0 + 600 + 900)/2500 = 0.60$$

Under certain assumptions, the estimation of attributable fractions can be based on rates as well as risks. Thus, in formula 3–6, which uses the risk ratio to calculate the attributable fraction, the rate ratio could be used instead, provided that the conditions are met for the rate ratio to approximate the risk ratio. If exposure results in an increase in disease occurrence at some levels of exposure and a decrease at other levels of exposure, compared with no exposure, the net attributable fraction will be a combination of the prevented cases and the caused cases at the different levels of exposure. The net effect of exposure in such situations can be difficult to assess and may obscure the components of the exposure effect. This topic is discussed in greater detail by Rothman and Greenland.[8]

Table 3–8. Hypothetical data giving 1-year disease risks for people at three levels of exposure

	Exposure			
	None	Low	High	Total
Disease	100	1200	1200	2500
No disease	9900	58,800	28,800	97,500
Total	10,000	60,000	30,000	100,000
Risk	0.01	0.02	0.04	0.025
Risk ratio	1.00	2.00	4.00	
Proportion of all cases	0.04	0.48	0.48	

Questions

1. Suppose that in a population of 100 people 30 die. The risk of death could be calculated as 30/100. What is missing from this measure?
2. Can we calculate a rate for the data in question 1? If so, what is it? If not, why not?
3. Eventually all people die. Why should we not state that the mortality rate for any population is always 100%?
4. If incidence rates remain constant with time and if exposure causes disease, which will be greater, the risk ratio or the rate ratio?

5. Why is it incorrect to describe a rate ratio of 10 as indicating a high risk for disease among the exposed?
6. A newspaper article states that a disease has increased by 1200% in the past decade. What is the rate ratio that corresponds to this level of increase?
7. Another disease has increased by 20%. What is the rate ratio that corresponds to this increase?
8. From the data in Table 3–5, calculate the fraction of diarrhea cases among infants exposed to a low antibody level that is attributable to the low antibody level. Calculate the fraction of all diarrhea cases attributable to exposure to low antibody levels. What assumptions are needed to interpret the result as an attributable fraction?
9. What proportion of the 56 breast cancer cases in Table 3–6 is attributable to radiation exposure? What are the assumptions?
10. Suppose you worked for a health agency and had collected data on the incidence of lower back pain among people in different occupations. What measures of effect would you choose and why?
11. Suppose that the rate ratio measuring the relation between an exposure and a disease is 3 in two different countries. Would this situation imply that exposed people have the same risk in the two countries? Would it imply that the effect of the exposure is the same in the two countries? Why or why not?

References

1. Cole, P: The evolving case-control study. *J Chron Dis* 1979;32:15–27.
2. Iskrant AP, Joliet PV: Table 24 in *Accidents and Homicides* Vital and Health Statistics Monographs, American Public Health Association, Harvard University Press, Cambridge, 1968.
3. Snow, J: *On the Mode of Communication of Cholera,* 2nd ed. London: John Churchill, 1860. (Facsimile of 1936 reprinted edition by Hafner, New York, 1965.)
4. Kurtzke, JF, Hyllested, K: Multiple sclerosis in the Faroe Islands: clinical and epidemiologic features. *Ann Neurol* 1979;5:6–21.
5. Cole, TB, Chorba, TL, Horan, JM: Patterns of transmission of epidemic hysteria in a school. *Epidemiology* 1990;1:212–218.
6. Glass, RI, Svennerholm, AM, Stoll, BJ, et al: Protection against cholera in breast-fed children by antibiotics in breast milk. *N Engl J Med* 1983; 308:1389–1392.
7. Boice, JD, Monson, RR: Breast cancer in women after repeated fluoroscopic examinations of the chest. *J Natl Cancer Inst* 1977;59:823–832.
8. Rothman, KJ, Greenland, S: *Modern Epidemiology,* 2nd ed. Philadelphia: Lippincott-Raven, 1998.

4

Types of Epidemiologic Study

In the last chapter, we learned about measures of disease frequency, including risk, incidence rate, and prevalence; measures of effect, including risk and incidence rate differences and ratios; as well as attributable fractions. Epidemiologic studies may be viewed as measurement exercises undertaken to obtain estimates of these epidemiologic measures. The simplest studies aim only at estimating a single risk, incidence rate, or prevalence. More complicated studies aim at comparing measures of disease occurrence, with the goal of predicting such occurrence, learning about the causes of disease, or evaluating the impact of disease on a population. This chapter describes the two main types of epidemiologic study, the cohort study and the case-control study, along with several variants.

Cohort Studies

In epidemiology, a *cohort* is defined most broadly as "any designated group of individuals who are followed or traced over a period of time." [1] A cohort study, which is the archetype for all epidemiologic studies, involves measuring the occurrence of disease within one or more cohorts. Typically, a cohort comprises persons with a common characteristic, such as an exposure or ethnic identity. For simplicity, we refer to two cohorts, *exposed* and *unexposed,* in our discussion. In this context, we use the term *exposed* in its most general sense; for example, an exposed cohort could have in common the presence of a specific gene. The purpose of following a cohort is to measure the occurrence of one or more specific diseases during the period of follow-up, usually with the aim of comparing the disease rates for two or more cohorts.

The concept of following a cohort to measure disease occurrence may appear straightforward, but there are many complications involving who is eligible to be followed, what should count as an instance of disease, how the incidence rates or risks are measured, and how exposure ought to be defined. Before we explore these issues, let us examine, as an example, an elegantly designed epidemiologic cohort study.