Subjects No. 32, Ministry of Health. London: His Majesty's Stationery Office; 1926.

18. Winkelstein W Jr. Janet Lane-Claypon, a forgotten epidemiologic pioneer. *Epidemiology.* 2006;17:705.
19. Paneth N, Susser E, Susser M. Origins and early development of the case-control study. Part 2: The case-control study from Lane-Claypon to 1950. *Social Prev Med.* 2002;47:359–365.
20. Daniel TM. *Wade Hampton Frost, Pioneer Epidemiologist 1880–1938.* Rochester, NY: University of Rochester Press; 2004.

3

What Is Causation?

The acquired wisdom that certain conditions or events bring about other conditions or events is an important survival trait. Consider an infant whose first experiences are a jumble of sensations that include hunger, thirst, color, light, heat, cold, and many other stimuli. Gradually, the infant begins to perceive patterns in the jumble and to anticipate connections between actions such as crying and effects such as being fed. Eventually, the infant assembles an inventory of associated perceptions. Along with this growing appreciation for specific causal relations comes the general idea that some events or conditions can be considered causes of other events or conditions.

Thus, our first appreciation of the concept of causation is based on our own observations. These observations typically involve causes with effects that are immediately apparent. For example, changing the position of a light switch on the wall has the instant effect of causing the light to go on or off. There is, however, more to the causal mechanism for getting the light to shine than turning the light switch to the on position. If the electric lines to the building are down because of a storm, turning on the switch will have no effect. If the bulb is burned out, manipulating the switch also will have no effect. One cause of the light going on is having the switch in the proper place, but along with it we must include a supply of power to the circuit, a working bulb, and intact wiring. When all other factors are in place, turning the switch will cause the light to go on, but if one or more of the other factors is not playing its causal role, the light will not go on when the switch is turned. There is a tendency to consider the switch as the unique cause of turning on the light, but we can define a more intricate causal mechanism in which the switch is one component of several. The tendency to identify the switch as the unique cause stems from its usual role as the final factor that acts in the causal mechanism. The wiring can be considered part of the causal mechanism, but after it is installed, it seldom warrants further attention. The switch is typically the only part of the mechanism that needs to be activated to turn on the light. The effect usually occurs immediately after turning the switch, and as a result, we tend to identify the switch as a unique cause. The inadequacy of this assumption is emphasized when the bulb fails and must be replaced before the light will go on.

THE CAUSAL PIE MODEL

Causes of disease can be conceptualized in the same way as the causes of turning on a light. A helpful way to think about causal mechanisms for disease is depicted in Figure 3–1.[1] Each *pie* in the diagram represents a theoretical *causal mechanism* for a given disease, sometimes called a *sufficient cause.* The three pies illustrate that there are multiple mechanisms that cause any type of disease. Each individual instance of disease occurs through a single mechanism or sufficient cause. A given causal mechanism requires the joint action of many component factors, or *component causes.* Each component cause is an event or a condition that plays a necessary role in the occurrence of some cases of a given disease. For example, the disease may be cancer of the lung, and in the first mechanism in Figure 3–1, factor C may be cigarette smoking. Other factors include genetic traits or other environmental exposures that play a causal role in cancer of the lung. Some component causes presumably act in many different causal mechanisms. (Terminology note: the *causal pie model* has also been described as the *sufficient-component cause model.*)

Implications of the Causal Pie Model

Multicausality

The model of causation shown in Figure 3–1 illuminates several important principles of causation, the most important of which is that every causal mechanism involves the joint action of a multitude of component causes. Consider as an example the cause of a broken hip. Suppose that someone experiences a traumatic injury to the head that leads to a permanent disturbance in equilibrium. Many years later, faulty equilibrium plays a causal role in a fall that occurs while the person is walking on an icy path. The fall results in a broken hip. Other factors playing a causal role for the broken hip may include the type of shoe the person was wearing, the lack of a handrail along the path, a sudden gust of wind, and the weight of the person. The complete causal mechanism involves a multitude of factors. Some factors, such as the earlier injury that resulted in the equilibrium disturbance and the weight of the person, reflect earlier events that have had a lingering effect. Some causal components of the broken hip are *genetic.* Genetic factors affect the person's weight, gait, behavior, and recovery from the earlier trauma. Other factors, such as the force of the wind, are *environmental* (nongenetic). There usually are some genetic and some environmental component causes in every causal mechanism. Even an event such as a fall on an icy path that results in a broken hip is part of a complicated causal mechanism that involves many component causes.

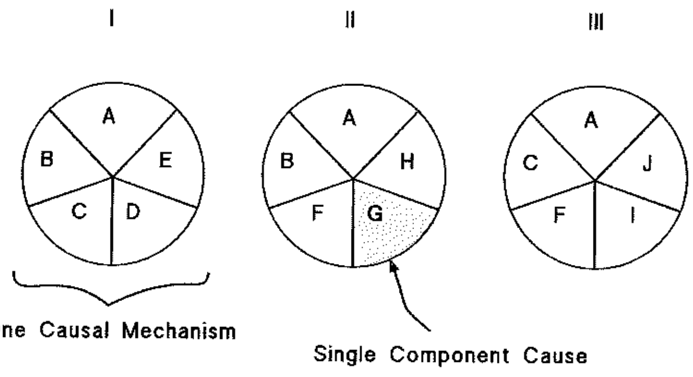

Figure 3–1 Three sufficient causes of a disease.

Genetic Versus Environmental Causes

It is a strong assertion that every case of every disease has both genetic and environmental causes. Nevertheless, if all genetic factors that determine disease are taken into account, essentially 100% of disease can be said to be inherited, in the sense that nearly all cases of disease have some genetic component causes. What would be the genetic component causes of someone who gets drunk and is killed in an automobile after colliding with a tree? Genetic traits may lead to psychiatric problems such as alcoholism, which may lead to drunk driving and consequent fatality. It is also possible to claim that essentially 100% of any disease is environmentally caused, even diseases that often are considered to be purely genetic. Phenylketonuria, for example, is considered by many to be purely genetic. Nonetheless, if we consider the disease that phenylketonuria represents to be the mental retardation that may result from it, we can prevent the disease by appropriate dietary intervention. The disease therefore has environmental determinants, and its causes are both environmental and genetic. Although it may seem like an exaggeration to claim that 100% of cases of any disease are environmental and genetic at the same time, it is a good approximation. It may seem counterintuitive, because we cannot manipulate many of the causes in most situations and the ones that can be controlled are usually solely environmental causes, as in the manipulation of diet to prevent the mental retardation of phenylketonuria.

Strength of Causes

It is common to think that some component causes play a more important role than other factors in the causation of disease. One way this concept is expressed is by the strength of a causal effect. We say that smoking has a strong effect on lung cancer risk because smokers have about 10 times the risk of lung cancer as nonsmokers. We say that smoking has a weaker effect on myocardial infarction because the risk of a heart attack is only about twice as great in smokers as in nonsmokers. With respect to an individual case of disease, however, every component cause that played a role was necessary to the occurrence of that case.

According to the causal pie model, for a given case of disease, there is no such thing as a strong cause or a weak cause. There is only a distinction between factors that were causes and factors that were not causes.

To understand what epidemiologists mean by *strength* of a cause, we need to shift from thinking about an individual case to thinking about the total burden of cases occurring in a population. We can then define a *strong cause* to be a component cause that plays a causal role in a large proportion of cases and a *weak cause* to be a causal component in a small proportion of cases. Because smoking plays a causal role in a high proportion of the lung cancer cases, we call it a strong cause of lung cancer. For a given case of lung cancer, smoking is no more important than any of the other component causes for that case; but on the population level, it is considered a strong cause of lung cancer because it causes such a large proportion of cases.

The strength of a cause defined in this way necessarily depends on the prevalence of other causal factors that produce disease. As a result, the concept of a strong or weak cause cannot be a universally accurate description of any cause. Suppose we say that smoking is a strong cause of lung cancer because it plays a causal role in a large proportion of cases. Exposure to ambient radon gas is considered to be a weaker cause because it has a causal role in a much smaller proportion of lung cancer cases. Imagine that society eventually succeeds in eliminating tobacco smoking, with a consequent reduction in smoking-related cases of lung cancer. One result is that a much larger proportion of the lung cancer cases that continue to occur will be caused by exposure to radon gas; eliminating smoking would strengthen the causal effect of radon gas on lung cancer. This example illustrates that *strength of effect* is not a biologically stable characteristic of a factor. From a biologic perspective, the causal role of a factor in producing disease is neither strong nor weak; the biology of causation corresponds to the identity of the component causes in a causal mechanism and the ways in which they interact to produce disease. The proportion of the population burden of disease that a factor causes, which we use to define the strength of a cause, can change from population to population and over time if there are changes in the distribution of other causes of the disease. The strength of a cause does not portray the biology of causation.

Interaction Between Causes

The causal pie model posits that several causal components act in concert to produce an effect. *Acting in concert* does not imply that factors must act at the same time. Consider the earlier example of the person who sustained trauma to the head that resulted in an equilibrium disturbance, which led years later to a fall on an icy path. The earlier head trauma played a causal role in the later hip fracture, as did the weather conditions on the day of the fracture. If both factors played a causal role in the hip fracture, they interacted with one another to cause the fracture, despite the fact that their time of action was many years apart. We would say that any and all of the factors in the same causal mechanism interact with one another to cause disease. The head trauma interacted with the weather conditions and with the other component causes, such as the type of footwear, the absence of a handhold, and any other conditions that were necessary to the causal mechanism of the fall and the broken hip that resulted. Each causal pie can

be considered as a set of interacting causal components. This model provides a biologic basis for the concept of interaction that differs from the more traditional statistical view of interaction. The implication of this difference is discussed in Chapter 11.

Sum of Attributable Fractions

Consider the data in Table 3–1, which shows the rates of head and neck cancer according to smoking status and alcohol exposure. Suppose that the differences in the rates reflect causal effects, so that confounding can be ignored. Among those who are smokers and alcohol drinkers, what proportion of the cases of head and neck cancer that occur is attributable to the effect of smoking? We know that the rate for these people is 12 cases per 10,000 person-years. If these same people were not smokers, we can infer that their rate of head and neck cancer would be 3 cases per 10,000 person-years. If this difference reflects the causal role of smoking, we can infer that 9 of every 12 cases (75%) are attributable to smoking among those who smoke and drink alcohol. If we turn the question around and ask what proportion of disease among these same people is attributable to alcohol drinking, we would be able to attribute 8 of every 12 cases (67%) to alcohol drinking.

Can we attribute 75% of the cases to smoking and 67% to alcohol drinking among those who are exposed to both? The answer is yes, because some cases are counted more than once as a result of the interaction between smoking and alcohol consumption. These cases are attributable to both smoking and alcohol drinking because both factors played a causal role in producing them. One consequence of interaction is that the proportions of disease attributable to various component causes do not sum to 100%.

A widely discussed but unpublished paper from the 1970s written by scientists at the National Institutes of Health proposed that as much as 40% of cancer is attributable to occupational exposures. Many scientists thought that this fraction was an overestimate and argued against this claim.[2,3] One of the arguments used in rebuttal was as follows: *x* percent of cancer is caused by smoking, *y* percent by diet, *z* percent by alcohol, and so on; when all of these percentages are summed, only a small percentage, much less than 40%, is left for occupational causes. This rebuttal, however, is fallacious because it is based on the naive view that every case of disease has a single cause and that two causes cannot both contribute to the same case of cancer. Because diet, smoking, asbestos, and various occupational exposures and other factors interact with one another and with genetic

Table 3–1 Hypothetical Rates of Head and Neck Cancer (Cases per 10,000 Person-Years) According to Smoking Status and Alcohol Drinking

Smoking Status	Alcohol Drinking	
	No	Yes
Nonsmoker	1	3
Smoker	4	12

factors to cause cancer, each case of cancer can be attributed repeatedly to many separate component causes. The sum of disease attributable to various component causes has no upper limit.

Induction Time

Because the component causes in a given causal mechanism do not act simultaneously, there usually is a period of time between the action of a component cause and the completion of a sufficient cause. The only exception is the last component cause to act in a given causal mechanism. The last-acting component cause completes the causal mechanism, and we can say that disease begins concurrently with its action. For earlier-acting component causes, we can define the *induction period* as the time interval that begins concurrently with the action of a component cause and ends when the final component cause acts and the disease occurs. For example, in the illustration of the fractured hip, the induction time between the head trauma that resulted in an equilibrium disturbance and the later hip fracture was many years. The induction time between the decision to wear nongripping shoes and the hip fracture might have been a matter of minutes or hours. The induction time between the gust of wind that triggered the fall and the hip fracture might have been seconds or less.

In an individual instance, we usually cannot know the exact length of an induction period, because we cannot be sure of the causal mechanism that produces disease in an individual instance nor when all the relevant component causes in that mechanism exerted their causal action. With research data, however, we can learn enough to characterize the induction period that relates the action of a single component cause to the occurrence of disease in general. An example of a lengthy induction time is the cause-effect relation between exposure of a female fetus to diethylstilbestrol (DES) and her subsequent development of adenocarcinoma of the vagina. The cancer generally occurs after the age of 15 years. Because the causal exposure to DES occurs during gestation, there is an induction time of more than 15 years for carcinogenesis. During this time, other causes presumably operate; some evidence suggests that hormonal action during adolescence may be part of the mechanism.[4]

The causal pie model makes it clear that it is incorrect to characterize a disease itself as having a lengthy or brief induction time. The induction time can be conceptualized only in relation to a specific component cause. We can say that the induction time relating DES to clear cell carcinoma of the vagina is at least 15 years, but we cannot say that 15 years is the minimum induction time for clear cell carcinoma in general. Because each component cause in any causal mechanism can act at a time different from the other component causes, each can have its own induction time. For the component cause that acts last, the induction time always equals zero. If another component cause of clear cell carcinoma of the vagina that acts during adolescence were identified, it would have a much shorter induction time than that of DES. Induction time characterizes a specific cause-effect pair rather than only the effect.

In carcinogenesis, the terms *initiator* and *promoter* are used to refer to component causes of cancer that act early and late, respectively, in the causal mechanism. Cancer itself has often been characterized as a disease process with a long induction time, but this characterization is a misconception. Any late-acting component

in the causal process, such as a promoter, will have a short induction time, and the induction time will always be zero for the last component cause (eg, the gust of wind causing the broken hip in the earlier example), because after the final causal component acts, disease has occurred. At that point, however, the presence of disease is not necessarily apparent. A broken hip may be apparent immediately, but a cancer that has just been caused may not become noticed or diagnosed for an appreciable time. The time interval between disease occurrence and its subsequent detection, whether by medical testing or by the emergence of symptoms, is called the *latent period*.[4] The length of the latent period can be reduced by improved methods of disease detection. The induction period, however, cannot be reduced by early detection of disease, because there is no disease to detect until after the induction period is over. Practically, it may be difficult to distinguish between the induction period and the latent period, because there may be no way to establish when the disease process began if it is not detected until later. Diseases such as slow-growing cancers may appear to have long induction periods with respect to many causes, in part because they have long latent periods.

Although it is not possible to reduce the induction period by earlier detection of disease, it may be possible to observe intermediate stages of a causal mechanism. The increased interest in biomarkers such as DNA adducts is an example of focusing on causes that are more proximal to the disease occurrence. Biomarkers may reflect the effects on the organism of agents that have acted at an earlier time.

Is a Catalyst a Cause?

Some agents may have a causal action by shortening the induction time of other agents. Suppose that exposure to factor A leads to epilepsy after an average interval of 10 years. It may be that exposure to drug B can shorten this interval to 2 years. Is B acting as a catalyst or as a cause of epilepsy? The answer is both; a catalyst *is* a cause. Without B, the occurrence of epilepsy comes 8 years later than it comes with B, so we can say that B causes the epilepsy to occur earlier. It is not sufficient to argue that the epilepsy would have occurred anyway and therefore that B is not a cause of its occurrence. First, it would not have occurred at that time, and the time of occurrence is considered part of the definition of an event. Second, epilepsy will occur later only if the person survives an additional 8 years, which is not certain. Agent B therefore determines when the epilepsy occurs, and it can determine whether it occurs at all. For this reason, we consider any agent that acts as a catalyst of a causal mechanism, shortening the induction period for other agents, to be a cause. Similarly, any agent that postpones the onset of an event, drawing out the induction period for another agent, we consider to be a preventive. It should not be too surprising to equate postponement with prevention; we routinely use such an equation when we employ the euphemism that we prevent death, which can only be postponed. We prevent death at a given time in favor of death at a later time. Similarly, slowing the process of atherosclerosis can result in postponement (and thereby prevention) of cardiovascular disease and death.

THE PROCESS OF SCIENTIFIC INFERENCE

Much epidemiologic research is aimed at uncovering the causes of disease. Now that we have a conceptual model for causes, how do we determine whether a given relation is causal? Some scientists refer to checklists for causal inference, and others focus on complicated statistical approaches, but the answer to this question is not to be found in checklists or statistical methods. The question itself is tantamount to asking how we apply the scientific method to epidemiologic research. This question leads directly to the philosophy of science, a topic that goes well beyond the scope of this book. Nevertheless, it is worthwhile to summarize two of the major philosophic doctrines that have influenced modern science.

Induction

Since the rise of modern science in the 17th century, scientists and philosophers have puzzled over the question of how to determine the truth about assertions that deal with the empirical world. From the time of the ancient Greeks, deductive methods have been used to prove the validity of mathematic propositions. These methods enable us to draw airtight conclusions because they are self-contained, starting with a limited set of definitions and axioms and applying rules of logic that guarantee the validity of the method. Empirical science is different, however. Assertions about the real world do not start from arbitrary axioms, and they involve observations on nature that are fallible and incomplete. These stark differences from deductive logic led early modern empiricists, such as Francis Bacon, to promote what they considered a new type of logic, which they called *induction* (not to be confused with the concept of an induction period). *Induction* was an indirect method used to gain insight into what has been metaphorically described as the fabric of nature.

The method of induction starts with observations on nature. To the extent that the observations fall into a pattern, they are said to induce in the mind of the observer a suggestion of a more general statement about nature. The general statement can range from a simple hypothesis to a more profound natural law or natural relation. The statement about nature is reinforced with further observations or refuted by contradictory observations. For example, suppose an investigator in New York conducts an experiment to determine the boiling point of water and observes that the water boils at 100°C. The experiment is repeated many times, each time showing that the water boils at about 100°C. By induction, the investigator concludes that the boiling point of water is 100°C. The induction itself involves an inference beyond the observations to a general statement that describes the nature of boiling water. As induction became popular, it was seen to differ considerably from deduction. Although not as well understood as deduction, the approach was considered a new type of logic, *inductive logic*.

Although induction, with its emphasis on observation, represented an important advance over the appeal to faith and authority that characterized medieval scholasticism, it was not long before the validity of the new logic was questioned. The sharpest criticism came from the skeptical philosopher David Hume, who pointed out that induction had no logical force. Rather, it amounted to the assumption that what had been observed in the past would continue to occur in the future. When supporters of induction argued that induction was a valid process because it had been seen to work on numerous occasions, Hume countered that the argument was an example of circular reasoning that relied on induction to justify itself. Hume was so profoundly skeptical that he distrusted any inference based on observation because observations depend on sense perceptions and are therefore subject to error.

Refutationism

Hume's criticisms of induction have been a powerful force in modern scientific philosophy. The most influential reply to Hume was offered by Karl Popper. Popper accepted Hume's point that in empirical science one cannot prove the validity of a statement about nature in any way that is comparable with a deductive proof. Popper's philosophy, known as *refutationism*, held that statements about nature can be "corroborated" by evidence, but corroboration does not amount to a logical proof. On the other hand, Popper also asserted that statements about nature can be refuted by deductive logic. To grasp the point, consider the earlier example of observing the boiling point of water. The refutationist view is that the repeated experiments showing that water boils at 100°C corroborate the hypothesis that water boils at this temperature, but they do not prove it.[5] A colleague of the New York researcher who works in Denver, a city located at high altitude, would find that water there boils at 94°C. This single contrary observation carries more weight regarding the hypothesis about the boiling point of water than thousands of repetitions of the initial experiment at sea level.

The asymmetric implications of a refuting observation compared with supporting observations are the essence of the refutationist view. This school of thought encourages scientists to subject a new hypothesis to rigorous tests that may falsify the hypothesis in preference to repetitions of the initial observations that add little beyond the weak corroboration that replication can supply. The implication for the method of science is that hypotheses should be evaluated by subjecting them to crucial tests. If a test refutes a hypothesis, a new hypothesis needs to be formulated that can then be subjected to further tests. After finding that water boils in Denver at a lower temperature than it boils in New York, the investigator must discard the hypothesis that water boils at 100°C and replace it with a more refined hypothesis, such as one that will explain the difference in boiling points under different atmospheric pressures. This process describes an endless cycle of *conjecture and refutation*. The conjecture, or hypothesis, is the product of scientific insight and imagination. It requires little justification except that it can account for existing observations. A useful approach is to pose competing hypotheses to explain existing observations and to test them against one another. The refutationist philosophy postulates that all scientific knowledge is tentative because it may one day need to be refined or even discarded. In this philosophy, what we call scientific knowledge is a body of currently unrefuted hypotheses that appear to explain existing observations.

How can an epidemiologist apply refutationist thinking to his or her work? If causal mechanisms are stated specifically, an epidemiologist can construct crucial tests of competing hypotheses. For example, when toxic shock syndrome was first studied, there were two competing hypotheses about the origin of the toxin. In one, the toxin responsible for the disease was a chemical in the tampon, and women using tampons were exposed to the toxin directly from the tampon. In the other hypothesis, the tampon acted as a culture medium for staphylococci that produced the toxin. Both hypotheses explained the correlation of toxic shock occurrence and tampon use. The two hypotheses, however, led to opposite predictions about the relation between the frequency of changing tampons and the risk of toxic shock. If chemical intoxication were the cause, more frequent tampon changes would lead to more exposure to the toxin and possible absorption of a greater overall dose. This hypothesis predicted that women who changed tampons more frequently would have a higher risk of toxic shock syndrome than women who changed tampons infrequently. The culture-medium hypothesis predicted that the women who change tampons frequently would have a lower risk than those who left the tampon in for longer periods, because a short duration of use for each tampon would prevent the staphylococci from multiplying enough to produce a damaging dose of toxin. Epidemiologic research, which showed that infrequent changing of tampons was associated with greater risk of toxic shock, refuted the chemical theory.

Critics of refutationism point out that refutation is not logically certain because it depends on theories, assumptions, and observations, all of which are susceptible to error. In epidemiology, for example, any study result may be influenced by an obscure bias, which is an inescapable source of uncertainty. Among the dissenting philosophic views is that of Thomas Kuhn,[6] who held that it is ultimately the collective beliefs of the community of scientists that determines what is accepted as truth about nature. According to Kuhn, the truth is not necessarily objective but rather something determined by consensus. Feyerabend,[7] another skeptic, held that science proceeds through intellectual anarchy, without any coherent method. A more moderate although still critical view was taken by Haack.[8,9] She saw science as an extension of everyday inquiry, employing pragmatic methods that she likened to solving a crossword puzzle, integrating clues with other answers in a trial-and-error approach. Despite these criticisms, refutationism has been a positive force in science by encouraging bold, testable theories and then fostering a valuable skeptical outlook by subjecting those theories to rigorous challenges.

Causal Criteria

Earlier we said that there is no simple checklist that can determine whether an observed relation is causal. Nevertheless, attempts at such checklists have appeared. Most of these lists stem from the canons of inference described by John Stuart Mill.[10] The most widely cited list of causal criteria, originally posed as a list of standards, is attributed to Hill,[11] who adapted them from the U.S. Surgeon General's 1964 report on Smoking and Health.[12] The Hill standards, often labeled the Hill criteria, are listed in Table 3–2, along with some problems related to each of the criteria.

Table 3–2 Causal Criteria of Hill

Criterion	**Problems with the Criterion**
1. Strength	Strength depends on the prevalence of other causes; it is not a biologic characteristic and can be confounded.
2. Consistency	Causal relations have exceptions that are understood best with hindsight.
3. Specificity	A cause can have many effects.
4. Temporality	It may be difficult to establish the temporal sequence between cause and effect.
5. Biologic gradient	It can be confounded; threshold phenomena would not show a progressive relation.
6. Plausibility	Too subjective
7. Coherence	How does it differ from consistency or plausibility?
8. Experimental evidence	Not always available
9. Analogy	Analogies abound.

Although Hill did not propose these criteria as a checklist for evaluating whether a reported association could be interpreted as causal, many others have attempted to apply them in that way. Admittedly, the process of causal inference as described earlier is difficult and uncertain, making the appeal of a simple checklist undeniable. Unfortunately, this checklist, like all others with the same goal, fails to deliver on the hope of clearly distinguishing causal from noncausal relations. Consider the first criterion, strength. It is tempting to believe that strong associations are more likely to be causal than weak ones, but as we saw in our discussion of causal pies, not every component cause has a strong association with the disease that it produces; strength of association depends on the prevalence of other factors. Some causal associations, such as the association between cigarette smoking and coronary heart disease, are weak. Furthermore, a strong association can be noncausal, a confounded result stemming from the effect of another risk factor for the disease that is highly correlated with the one under study. For example, birth order is strongly associated with the occurrence of Down syndrome, but it is a confounded association that is completely explained by the effect of maternal age. If weak associations can be causal and strong associations can be noncausal, it does not appear that strength of association can be considered a criterion for causality.

The third criterion (see Table 3–2), specificity, suggests that a relation is more likely to be causal if the exposure is related to a single outcome rather than myriad outcomes. This criterion is misleading because it implies, for example, that the more diseases with which smoking is associated, the greater the evidence that smoking is not causally associated with any of them. The fifth criterion, biologic gradient, is often taken as a sign of a causal relation, but it can just as well result from confounding or other biases as from a causal connection. The relation between Down syndrome and birth order, mentioned earlier, shows a biologic gradient despite being completely explained by confounding from maternal age.

Other criteria from Hill's list are vague (eg, consistency, plausibility, coherence, analogy) or do not apply in many settings (eg, experimental evidence). The only characteristic on the list that is truly a causal criterion is temporality, which implies that the cause comes before the effect. This criterion, which is part of the definition of a cause, is a useful one, although it may be difficult to establish the proper time sequence for cause and effect. For example, does stress lead to overeating, or does overeating lead to stress? It usually is better to avoid a checklist approach to causal inference and instead consider approaches such as conjecture and refutation. Checklists lend a deceptive kind of mindless authority to an inherently imperfect and creative process. In contrast, causal inference based on conjecture and refutation fosters a highly desirable critical scrutiny.

Although checklists may not be appropriate for causal inference, the points laid out by Hill are still important considerations. The criteria may be useful when applied in the context of specific hypotheses. For example, Weiss observed that the specificity of effects might be important in inferring the beneficial effect of sigmoidoscopy in screening for colorectal cancer if the association between sigmoidoscopy and reduced death from colorectal cancer is stronger for cancer occurring at sites within reach of a sigmoidoscope.[13]

Generalization in Epidemiology

A useful way to think of scientific generalization is to consider a generalization to be the elaboration of a scientific theory. A given study may test the viability of one or more theories. Theories that survive such tests can be viewed as general statements about nature that tell us what to expect in people or settings that were not studied. Because theories can be incorrect, scientific generalization is not a perfect process. Formulating a theory is not a mathematical or statistical process, and generalization should not be considered a statistical exercise. It is the process of causal inference itself.

Many people believe that generalizing from an epidemiologic study involves a mechanical process of making an inference about a target population of which the study population is considered a sample. This type of generalization does exist, in the field of survey sampling. In survey sampling, researchers draw samples from a population to avoid the expense of studying the entire population, which makes the statistical representativeness of the sample the main concern for generalizing to the source population.

Although survey sampling is an important tool for characterizing a population efficiently and may be used in some epidemiologic applications, such as prevalence surveys, it is a mechanical tool that does not always share the same goals as science. Survey sampling is useful for problems such as trying to predict how a population will vote in an election or what type of laundry soap the people in a region prefer. These are characteristics that depend on attitudes and for which there is little coherent biologic theory on which to base a scientific generalization. Survey results may be quickly outdated (eg, election polls may be repeated weekly or even daily) and do not apply outside the populations from which the surveys were conducted. (Disclaimer: I am not saying that social science is not science or that we cannot develop theories about social behavior. I am saying only that surveys about the current attitudes of a specific group of people are not the same as social theories.) Even if survey sampling is used to characterize the prevalence of disease or the medical needs of a population, the objectives are pragmatic rather than scientific and may not apply outside the study population. Scientific results from epidemiologic studies, in contrast, seldom need to be repeated weekly to see if they still apply. An epidemiologic study conducted in Chicago showing that exposure to ionizing radiation causes cancer does not need to be repeated in Houston to determine whether ionizing radiation also causes cancer in people living in Houston. Generalization about ionizing radiation and cancer is based on understanding of the underlying biology rather than on statistical sampling.

It may be helpful to consider the problem of scientific generalization about causes of cancer from the point of view of a biologist studying carcinogenesis in mice. Most researchers who study cancer in animals do so because they would like to understand better the causes of human cancer. If scientific generalization depended on having studied a statistically representative sample of the target population, researchers studying mice would have nothing to contribute to the understanding of human cancer. Mouse researchers obviously do not study representative samples of people; they do not even study representative samples of mice. Instead, they seek mice that have uniformly similar genes and perhaps certain biologic characteristics. In choosing mice to study, they have to consider mundane issues such as the cost of the mice. Although researchers studying animals are unlikely to worry about whether their mouse or rabbit subjects are statistically representative of all mice or rabbits, they may consider whether the biology of the animal population they are studying is similar to (and representative of) that of humans. This type of representativeness, however, is not statistical representativeness based on sampling from a source population; it is a biologic representativeness based on scientific knowledge. Despite the absence of statistical representativeness, no one seriously doubts the contribution that animal research can make to the understanding of human disease.

Many epidemiologic activities, such as measuring the prevalence of patients in need of dialysis, do require surveys to characterize a specific population, but these activities are usually examples of applied epidemiology rather than the science of epidemiology. The activities of applied epidemiology involve taking already established epidemiologic knowledge and applying it to specific settings, such as preventing malaria transmission by reducing the mosquito vector population or reducing lung cancer and cardiovascular disease occurrence by implementing an antismoking campaign. The activities of epidemiologic research, as in laboratory science, move away from the specific toward the general. We make specific observations in research studies and then hope to generalize from them to a broader base of understanding. This process is based more on scientific knowledge, insight, and conjecture about nature than it is on the statistical representativeness of the actual study participants. This principle has important implications for the design and interpretation of epidemiologic studies (see Chapter 7).

QUESTIONS

1. Criticize the following statement: The cause of tuberculosis is infection with the tubercle bacillus.

2. A trait in chickens called yellow shank occurs when a specific genetic strain of chickens is fed yellow corn. Farmers who own only this strain of chickens

observe the trait to depend entirely on the nature of the diet, specifically whether they feed their chickens yellow corn. Farmers who feed all of their chickens only yellow corn but own several strains of chicken observe the trait to be genetic. What argument could you use to explain to both kinds of farmer that the trait is both environmental and genetic?

3. A newspaper article proclaims that diabetes is neither genetic nor environmental but multicausal. Another article announces that one half of all colon cancer cases are linked to genetic factors. Criticize both messages.

4. Suppose a new treatment for a fatal disease defers the average time before onset of death among those with the disease for 20 years beyond the time when they would have otherwise died. Is it proper to say that this new treatment reduces the risk of death, or does it merely postpone death?

5. It is typically more difficult to study an exposure-disease relation that has a long induction period than one that has a short induction period. What difficulties ensue because the exposure-disease induction period is long?

6. Suppose that both A and B are causes of a disease that is always fatal, so that the disease can occur only once in a single person. Among people exposed to both A and B, what is the maximum proportion of disease that can be attributed to either A or B? What is the maximum for the sum of the amount attributable to A and the amount attributable to B? Suppose that A and B exert their causal influence only in different causal mechanisms, so that they never act through the same mechanism. Would that change your answer?

7. Adherents of induction claim that we all use this method of inference every day. We assume, for example, that the sun will rise tomorrow as it has in the past. Critics of induction claim that this knowledge is based on belief and assumption and that it is no more than a psychological crutch. Why should it matter to a scientist whether scientific reasoning is based on induction or on a different approach, such as conjecture and refutation?

8. Give an example of competing hypotheses for which an epidemiologic study would provide a refutation of at least one.

9. Could a causal association fail to show evidence of a biologic gradient (ie, Hill's fifth criterion)? Explain.

10. Suppose you are studying the influence of socioeconomic factors on cardiovascular disease. Would the study be more informative if (1) the study participants had the same distribution of socioeconomic factors as the general population or (2) the study participants were recruited so that there were equal numbers of participants in each category of the socioeconomic variables? Why?

REFERENCES

1. Rothman KJ. Causes. *Am J Epidemiol.* 1976;104:587–592.
2. Higginson J. Proportion of cancer due to occupation. *Prev Med.* 1980;9:180–188.
3. Ephron E. *The Apocalyptics. Cancer and the Big Lie.* New York: Simon & Schuster, 1984.
4. Rothman KJ: Induction and latent period. *Am J Epidemiol.* 1981;114:253–259.
5. Magee B. *Philosophy and the Real World. An Introduction to Karl Popper.* La Salle, Illinois, Open Court, 1985.
6. Kuhn T. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press, 1962.
7. Feyerabend P. *Against Method.* New York: New Left Books, 1975.
8. Haack S. *Defending Science Within Reason.* Amherst: Prometheus Books, 2003.
9. Haack S. Trial and error: the Supreme Court's philosophy of science. *Am J Public Health.* 2005;95(suppl 1):S66-S73.
10. Mill JS. *A System of Logic, Ratiocinative and Inductive.* 5th ed. London: Parker, Son & Bowin, 1862.
11. Hill AB. The environment and disease: Association or causation? *Proc R Soc Med.* 1965;58:295–300.
12. U.S. Department of Health, Education and Welfare. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service.* Public Health Service Publication No. 1103. Washington, DC: Government Printing Office, 1964.
13. Weiss NS. Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology.* 2002;13:6–8.

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

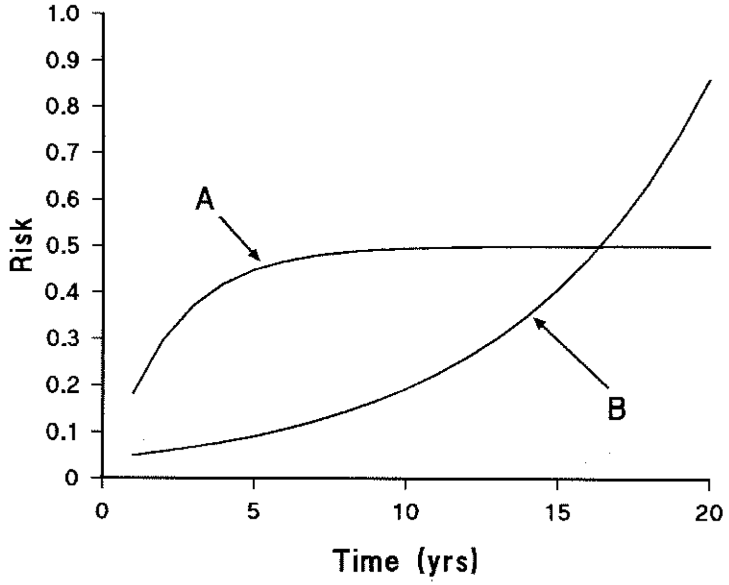

Figure 4–1 Two possible patterns of disease risk with time.

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

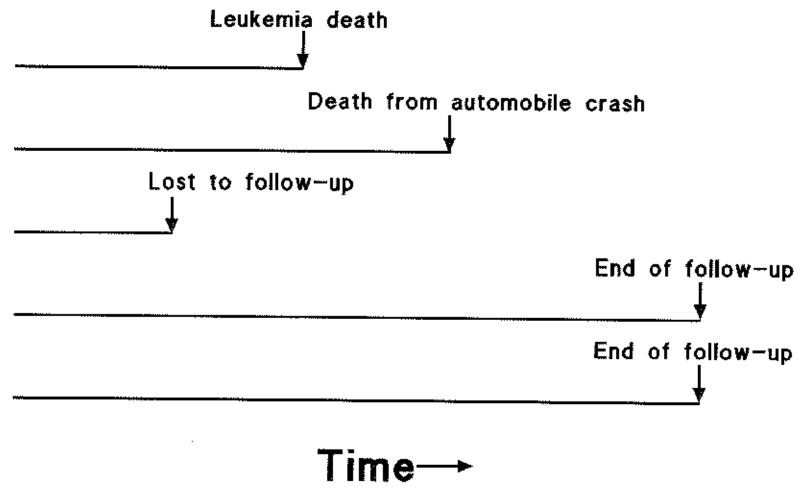

Figure 4–2 Time at risk for leukemia death for five people.

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 $yr^{-1} \times 20$ 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

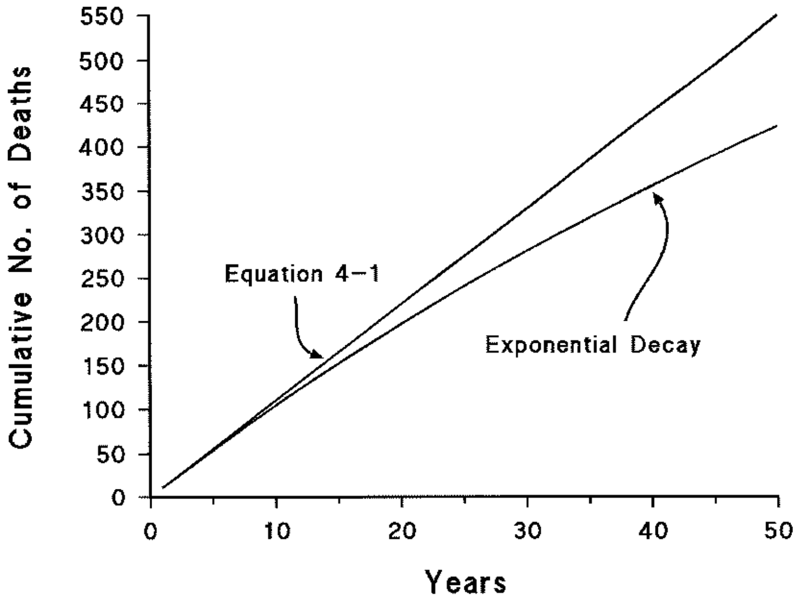

Figure 4–3 Cumulative number of deaths among 1000 people with a mortality rate of 11 deaths per 1000 person-years, presuming no population shrinkage (see Equation 4–1) and taking the population shrinkage into account (ie, exponential decay).

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.

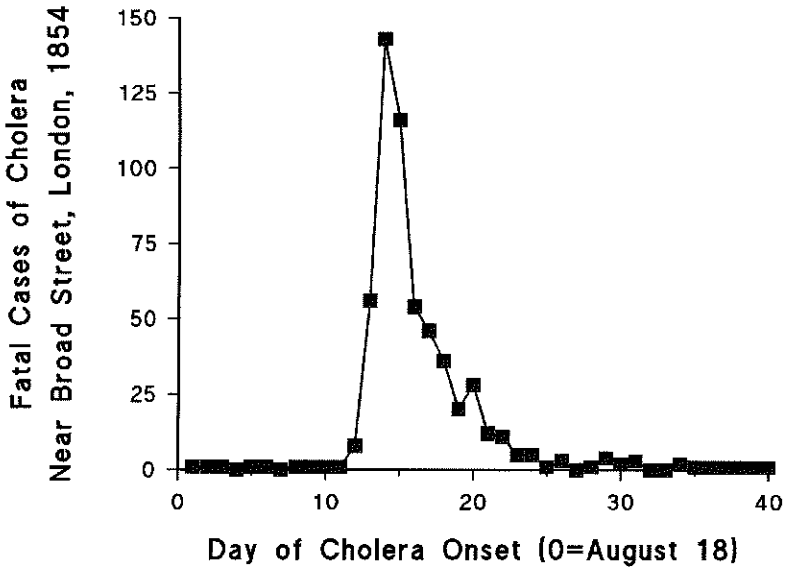

Figure 4–4 Epidemic curve for fatal cholera cases during the Broad Street outbreak in London in 1854.

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,

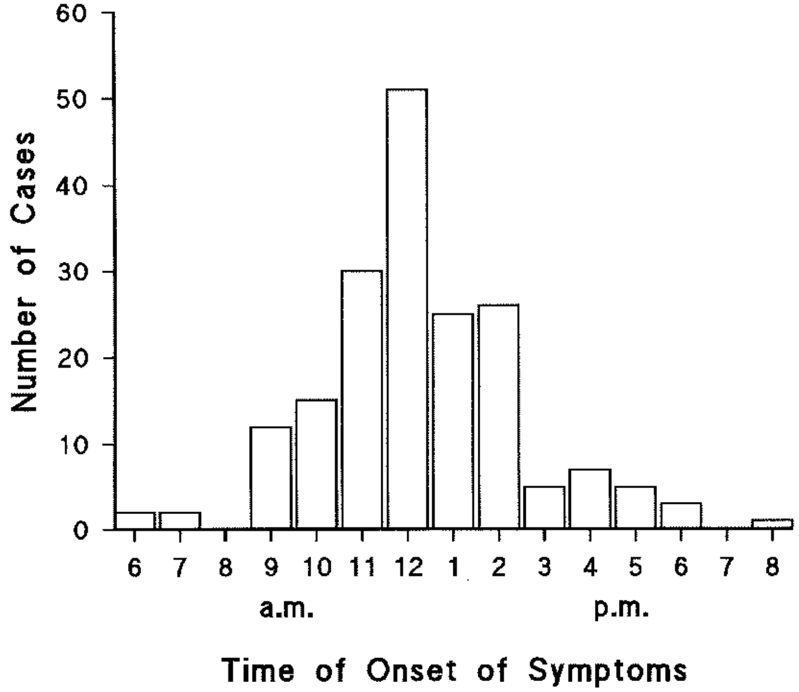

Figure 4–5 Epidemic curve for an outbreak of hysteria among elementary school children on November 6, 1985.

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 one study group at a given time. The time sequence may affect the interpretation, and the passage of time means that the two experiences that are compared may differ by factors other than the exposure. The counterfactual setting is impossible, because it implies that a person lives through the same experience twice during the same time period, once with exposure and once without exposure.

In the theoretical ideal of a counterfactual study, each exposed person would be compared with his or her unexposed counterfactual experience. Everyone is exposed, and in a parallel universe everyone is also unexposed, with all other factors remaining the same. The effect of exposure could then be measured by comparing the incidence proportion among everyone while exposed with the incidence proportion while everyone is unexposed. Any difference in these proportions would have to be an effect of exposure. Suppose we observed 100 exposed people and found that 25 developed disease in 1 year, providing 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 being unexposed. Suppose that 10 people developed disease in those counterfactual conditions. 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 to it in the design of epidemiologic studies. Instead of comparing the experience of an exposed group with its counterfactual ideal, we must compare their experience with that of a real unexposed population. The goal is to find an unexposed population that would give a result that is close, if not identical, to that from a counterfactual comparison.

Suppose we consider the same 100 exposed people mentioned earlier, 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 represent 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 cases in the unexposed group is attributable to the exposure? Perhaps the exposure had no effect but the unexposed group was 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 had a strong effect that was balanced by the fact that the unexposed group had a higher risk for disease.

To achieve a valid substitution for the counterfactual experience, we resort to various design methods that promote comparability. One example is the crossover trial, which is based on comparison of the experience of each exposed person with himself or herself at a different time. But a crossover trial is feasible only for an exposure that can be studied in an experimental setting (ie, assigned by the investigator according to a study protocol) and only if it has a brief effect. A persistent exposure effect would distort the effect of crossing over from the exposed to the unexposed group. 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 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 to evaluate 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* (*RD*) is the difference in incidence proportion or risk between the exposed and the unexposed groups. If the incidence proportion is 0.25 for the exposed and 0.10 for the unexposed, the *RD* is 0.15. With an incidence rate instead of a risk to measure disease occurrence, we can likewise calculate the *incidence rate difference* (*IRD*) for the two measures. (Terminology note: In older texts, the *RD* is sometimes referred to as the *attributable risk.* The *IRD* also has been called the *attributable rate.*)

Difference measures such as *RD* and *IRD* measure the absolute effect of an exposure. It is also possible to measure the relative effect. As an analogy, consider how to assess the performance of an investment over a period of time. Suppose that an initial investment of $100 became $120 after 1 year. The difference in the value of the investment at the end of the year and the value at the beginning, $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, but 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. We first obtain an absolute measure of effect, which can be the *RD* or the *IRD*, and we then divide that by the measure of occurrence of disease among unexposed persons. For risks, the relative effect is given by the following equation:

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

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

$$\text{Relative effect} = \frac{RD}{R_0} = \frac{R_1 - R_0}{R_0} = RR - 1 \qquad [4\text{–}5]$$

In Equation 4–5, the *risk ratio* (*RR*) is defined as R_1/R_0. The relative effect is the risk ratio minus 1 (*RR* – 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 gives (\$120/\$100) – 1 = 1.2 – 1 = 20%. If the risk in exposed people is 0.25 and that in unexposed people is 0.10, 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 (which is the value of the *RR* when there is no effect). By defining the relative effect in this way, we ensure that we have a relative effect of zero when the absolute effect is also zero.

Although the relative effect is *RR* – 1, it is common for epidemiologists to refer to the *RR* itself as a measure of relative effect, without subtracting 1. When the *RR* is used in this way, it is important to remember that a value of 1 corresponds to the absence of an effect. For example, an *RR* of 3 represents twice as great an effect as an *RR* of 2. Sometimes, epidemiologists refer to the percentage increase in risk to convey the magnitude of relative effect. For example, they may describe an effect that represents a 120% increase in risk. 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 the same as the relative effect defined previously. An increase of 120% corresponds to an *RR* of 2.2, which is 2.2 – 1.0 = 120% greater than 1. 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, for example, it means that the *RR* is 5. If the effect is described as a 10% increase in risk, it corresponds to an *RR* of 1.1, which is 1.1 – 1.0.

Effect measures that involve the *IRD* and the incidence rate ratio are defined analogously to those involving the *RD* and the risk ratio. Table 4–5 compares absolute and relative measures constructed from risks and from rates.

The range of the *RD* measure derives from the range of risk itself, which is [0,1]. The lowest possible *RD* 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 *RD*, 1, comes from an exposed group with 100% risk and an unexposed group with zero risk. *RD* has no dimensionality (ie, 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 the dimensionality of the underlying measure.

Table 4–5 Comparison of Absolute and Relative Effect Measures

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

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 ∞. 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, the numerator, the denominator, and the ratio are all dimensionless.

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

When to Use Absolute and Relative Effect Measures

Absolute and relative effect measures provide different messages. When measuring the effect of an exposure on the health of a population, an absolute effect measure is needed. It reflects added or diminished disease burden in that population in terms of an increased risk or incidence rate or, for protective exposures, a decreased risk or incidence rate. The public-health implications of any exposure need to be assessed in terms of the absolute effect measures.

Relative effect measures convey a different message. The attributable fraction among exposed people, (*RR* –1)/*RR*, is purely a function of the relative effect measure, which gives a clue about the message of relative effect measures. These measures indicate the extent to which the exposure in question accounts for the occurrence of disease among the exposed people who get disease. The relative measure itself expresses this relation on a scale that goes from zero to infinity, and the attributable fraction converts it to a proportion, but both convey a message about the extent to which disease among the exposed population is a consequence of exposure.

It is important to realize that a relative effect may be extremely large but with little public-health consequence. If an exposure has a rate ratio of 10 for an extremely rare disease, the 10-fold increase in disease implies that the exposure accounts for almost all the disease among the exposed; however, even among exposed the disease may remain rare. Such an exposure may have less public-health consequence than another exposure that merely doubles the rate of a much more common disease.

In case-control studies (see Chapter 5), usually only relative effects are directly obtainable. Nevertheless, by taking into account the overall rate or risk of disease occurrence in a population, the relative measures obtained from case-control studies can be converted into absolute measures, which are needed to assess appropriately the public-health impact of an exposure.

Examples

Table 4–6 presents data on the risk of diarrhea among breast-fed infants during a 10-day period after their infection with *Vibrio cholerae 01* according to the level

Table 4–6 Diarrhea During a 10-day Follow-up Period in Breast-fed Infants Colonized with *Vibrio cholera 01* According to 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

Reproduced with permission from Glass RI et al.[7]

of antipolysaccharide antibody titers in their mother's breast milk.[7] 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 *RD* for infants exposed to milk with low compared with high levels of antibody is 0.86 − 0.44 = 0.42. This *RD* 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; it assumes that the infants exposed to low titers would have experienced a risk equal to that of those exposed to high titers were it not for the lower antibody levels.

We can also measure the effect of low titers on diarrhea risk in relative terms. The risk ratio, *RR*, is 0.86/0.44 = 1.96. The relative effect is 1.96 − 1, or 0.96, indicating a 96% greater risk of diarrhea among infants exposed to low antibody titers in breast milk. Commonly, we would describe the risk among the 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 4–7 gives data for the incidence rate of breast cancer among women who were treated for tuberculosis early in the 20th century.[8] Some women received a treatment that involved repeated fluoroscopy of the lungs, with a resulting high dose of ionizing radiation to the chest.

Table 4–7 Breast Cancer Cases and Person-Years of Observation for Women with Tuberculosis Who Were Repeatedly Exposed to Multiple X-ray Fluoroscopies and for 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-years)	14.6	7.9	11.9

Reproduced with permission from Boice and Monson.[8]

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 *IRD* 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 the radiation exposure and assumes that the exposed women would have had a rate equal to that among the unexposed women were it not for the exposure. 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 can be expressed as an 86% greater rate of breast cancer among women exposed to the radiation. Alternatively, the incidence rate ratio can be described as indicating a rate of breast cancer among exposed women that is 1.86 times that of the rate among unexposed women.

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 may be reported as 4.1; in others, the same number may be reported as 4.0846. 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 can 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 that, for a given study, reporting should carry into the first decimal (eg, 4.1). If the study reported risk ratios and took on values lower than 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. Therefore, a simple rule such as one decimal place (for example) 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 not 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. Because 1.0 is the zero point for ratio measures of effect, this rule treats positive effects near zero differently from negative effects. If all 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. Values used in intermediate calculations should never be rounded; one should round only in the final step before reporting. 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. Under such a strategy, both 1.75 and 1.85 would be rounded to 1.8.

The 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 can 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 4–1, we know that a constant incidence rate will result in a risk that is approximately equal to the product of the rate and the time period, provided that the time period is short enough so that the risk remains less than about 0.20. For greater values, the approximation does not work well. Assuming that we are dealing with short time periods, the ratio of the risk among the exposed to the risk among the unexposed, R_1/R_0, will be as follows:

$$\text{Risk ratio} = \frac{R_1}{R_0} \approx \frac{I_1 \times \text{time}}{I_0 \times \text{time}} = \frac{I_1}{I_0}$$

This relation shows that the risk ratio is nearly 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 4–1 to apply. The shorter the time period or the lower the rates, the better the approximation represented by Equation 4–1 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 4–6, for example. The risk in the high-titer antibody group (considered 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. 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, and when risk among the unexposed is high, we can expect there to be a divergence between the incidence rate ratio and the risk ratio. Suppose the incidence rates that gave rise to the risks in Table 4–6 were constant over the 10-day follow-up period. If we take into account the exponential-decay relation between risk and rate, we can back-calculate from the risks in Table 4–6 to the underlying rates based on the exponential decay curve, and from that result, we can calculate that the ratio of those underlying rates is 3.4, compared with the 1.96 for the ratio of risks. This large discrepancy arises because the risks are large.

If the time period over which a risk is calculated approaches 0, the risk itself also approaches 0; the risk of a given person having a myocardial infarction may be 10% over a decade, but over the next 10 seconds, it will be extremely small, and its value will shrink 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 these risks approach 0, the risk ratio approaches the value of the incidence rate ratio. The incidence rate ratio is the limiting value for the risk ratio as the time interval over which the risks are taken approaches 0. We therefore can describe the incidence rate ratio as an *instantaneous risk ratio*. This equivalence of the two types of ratios 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 the term relative risk, because 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. The often-used notation *RR* is sometimes read to mean relative risk, which equally can 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 people 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. These effect values may be 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 *RD* between those exposed and those unexposed to asbestos is sure to be greater among older adults than younger adults, and 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, and 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 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 *RD* 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 4–5). We can also divide the *RD* by the risk in exposed people to get an expression that we refer to as the *attributable fraction*:

$$\text{Attributable fraction} = \frac{RD}{R_1} = \frac{R_1 - R_0}{R_1} = 1 - \frac{1}{RR} = \frac{RR - 1}{RR} \qquad [4\text{–}6]$$

If the *RD* reflects a causal effect that is not distorted by any bias, the attributable fraction is a measure that quantifies the proportion of the disease burden among exposed people that is caused by the exposure. Consider the hypothetical data in Table 4–8. The risk of disease during a 1-year period is 0.05 among the exposed and 0.01 among the unexposed. Suppose that this difference can

Table 4–8 Hypothetical Data Giving 1-Year Disease Risks for Exposed and Unexposed People

	Exposure		
	No	Yes	Total
Disease	900	500	1,400
No disease	89,100	9,500	98,600
Total	90,000	10,000	100,000
Risk	0.01	0.05	0.014

be reasonably attributed to the effect of the exposure (because we believe that we have accounted for all substantial biases). The *RD* is 0.04, which is 80% of the risk among the exposed. We would then say that the exposure appears to account for 80% of the disease that occurs among exposed people during the 1-year period. Another way to calculate the attributable fraction is from the risk ratio: $(5 - 1)/5 = 80\%$. (Terminology note: The *attributable fraction* sometimes is referred to in older texts as the *attributable risk percent* or *attributable risk.*)

To calculate the attributable fraction for the entire population of 100,000 people in Table 4–8, we first calculate the attributable fraction for exposed people. To get the overall attributable fraction for the total population, the fraction among the exposed is multiplied by the proportion of all cases in the total population who are exposed. There are 1400 cases in the entire population, of whom 500 are exposed. 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 cases who are exposed: $0.8 \times 0.357 = 0.286$; 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. Among all cases, only some of the exposed cases can be attributable to the exposure. There are 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. Therefore, 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 the following equation, which takes into account each of the exposure levels:

$$\text{Total attributable fraction} = \sum_i (AF_i \times P_i) \qquad [4\text{–}7]$$

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 is 0.

Equation 4–7 can be applied to the hypothetical data in Table 4–9, which describe 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

Table 4–9 Hypothetical Data Giving 1-Year Disease Risks for People at Three Levels of Exposure

	Exposure			
	None	Low	High	Total
Disease	100	1,200	1,200	2,500
No disease	9,900	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	

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

The same result can be calculated directly from the number of attributable cases at each of the exposure levels:

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

Under certain assumptions, estimation of attributable fractions can be based on rates as well as risks. In Equation 4–6, which uses the risk ratio to calculate the attributable fraction, the rate ratio can 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, Greenland and Lash.[9]

QUESTIONS

1. Suppose that in a population of 100 people, 30 die. The risk of death can 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 of 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 4–6, 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 4–7 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 to look at, 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 magnitude in the two countries? Why or why not?

REFERENCES

1. Gaylord Anderson, as cited in Cole P. The evolving case-control study. *J Chron Dis.* 1979;32:15–27.
2. Iskrant AP, Joliet PV. *Accidents and Homicides.* Cambridge, MA: Harvard University Press; 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. Poser CM, Hibberd PL, Benedikz J, Gudmundsson G. *Neuroepidemiology.* 1988;7:168–180.
6. Cole TB, Chorba TL, Horan JM. Patterns of transmission of epidemic hysteria in a school. *Epidemiology.* 1990;1:212–218.
7. 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.
8. Boice JD, Monson RR. Breast cancer in women after repeated fluoroscopic examinations of the chest. *J Natl Cancer Inst.* 1977;59:823–832.
9. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

5

Types of Epidemiologic Studies

Chapter 4 described measures of disease frequency, including risk, incidence rate, and prevalence; measures of effect, including risk and incidence rate differences and ratios; and 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. More specialized study designs, such as two-stage designs and ecologic studies, are discussed in *Modern Epidemiology.*[1]

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."[2] 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 may 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 exploring these