3. MEASURES OF DISEASE FREQUENCY

The clearest of many definitions of epidemiology that has been proposed has been attributed to Gaylord Anderson [Cole, 1979]. His definition is

> Epidemiology: the study of the occurrence of illness

Other sciences are also directed toward the study of illness, but in epidemiology the focus is on the *occurrence* of illness. As a branch of science, epidemiology deals with the evaluation of scientific hypotheses. These hypotheses are often posed as qualitative propositions. The "null" form of such propositions is highly refutable and, as discussed in the previous chapter, derives its empirical content from this characteristic. Unlike the framing of hypotheses, scientific research, which comprises the activity of attempted refutation of hypotheses, is predicated on measurement. Qualitatively stated hypotheses about evolution, the formation of the earth, the effect of gravity on light waves, or the method by which birds find their way during migration are all tested by measurements of the phenomena that relate to the hypotheses. The physicist Kelvin aptly stated the importance of measurement in science [cited in Beiser, 1960]:

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of Science, whatever the matter may be.

From Hippocrates to Sydenham, physicians have considered the causes of disease, but it was only when measurement of the occurrence of disease replaced reflection about causation that scientific knowledge about causation made impressive strides. The fundamental task in epidemiologic research is thus to quantify the occurrence of illness. The goal is to evaluate hypotheses about the causation of illness and its sequelae and to relate disease occurrence to characteristics of people and their environment.

There are three basic measures of disease frequency. *Incidence rate* is a measure of the instantaneous force of disease occurrence. *Cumulative incidence* measures the proportion of people who convert, during a specified period of time, from nondiseased to diseased. *Prevalence* measures the proportion of people who have disease at a specific instant. These measures and their interrelation will be described in detail.

INCIDENCE

In attempting to measure the frequency of disease occurrence in a population, it is insufficient merely to record the number of people or the proportion of the population that is affected. It is also necessary to take into

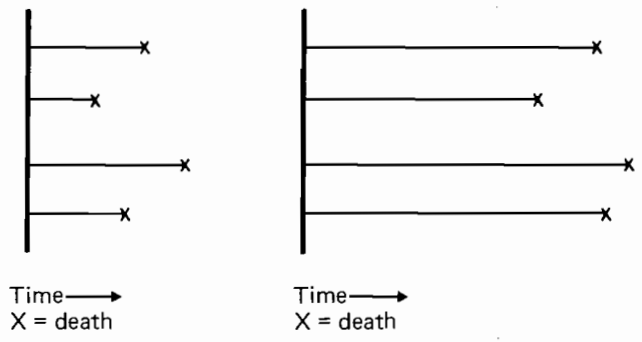

Fig. 3-1. Two different patterns of disease occurrence.

sider the frequency of a disease that ultimately affects all people, namely, death. Since all people are eventually affected, the time from birth to death becomes the determining factor in measuring the occurrence of death. Time differentiates between the two situations shown in Figure 3-1.

Thus, an incidence measure must take into account the number of individuals in a population that becomes ill and the time periods experienced by members of the population during which these events occur. *Incidence rate* is therefore defined as the number of disease onsets in the population divided by the sum of the time periods of observation for all individuals in the population:

$$\text{Incidence rate} = \frac{\text{no. disease onsets}}{\Sigma \text{ time periods}}$$

where Σ indicates the sum of time periods for all individuals.

For many epidemiologic applications, the possibility of a person getting a disease more than once is ruled out by either convention or biology. If the disease is rhinitis, we may simply wish to measure the incidence of "first" occurrence, even though disease can occur repeatedly; for cancer, heart disease, and many other illnesses, first occurrence is often of greater interest for study than subsequent occurrences in the same individual. For an outcome such as death or a disease such as diabetes, which is considered not to recur but to be a permanent state once diagnosed, only first occurrence can be studied. When the events tallied are first occurrences of disease, then the observation period for each individual who develops the disease terminates with the onset of disease.

Because incidence rate is a quotient with a frequency in the numerator and a measure of time in the denominator, its dimensionality is time^{-1}, that is, the reciprocal of time. The

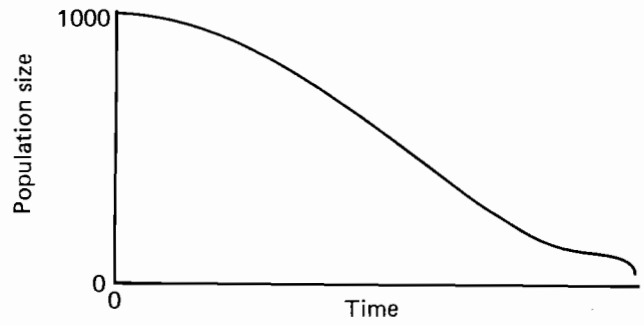

Fig. 3-2. Size of a fixed population of 1,000 people, by time.

considered a product of population size by the average time period of observation for a member of the population, although this product is, like any product, only a shorthand description of the appropriate summation. The denominator of the incidence rate is often referred to as a measure of "person-time" to distinguish the time summation from ordinary clock time. The person-time measure forms the observational experience in which disease onsets can be observed. Implicit in the measure is the concept that a given amount of person-time, say 100 person-years, can be derived from observing a variety of populations in a variety of circumstances. That is, the observations of 100 persons for 1 year, 50 persons for 2 years, 200 persons for 6 months, or one person for 100 years are assumed to be equivalent. One unit of person-time is assumed to be equivalent to and independent of another unit of person-time. This assumption, although generally a reasonable one, could be unwarranted in extreme situations—for example, observing one individual for 100 years to obtain 100 person-years. Usually the units of person-time are restricted by age, which eliminates extreme departures from independence of the person-time units. One could not obtain 100 person-years of experience in the age range 50 to 54 years with fewer than 20 individuals.

Conceptually we can imagine the person-time experience of two distinct types of populations, the *fixed population* and the *dynamic population*. A fixed population adds no new members, whereas a dynamic population does. Suppose we are measuring the *mortality rate,* defined as the incidence rate of death, in a fixed population of 1,000 people. After a period of sufficient time, the original 1,000 will have dwindled to zero. A graph of the size of the population with time might look like that in Figure 3-2.

The curve slopes downward because the 1,000 individuals eventually all die. The population is fixed in the sense that we consider the fate of only the 1,000 individuals initially identified. The person-time experience of

sloping curve in the diagram. As each individual dies, the curve notches downward; that individual no longer contributes to the person-time observation pool of the fixed population. Each individual's contribution is exactly equal to the length of time that individual is followed from start to finish; in this example, since the entire population is followed until death, the finish is the individual's death. In other instances, the contribution to the person-time experience would continue until the onset of disease or some arbitrary cutoff time for observation, whichever came sooner.

Suppose we added up the total person-time experience of this fixed population of 1,000 and obtained a total of 75,000 person-years. The mortality rate would be (1,000/75,000)year^{-1} since the 75,000 person-years represent the experience of all 1,000 people until their deaths. A fixed population facing a constant death rate would decline exponentially in size, but in practice "exponential decay" virtually never occurs. Because a fixed population ages steadily during the observation period, the death or disease rate in a fixed population generally changes with time because of the change in age. *Life-table* methodology is a procedure by which the mortality (or morbidity) of a fixed population is evaluated within successive small time intervals so that the time dependence of mortality can be elucidated.

A dynamic population differs from a fixed population in that we do not restrict the observations to any fixed group. Instead, we extend the observations to those entering the population as observation time proceeds. People enter a population in various ways. Some are born into it; others migrate into it. For a population of people of a specific age, individuals also enter the population by aging into it. Similarly, individuals can exit from the person-time observational experience by dying, aging out of a defined age group, emigrating, and becoming diseased, if only first bouts of a disease are being studied. If the number of people entering a population is exactly balanced by the number exiting the population in any period of time, the population is said to be in a *steady state*. Steady state is a property that applies only to dynamic populations, not to fixed populations.

The graph of the size of a dynamic population in steady state is simply a horizontal line. People are continually entering and leaving the person-time experience in a way that might be diagrammed as shown in Figure 3-3.

In the diagram, the symbol > represents an individual entering the person-time experience, a line segment represents that individual's contribution to the person-time experience, the termination of a line segment indicates removal from the person-time experience, and X indicates removal from the person-time experience because of disease onset. In theory, if the incidence rate is constant during time, any portion of the population-time experience of a dynamic population in a steady state will

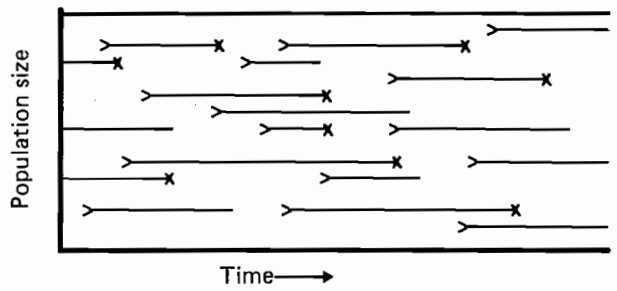

Fig. 3-3. Size of a dynamic population, by time, with an indication of population turnover.

provide a good estimate of disease incidence. The value of incidence will be the ratio of the number of cases of disease onset, indicated by X, to the two-dimensional (population × time) area. Because this ratio is equivalent to the density of disease onsets in the observational area, the incidence rate has also been referred to as *incidence density* [Miettinen, 1976]. Another synonym for the measure is *force of morbidity* (or *force of mortality* in reference to deaths).

The numerical range for incidence rate is zero to infinity, corresponding to the range of densities of points in two-dimensional space. How can disease incidence be infinite? Infinity is the theoretical upper limit for a disease that is universal and strikes quickly. If a population in a space colony were suddenly all exposed without protective gear to the environment of space, the incidence rate of death would be extremely high, though not quite at infinity, because death would not be instantaneous. The limiting value of infinity is approached only at the instant of some sudden holocaust. To some it may be surprising that an incidence rate can exceed the value of 1.0, which would seem to indicate that more than 100 percent of a population is affected. It is true that at most only 100 percent of a population can get a disease, but the incidence rate does not measure the proportion of a population with illness. The measure is not a proportion—recall that incidence rate is measured in units of the reciprocal of time. Among 100 people, no more than 100 deaths can occur, but those 100 deaths can occur in 10,000 person-years, in 1,000 person-years, in 100 person-years, or even in 1 person-year (if the 100 deaths occur after an average of 3.65 days each). An incidence rate of 100 cases (or deaths) per 1 person-year might be expressed as

$$100 \ \frac{\text{cases}}{\quad}$$

It might also be expressed as

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

$$8.33 \frac{\text{cases}}{\text{person-month}} \quad \text{or}$$

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

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

The numerical value of an incidence rate in itself has no interpretability because it depends on the arbitrary selection of the time unit. It is essential in presenting incidence rates to give the appropriate time units, either as the examples given above or as in 8.33 month^{-1} or 1.92 week^{-1}. In epidemiologic writing, the units are often given only implicitly rather than explicitly, as in "an annual incidence of 50 per 100,000." The latter quantity is equivalent to

$$\frac{50}{100{,}000} \frac{\text{cases}}{\text{person-years}} \quad \text{or} \quad 5 \times 10^{-4}\ \text{year}^{-1}$$

It is preferable, however, not to use an expression such as "annual incidence of"; this description is analogous to describing a velocity of 60 miles/hr as "an hourly velocity of 60 miles." Aside from being clumsy, it makes an inappropriate implication about time, as if the measure applied to the entire stated interval of time when in fact it does not. A velocity of 60 miles/hr does not apply to an hour of time; one need not travel at the velocity for an hour nor spend an hour to measure it. The velocity of 60 miles/hr is an instantaneous concept: One can readily conceive of traveling at that velocity at a specific instant in time. Whether the velocity is expressed as 60 miles/hr or 88 feet/sec or 0.57 astronomical units/century makes no difference; the same speed is indicated, and the units of time used to express it have no bearing on the instantaneous nature of the measure. The same principle applies to incidence rate [Elandt-Johnson, 1975]. Like velocity, it is always an instantaneous concept, even with units of person-years or person-centuries. Thus, there is nothing annual about an "annual incidence," and it would be preferable not to use such terminology.

The dimensionality of incidence rate, that is, the reciprocal of time, makes it an awkward measure to absorb intuitively. The measure does,

however, have an interpretation. Referring back to Figure 3-2, one can see that the area under the curve is equal to N(T), where N is the number of people in the fixed population and T is the average time until death. This is equivalent to saying that the area under the curve is equal to the area of a rectangle with height N and width T. Since T is the average time until death for N people, the total person-time experience is N(T). The time-averaged mortality rate at complete follow-up, then, is N/[N(T)] = 1/T; that is, the mortality rate equals the reciprocal of the average time until death, or, more generally, incidence rate equals the reciprocal of the average time until disease onset [Morrison, 1979]. Thus, a mortality rate of 0.04 yr^{-1} indicates an average time until death of 25 years. If the outcome is not death but either disease onset or death only from a specific cause, the interpretation above must be modified slightly. The time period at issue is then the average time until disease onset, assuming that a person is not at risk of other causes of death. That is, the measure is a time conditional on no other *competing risks* of death. This interpretation of incidence rates as the inverse of the average "waiting time" will not be valid unless the incidence rate can be used to describe a population in steady state or a fixed population with complete follow-up. For example, the mortality rate for the United States in 1977 was 0.0088 year^{-1}, suggesting a mean lifespan, or expectation of life, of 114 years. Other analyses indicate that the actual expectation of life in 1977 was 73 years. The discrepancy is due to the lack of a steady state.

CUMULATIVE INCIDENCE

Despite the interpretation that can be given to incidence rate, it is occasionally more convenient to use a more readily interpretable measure of disease occurrence. Such a measure is the *cumulative incidence,* which may be defined as the proportion of a fixed population that becomes diseased in a stated period of time. If *risk* is defined as the probability of an individual developing disease in a specified time interval, then cumulative incidence is a measure of average risk. Like any proportion, the value of cumulative incidence ranges from zero to 1 and is dimensionless. It is uninterpretable, however, without specification of the time period to which it applies. A cumulative incidence of death of 3 percent may be low if it refers to a 40-year period, whereas it would be high if it applies to a 40-day period.

It is possible to derive estimates of cumulative incidence from incidence rate. Consider a fixed population (Fig. 3-4).

At time t, $CI_t = (P_0 - P_t)/P_0$; in words, the cumulative incidence at time t equals the number of people who have exited the fixed population by time t because of disease ($P_0 - P_t$) divided by the initial number of people

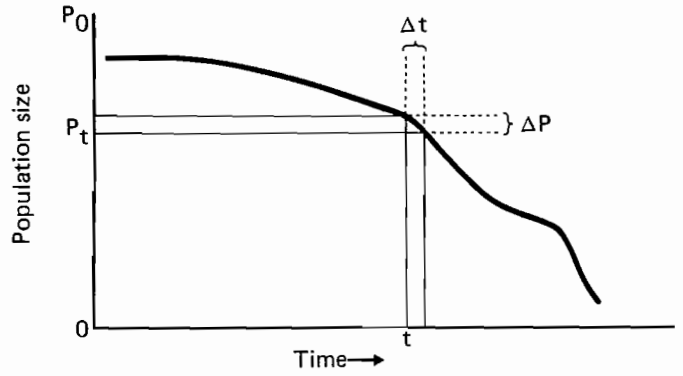

Fig. 3-4. Size of a fixed population, by time, indicating a small decrement at time t.

in the population. The incidence rate at time t is the ratio of new cases to the person-time observation experience; thus

$$I_t = \frac{-\Delta P}{P_t \Delta t}$$

or, written in terms of differential calculus,

$$I_t = \frac{-dP}{P_t dt} \qquad -I_t dt = \frac{dP}{P_t}$$

(The minus sign is used because the change in P is negative in relation to t; without the minus sign, the incidence measure would be negative.) Integrating both sides,

$$-\int_0^t I_t dt = \ln(P_t) - \ln(P_0)$$

Taking antilogs,

$$\exp(-\int_0^t I_t dt) = P_t/P_0$$

and since

$$CI_t = \frac{P_0 - P_t}{P_0}$$

we have

$$CI_t = 1 - \exp(-\int_0^t I_t dt)$$

This is estimated as

$$CI_t = 1 - \exp(-\sum_i I_i \Delta t_i)$$

where the summation of the index, i, is over categories of time covering the interval [0,t].

For a constant incidence rate,

$$CI_t = 1 - e^{-I\Delta t}$$

Because $e^x \doteq 1 + x$ for $|x|$ less than about 0.1, a good approximation for a small cumulative incidence (less than 0.1) is

$$CI_t \doteq \sum_i I_i \Delta t_i \qquad \text{or} \qquad CI_t \doteq I\Delta t$$

if the rate is constant with time. Thus, to estimate small risks, one can simply multiply the incidence rate by the time period. The above approximation offers another interpretation for the incidence rate; it can be viewed as the ratio of a short-term risk to the time period for the risk as the duration of the time period approaches zero.

The cumulative incidence measure is premised on the assumption that there are no competing risks of death. Thus, if an individual at age 40 faces a cumulative incidence, or risk, of 35 percent in 30 years for cardiovascular death, this is interpreted as the probability of dying from cardiovascular disease given that the individual is free from other risks of death. Because no one is actually free from competing risks, the cumulative incidence measure for any outcome other than death from all causes is a hypothetical measure. In principle, cumulative incidence for lengthy periods is unobservable and must be inferred because of the influence of competing risks.

A specific type of cumulative incidence is the *case fatality rate,* which is the cumulative incidence of death among those who develop an illness (it is therefore technically not a rate but a proportion). The time period for measuring the case fatality rate is often unstated, but it is always better to specify it. When unstated, presumably there is a short period of increased risk. For long periods of risk of death after disease onset, it is preferable to use the mortality rate among those with the illness rather than the case fatality rate, so that the actual time at risk for each individual can be taken

erage time elapsed until the event, the overall mortality rate of a disease in a population is related to the incidence rate and the mortality rate among cases as follows [Morrison, 1979]:

$$M_T = \frac{1}{T} = \frac{1}{T_1 + T_2} = \frac{1}{1/I + 1/M_c}$$

where M_T is the total population mortality rate, T is the life expectancy, T_1 is the average time until disease onset, T_2 is the average time from disease onset to death, I is the incidence rate of disease, and M_c is the mortality rate among cases.

PREVALENCE

Unlike incidence measures, which focus on events, *prevalence* focuses on disease status. Prevalence may be defined as the proportion of a population that is affected by disease at a given point in time. The term *point prevalence* is sometimes used to mean the same thing. An individual that dies from an illness is thereby removed from the group that constitutes the numerator of prevalence; consequently, mortality from an illness decreases prevalence. Diseases with large incidence rates may have low prevalences if they are soon fatal. People may also exit from the prevalence pool by recovering from disease.

Earlier it was stated that a population in steady state has an equal number of people entering and exiting during any unit of time. This concept can be extended to refer to a subpopulation of ill people, or a *prevalence pool* (i.e., the numerator of a prevalence). In a steady state, the number of people entering the prevalence pool is balanced by the number exiting from it:

Inflow (to prevalence pool) = outflow (from prevalence pool)

People enter the prevalence pool from the nondiseased population. If the total number of people in a population is N and the prevalence pool is P, then the size of the nondiseased population that "feeds" the prevalence pool is N − P. During any time interval, Δt, the number of people who enter the prevalence pool is

$$I\Delta t(N - P)$$

where I is the incidence rate. During the same time interval Δt, the outflow from the prevalence pool is

where I′ represents the incidence rate of exiting from the prevalence pool, that is, the number who exit divided by the person-time experience of those in the prevalence pool. Earlier we saw that the reciprocal of an incidence rate in a steady state equals the mean duration of time spent before the incident event. Therefore, the reciprocal of I′ is the mean duration of illness, $\bar{D}$. Thus,

$$\text{Inflow} = I\Delta t(N - P) = \text{outflow} = (1/\bar{D})\Delta tP$$

$$I\Delta t(N - P) = (1/\bar{D})\Delta tP$$

$$P/(N - P) = I\bar{D}$$

$P/(N - P)$ is the ratio of ill to not-ill (we could call them healthy except that we mean they are not ill from a specific illness, which doesn't imply an absence of all illness) people in the population, or equivalently, the ratio of prevalence to the complement of prevalence (1 − prevalence). The ratio of a proportion to the quantity 1 minus the proportion is referred to as *odds*. In this case, $P/(N - P)$ is the *prevalence odds,* or odds of having a disease relative to not having the disease. Thus, the prevalence odds equals the incidence rate times the mean duration of illness. If the prevalence is small, say less than 0.1, then it follows that

$$\text{Prevalence} \doteq I\bar{D}$$

since prevalence will approximate the prevalence odds for small values of prevalence. More generally [Freeman and Hutchison, 1980],

$$\text{Prevalence} = \frac{I\bar{D}}{1 + I\bar{D}}$$

which can be obtained from the above expression for prevalence odds.

Prevalence, being a proportion, is dimensionless, with a range of zero to 1.0. The above equations are in accord with these requirements, because in each of them the incidence rate, with a dimensionality of the reciprocal of time, is multiplied by the mean duration of illness, giving a dimensionless product. Furthermore, the product has the range of zero to infinity, which corresponds to the range of prevalence odds, whereas the expression

$$\frac{I\bar{D}}{1 + I\bar{D}}$$

is always in the range zero to 1.0.

Seldom is prevalence of direct interest in etiologic applications of epi-

prevalence, studies of prevalence, or studies based on prevalent cases, yield associations that reflect the determinants of survival with disease just as well as the causes of disease. Better survival and therefore a higher prevalence might indeed be related to the action of preventives that somehow mitigate the disease once it occurs.

Nevertheless, for one class of diseases, namely, congenital malformations, prevalence is the measure usually employed. The proportion of babies born with some malformation is a prevalence, not an incidence rate. The incidence of malformations refers to the occurrence of the malformations among the susceptible populations of embryos. Many malformations lead to early embryonic or fetal death that is classified, if recognized, as a miscarriage rather than a birth. Thus, malformed babies at birth represent only those individuals who survived long enough with their malformations to be recorded as a birth. This is indeed a prevalence measure, the reference point in time being the moment of birth. Generally, it would be more useful and desirable to study the incidence than the prevalence of congenital malformations, but usually this is not possible. Consequently, in this area of research, prevalent rather than incident cases are studied.

Prevalence is sometimes used to measure the occurrence of nonlethal degenerative diseases with no clear moment of onset. In this and other situations, prevalence is measured simply for convenience, and inferences are made about incidence by using assumptions about the duration of illness. Of course, in epidemiologic applications outside of etiologic research, such as planning for health resources and facilities, prevalence may be a more germane measure than incidence.

REFERENCES

Cole, P. The evolving case-control study. *J. Chron. Dis.* 1979; 32:15–27.

Beiser, A. *The World of Physics.* New York: McGraw-Hill, 1960.

Elandt-Johnson, R. C. Definition of rates: Some remarks on their use and misuse. *Am. J. Epidemiol.* 1975; 102:267–271.

Freeman, J., and Hutchison, G. B. Prevalence, incidence and duration. *Am. J. Epidemiol.* 1980; 112:707–723.

Miettinen, O. S. Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* 1976; 103:226–235.

Morrison, A. S. Sequential pathogenic components of rates. *Am. J. Epidemiol.* 1979; 109:709–718.

4. MEASURES OF EFFECT

Epidemiologists use the term *effect* in two senses. In a general sense, any instance of disease may be the effect of a given cause. In a more particular and quantitative sense, an effect is the difference in disease occurrence between two groups of people who differ with respect to a causal characteristic; the characteristic is generally referred to as an *exposure.*

Absolute effects are differences in incidence rate, cumulative incidence, or prevalence. *Relative effects* involve ratios of these measures. An *attributable proportion* is the proportion of a diseased population for which the exposure is one of the component causes in the sufficient cause that caused the disease.

ABSOLUTE EFFECT

Suppose that all sufficient causes of a particular disease were divided into two sets, those that contain a specific cause and those that do not. We can summarize this situation with the following diagram (Fig. 4-1).

U and U′ represent different collections of causal factors. Note that disease can occur either with or without E, the exposure of interest. The absolute effect of exposure E corresponds biologically to the existence of sufficient causes that require E as a component. Epidemiologically, the effect of E can be assessed by measuring the incidence rate of sufficient causes that contain E. People who have the exposure can nevertheless develop the disease from a mechanism that does not include the exposure, so that it does not suffice to measure the incidence rate of disease among those exposed. The incidence rate among the exposed reflects the incidence of both sets of sufficient causes represented in the diagram. The incidence rate of sufficient causes containing E must be derived by subtraction of the incidence rate of the sufficient causes that lack E. This rate can be measured in a population that resembles the exposed population but lacks the exposure. Thus, if I_1 is the incidence rate of disease in an exposed population and I_0 is the rate in a comparable unexposed population, $I_1 - I_0$ represents the incidence rate of disease with the exposure as a component cause. The absolute effect is the difference in incidence rates between an exposed and an unexposed population.

This measure is also often referred to straightforwardly as *rate difference.* Synonyms include attributable risk [Walter, 1976], which derives from the closely related measure, *risk difference,* sometimes also used as a synonym for rate difference. Properly, however, risk difference should denote only a difference in risks or cumulative incidences rather than incidence rates. Thus, while rate difference has a range from minus infinity to plus infinity and the same dimensionality as the rate involved (time $^{-1}$ if incidence), risk difference has a range from -1 to $+1$ and is dimensionless.

The term *attributable risk* is unwarranted if no cause-effect relation exists between exposure and disease. If the