3

Measures of Disease Frequency

Kenneth J. Rothman and Sander Greenland

In this chapter, we begin to address the basic elements, concepts, and tools of epidemiology. A good starting point is to define epidemiology. Unfortunately, there seem to be more definitions of epidemiology than there are epidemiologists. Some have defined it in terms of its methods. While the methods of epidemiology may be distinctive, it is more typical to define a branch of science in terms of its subject matter rather than its tools. A widely cited definition was given by MacMahon and Pugh (1970): the study of the distribution and determinants of disease frequency in man. A similar subject-matter definition has been attributed to Gaylord Anderson (Cole, 1979), who defined epidemiology simply as the study of the occurrence of illness. (Although reasonable distinctions can be made between the terms *disease* and *illness,* we shall consider them equivalent here.) Other sciences, such as pathology, are also directed toward the study of disease, but in epidemiology the focus is on the occurrence of disease rather than on the natural history or some other aspect of the disease. If the subject of epidemiologic inquiry is taken to be the occurrence of disease and other health outcomes, it is reasonable to infer that the ultimate goal of most epidemiologic research is the elaboration of causes that can explain patterns of disease occurrence.

Most epidemiologic research is designed to evaluate scientific hypotheses. These hypotheses are often posed as qualitative propositions; the "null" form of such propositions are specific statements, such as "Eating small amounts of aluminum, compared with eating no aluminum, does not increase the rate of occurrence of Alzheimer's disease." ("Null" here implies that there is no relation between the postulated cause and effect, as in "null hypothesis.") Stated in the null form, these specific propositions are, in principle, highly refutable, as discussed in the previous chapter.

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

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

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

INCIDENCE TIME

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

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

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

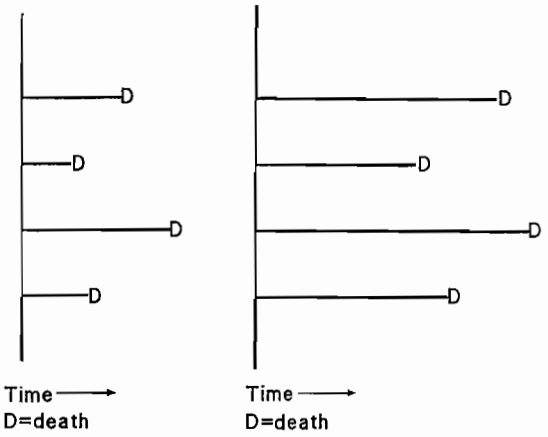

FIG. 3-1. Two different patterns of disease occurence.

INCIDENCE RATE

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

Person-Time

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

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

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

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

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

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

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

Closed and Open Populations

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

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

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

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

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

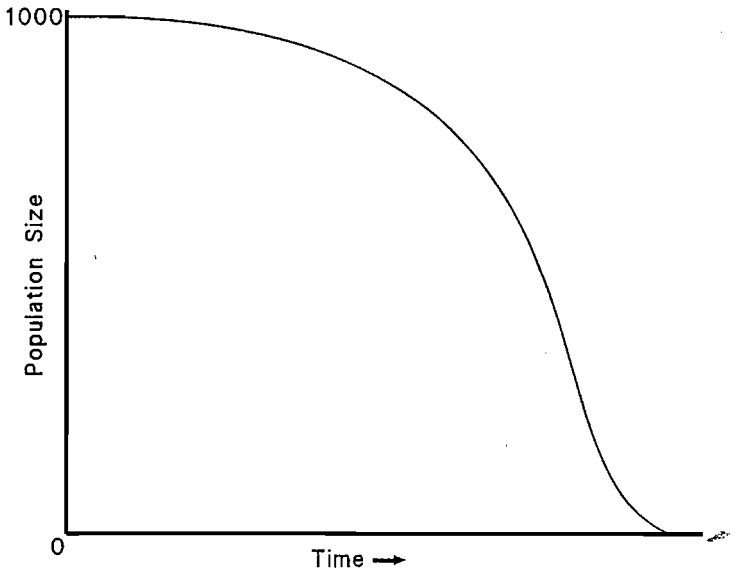

FIG. 3-2. Size of a closed population of 1000 people, by time.

tial. *Life-table* methodology is a procedure by which the death rate (or disease rate) of a closed population is evaluated within successive small age or time intervals, so that the age or time dependence of mortality can be elucidated. Even with such methods, it is important to distinguish any age-related effects from those related to other time axes, since each individual's age increases directly with an increase along any other time axis. For example, a person's age increases with increasing duration of employment, increasing calendar time, and increasing time from start of follow-up.

An open population differs from a closed population in that individual contributions need not begin at the same time. Instead, the population at risk is open to new members who become eligible with passing time. People can enter a population open in calendar time through various mechanisms. Some are born into it; others migrate into it. For a population of people of a specific age, individuals can become eligible to enter the population by aging into it. Similarly, individuals can exit from the person-time observational experience defining a given incidence rate by dying, aging out of a defined age group, emigrating, or becoming diseased (the latter method of exiting applies only if first bouts of a disease are being studied).

Steady State

If the number of people entering a population is balanced by the number exiting the population in any period of time within levels of age, sex, and other determinants of risk, the population is said to be *stationary,* or in a *steady state.* Steady state is a property that can occur only in open populations, not closed populations. It is, however, possible to have a population in steady state in which no immigration or emigration is occurring; this situation would require that births perfectly balance deaths in the population. The graph of the size of an open population in steady state is simply a horizontal line. People are continually entering and leaving the population in a way that might be diagrammed as shown in Fig. 3-3.

In the diagram, the symbol > represents a person entering the population, a line segment represents their person-time experience, and the termination of a line segment represents the end of their experience. A terminal D indicates that the experience ended because of disease onset, and a terminal C indicates that the experience ended for other reasons. In theory, any time interval will provide a good estimate of the incidence rate in a stationary population. The value of incidence will be the ratio of the number of cases of disease onset, indicated by D, to the area depicting the product of population × time.

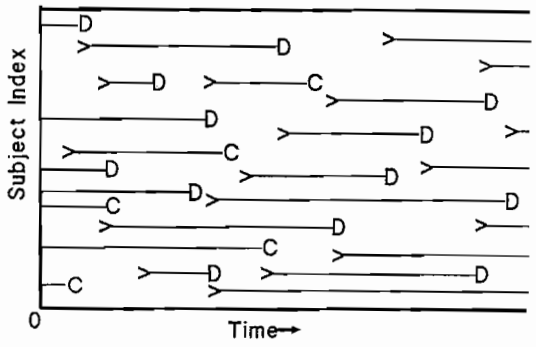

FIG. 3-3. Composition of an open population in approximate steady state, by time. > indicates entry into the population, D indicates disease onset, and C indicates exit from the population without disease.

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

Interpretation of an Incidence Rate

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

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

It might also be expressed as

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

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

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

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

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

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

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

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

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

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

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

OTHER TYPES OF RATES

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

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

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

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

miles, whereby the safety of commercial train and air travel can be compared. Here, person-miles replace person-time in the denominator of the rate. Like rates with time in the denominator, the numerical portion of such rates is completely dependent on choice of measurement units; a rate of 1.6 deaths per 10^6 passenger-mile equals a rate of 1 death per 10^6 passenger-kilometer.

The concept central to precise usage of the term *incidence rate* is that of expressing the change in incident number relative to the change in another quantity. Thus, a person-time rate expresses the increase in the incident number we expect per unit increase in person-time; an absolute rate expresses the increase in incident number we expect per unit increase in clock time; and a passenger-mile rate expresses the increase in incident number we expect per unit increase in passenger miles.

INCIDENCE PROPORTIONS AND SURVIVAL PROPORTIONS

When considering a given interval of time, we can also express the increase in incident number per unit increase in population size. If we measure size at the start of the interval and no one enters the population (immigrates) or leaves alive (emigrates) after the start of the interval, such a rate becomes the proportion of people who become cases among those who entered the interval. We call this quantity the *incidence proportion,* which may also be defined as the proportion of a closed population at risk that becomes diseased within a given period of time. This quantity is often called the *cumulative incidence* (Miettinen, 1976a), but this term is also used for another quantity we will discuss below. A more traditional term for incidence proportion is *attack rate,* but we reserve the term *rate* for person-time incidence rates.

If *risk* is defined as the probability of disease developing in an individual in a specified time interval, then incidence proportion is a measure, or estimate, of average risk. Although this concept of risk applies only to individuals and incidence proportion to populations, incidence proportion is sometimes called risk. "Average risk" is a more accurate synonym, one that we will sometimes use.

Like any proportion, the value of an incidence proportion ranges from zero to one and is dimensionless. It is uninterpretable, however, without specification of the time period to which it applies. An incidence proportion of death of 3% means something very different when it refers to a 40-year period than when it refers to a 40-day period.

A useful complementary measure to the incidence proportion is the *survival proportion,* which may be defined as the proportion of a closed population at risk that does *not* become diseased within a given period of time. If R and S denote the incidence and survival proportions, we have that $S = 1 - R$ and $R = 1 - S$. Another measure that is commonly used is the *incidence odds,* defined as $R/S = R/(1 - R)$, the ratio of the proportion getting the disease to the proportion not getting the disease. If R is small, $S \doteq 1$ and $R/S \doteq R$; that is, the incidence odds will approximate the incidence proportion when both quantities are small. Otherwise, because $S < 1$, the incidence odds will be greater than the incidence proportion.

Under certain conditions, there is a very simple relation between the incidence proportion and the incidence rate of a nonrecurrent event. Consider a closed population over an interval t_0 to t_1, and let $\Delta t = t_1 - t_0$ be the length of the interval. If N is the size of the population at t_0 and A is the number of disease onsets over the interval, then the incidence and survival proportions over the interval are $R = A/N$ and $S = (N - A)/N$. Now suppose the size of the population at risk declines only slightly over the interval. Then,

$N - A \doteq N$, $S \doteq 1$, and so $R/S \doteq R$. Furthermore, the average size of the population at risk will be approximately N, and so the total person-time at risk over the interval will be approximately $N\Delta t$. Thus, the incidence rate (I) over the interval will be approximately $A/N\Delta t$, and we obtain

$$R = A/N = (A/N\Delta t)\Delta t \doteq I\Delta t \text{ and } R \doteq R/S.$$

In words, the incidence proportion, incidence odds, and the quantity $I\Delta t$ will all approximate one another if the population at risk declines only slightly over the interval. We can make this approximation hold to within an accuracy of $1/N$ by making Δt so short that no more than one person leaves the population at risk over the interval. Thus, given a sufficiently short time interval, one can simply multiply the incidence rate by the time period to approximate the incidence proportion. This approximation offers another interpretation for the incidence rate: It can be viewed as the limiting value of the ratio of the average risk to the time period for the risk as the duration of the time period approaches zero.

A specific type of incidence proportion is the *case fatality rate,* or *case fatality ratio,* which is the incidence proportion of death among those in whom an illness develops (it is therefore not a rate in our sense, but a proportion). The time period for measuring the case fatality rate is often unstated, but it is always better to specify it.

PRODUCT-LIMIT AND EXPONENTIAL FORMULAS

We will now derive a more general relation between the incidence proportion of an inevitable, nonrecurrent event (such as death) and the incidence rate in a closed population. Consider the small closed population shown in Fig. 3-4. The time at risk (risk history) of each member is graphed in order from the shortest on top to the longest at the bottom. Each history ends with a D, indicating the occurrence of the event of interest, or at the end of follow-up at $t_5 = 1999$. The starting time is denoted t_0 and is here equal to 1980. Each time that one or more events occur is marked by a vertical dashed line, the unique event times are denoted by t_1 (the earliest) to t_4, and the end of follow-up is denoted by t_5. Let us denote the number of events at time t_k by A_k, the total number of persons at risk

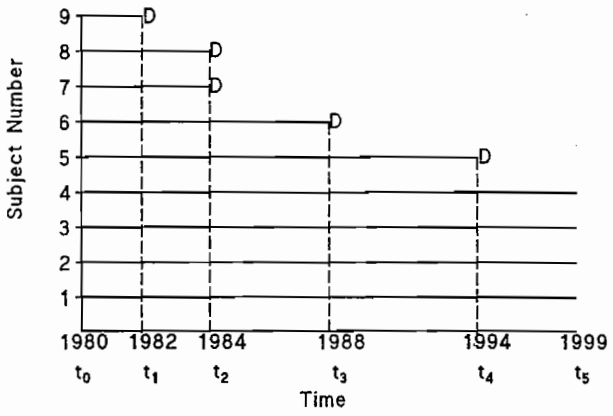

FIG. 3-4. Example of a small closed population with end of follow-up at 19 years.

TABLE 3-1. *Event times and intervals for the closed population in Fig. 3-4*

	Start	Outcome event times (t_k)				End
	1980	1982	1984	1988	1994	1999
Index (k)	0	1	2	3	4	5
No. outcome events (A_k)	0	1	2	1	1	0
No. at risk (N_k)	9	9	8	6	5	4
Proportion surviving (s_k)		8/9	6/8	5/6	4/5	4/4
Length of interval (Δt_k)		2	2	4	6	5
Person-time ($N_k \Delta t_k$)		18	16	24	30	20
Incidence rate (l_k)		1/18	2/16	1/24	1/30	0/20

at time t_k (including the A_k people who experience the event) by N_k, and the number of people alive at the end of follow-up by N_6.

Product-Limit Formula

Table 3-1 shows the history of the population over the 19-year follow-up period in Fig. 3-4, in terms of t_k, A_k, and N_k. Note that because the population is closed and the event is inevitable, the number remaining at risk after t_k, N_{k+1}, is equal to $N_k - A_k$, which is the number at risk up to t_k minus the number experiencing the event at t_k. The proportion of the population remaining at risk up to t_k that also remains at risk after t_k is thus:

$$s_k = \frac{N_k - A_k}{N_k} = \frac{N_{k+1}}{N_k} .$$

We can now see that the proportion of the original population that remains at risk at the end of follow-up is

$$S = N_5/N_1 = (N_5/N_4)(N_4/N_3)(N_3/N_2)(N_2/N_1) = s_4 s_3 s_2 s_1 ,$$

which for Table 3-1 yields

$$S = (4/5)(5/6)(6/8)(8/9) = 4/9.$$

This multiplication formula says that the survival proportion over the whole time interval in Fig. 3-4 is just the product of the survival proportions for every subinterval t_{k-1} to t_k. In its more general form,

$$S = \prod_{k=1}^{5} \frac{(N_k - A_k)}{N_k} . \qquad [3\text{-}1]$$

This multiplication formula is called the Kaplan-Meier or product-limit formula (Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984).

Exponential Formula

Now let T_k be the total person-time at risk in the population over the subinterval from t_{k-1} to t_k, and let $\Delta t_k = t_k - t_{k-1}$ be the length of the subinterval. Because the population is of constant size N_k over this subinterval and everyone still present contributes Δt_k person-

time units at risk, the total person-time at risk in the interval is $N_k\Delta t_k$, so that the incidence rate in the time following t_{k-1} up through (but not beyond) t_k is

$$I_k = \frac{A_k}{N_k\Delta t_k} .$$

But the incidence proportion over the same subinterval is equal to $I_k\Delta t_k$, so that the survival proportion over the subinterval is

$$s_k = 1 - I_k\Delta t_k .$$

Thus, we can substitute $1 - I_k\Delta t_k$ for s_k in the earlier equation for S, the overall survival proportion, to get

$$\begin{aligned} S &= (1 - I_5\Delta t_5)\,(1 - I_4\Delta t_4)\,(1 - I_3\Delta t_3)\,(1 - I_2\Delta t_2)\,(1 - I_1\Delta t_1) \\ &= (1 - (0)5)\,(1 - (1/30)6)\,(1 - (1/24)4)\,(1 - (2/16)2)\,(1 - (1/18)2) \\ &= 4/9, \end{aligned}$$

as before.

If each of the subinterval incidence proportions $I_k\Delta t_k$ is small (<0.10 or so), we can simplify the last formula by using the fact that, for small x,

$$1 - x \doteq \exp(-x).$$

Taking $x = I_k\Delta t_k$ in this approximation formula, we get $1 - I_k\Delta t_k \doteq \exp(-I_k\Delta t_k)$, and so

$$\begin{aligned} S &\doteq \exp(-I_5\Delta t_5)\exp(-I_4\Delta t_4)\ \ldots\ \exp(-I_1\Delta t_1) \\ &= \exp(-I_5\Delta t_5 - I_4\Delta t_4 - \ldots - I_1\Delta t_1) \\ &= \exp(-\sum_{k=1}^{5} I_k\Delta t_k), \end{aligned}$$

which for Table 3-1 yields

$$\exp(-\,0(5) - (1/30)6 - (1/24)4 - (2/16)2 - (1/18)2) \;=\; 0.483,$$

not too far from the earlier value of $4/9 = 0.444$. Finally, we use the fact that the incidence proportion for the whole period is $1 - S$ to get

$$R = 1 - S \doteq 1 - \exp(-\sum_{k=1}^{5} I_k\Delta t_k) . \qquad [3\text{-}2]$$

The last formula is cited in many textbooks and is sometimes called the *exponential formula* for relating rates and incidence proportions. The sum in the exponent, $\Sigma_k I_k\Delta t_k$, is sometimes called the *cumulative incidence* (Breslow and Day, 1980) or *cumulative hazard* (Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984). Confusingly, the term *cumulative incidence* is more often used to denote the incidence proportion. The cumulative hazard, although unitless, is *not* a proportion and will exceed 1.0 when the incidence proportion exceeds $1 - e^{-1} = 0.632$.

We wish to emphasize the assumptions we used to derive the exponential formula (equation 3-2):

1. The population is closed;
2. The event under study is inevitable (there is no competing risk); and
3. The number of events A_k at each event time t_k is a small proportion of the number at risk N_k at that time (i.e., A_k/N_k is always small).

If the population is not very small, we can almost always force assumption 3 to hold by measuring time so finely that every event occurs at its own unique time (so that only one event occurs at each t_k). In Table 3.1, the discrepancy between the true R of $5/9 = 0.555$ and the exponential formula value of $1 - 0.483 = 0.517$ is rather small considering that A_k/N_k gets as large as $2/8 = 0.25$ (at t_2).

Assumptions 1 and 2 were also used to derive the product-limit formula (equation 3-1). These assumptions are often not satisfied, yet are often overlooked in presentations and applications of the formulas. Some form of the closed-population assumption (No. 1) is essential because the incidence proportion is defined only in reference to closed populations. A major use of the product-limit and exponential formulas is, however, in translating incidence-rate estimates from open populations into incidence-proportion estimates for a closed population of interest. By assuming that the incidence rates in the two populations are the same at each time, one can justify substituting the survival proportions $(N_k - A_k)/N_k$ or the incidence rates observed in the open population into the product-limit formula or the exponential formula. This assumption is often plausible when the open population one observes is a subset of the closed population of interest, as in a cohort study with losses to follow-up that are unrelated to risk.

Applications with Competing Risks

Application of the product-limit and exponential formulas to outcomes with competing risks requires new concepts and assumptions. Consider the subinterval-specific incidence rates for our closed population of interest. When competing risks are present, the product-limit formula (equation 3-1) for the survival proportion S no longer holds, because N_{k+1} no longer need equal $N_k - A_k$: Competing risks may remove additional people between disease onset times, in which case N_{k+1} will be smaller than $N_k - A_k$. Also, when competing risks occur between t_{k-1} and t_k, the population size will not be constant over the subinterval; consequently, the person-time in interval k will not equal $N_k \Delta t_k$ and $I_k \Delta t_k$ will not equal A_k/N_k. Thus, the exponential formula (equation 3-2) given above will fail to hold if competing risks occur.

We can, however, ask the following questions: What would the incidence proportion over the total interval have been if no competing risk had occurred? The product-limit formula (equation 3-1) gives an estimate of this quantity under the assumption that the subinterval-specific incidence rates would not change if no competing risk occurred, and the exponential formula (equation 3-2) remains an approximation to it if A_k/N_k is always small. The assumption that the rates would not change if competing risks were removed requires careful scrutiny, however. Under conditions that would eliminate competing risks, the incidence rates of the outcome under study may be likely to change. Suppose for example the outcome of interest is colon cancer. Competing risks would include deaths from any causes other than colon cancer. Removal of so many risks would be virtually impossible, but an attempt to minimize them might involve such interventions as a high antioxidant, low-fat diet to prevent deaths from other cancers and from heart disease. Such preventive interventions would probably lower colon cancer rates, as well as the rates of the competing risks, and so violate the assumption that the specific rates would not change if no competing risk occurred.

For a general formula relating incidence proportions and rates in the presence of competing risks, see Benichou and Gail (1990a). For further discussion and debate of the con-

cept of removal of competing risks, see Kalbfleisch and Prentice (1980), Cox and Oakes (1984), Slud et al. (1988), Prentice and Kalbfleisch (1988), and Pepe and Mori (1993).

Relation of Survival Proportion to Average Incidence Time

Returning now to the simpler situation of an inevitable nonrecurrent outcome in a closed population, we will derive an equation relating survival proportions to average incidence time. First, we may write the total person-time at risk over the total interval in Fig. 3-4 as

$$N_1\Delta t_1 + \ldots + N_5\Delta t_5 = \sum_{k=1}^{5} N_k\Delta t_k$$
$$= 18 + 16 + 24 + 30 + 20 = 108 \text{ person-years.}$$

Thus, the average time at risk contributed by population members over the interval is

$$\frac{1}{N_0}\sum_{k=1}^{5} N_k\Delta t_k = \sum_{k=1}^{5} (N_k/N_0)\Delta t_k$$
$$= \frac{1}{9}\,108 = 12 \text{ years.}$$

Note that N_k/N_0 is just the proportion who remain at risk up to t_k, that is, the survival proportion from t_0 to t_k. If we denote this proportion N_k/N_0 by $S_{0,k}$ (to distinguish it from the subinterval-specific proportions s_k), the average time at risk can be written

$$\sum_{k=1}^{5} S_{0,k}\Delta t_k = (9/9)2 + (8/9)2 + (6/9)4 + (5/9)6 + (4/9)5 = 12 \text{ years,}$$

as before. Now suppose the interval is extended forward in time until the entire population has experienced the outcome of interest, as in Fig. 3-2. The average time at risk will then equal the average incidence time, so that the average incidence time will be computable from the survival proportions using the last formula. The survival proportions may in turn be computed from the subinterval-specific incidence rates, as described above.

Summary

The three broad types of occurrence measures—incidence time, incidence rate, and incidence (and survival) proportion—are all linked by simple mathematical formulas that apply when one considers an inevitable nonrecurrent event in a closed population followed until everyone has experienced the event. The mathematical relations become more subtle when one considers events with competing risks, open populations, or truncated risk periods, and interpretations become especially problematic when competing risks are present.

PREVALENCE

Unlike incidence measures, which focus on events, *prevalence* focuses on disease status. Prevalence may be defined as the proportion of a population that has disease at a specific point in time. The terms *point prevalence, prevalence proportion,* and *prevalence rate* are sometimes used to mean the same thing. The *prevalence pool* is the subset of the

population with the disease. An individual who dies with or from disease is removed from the prevalence pool; consequently, death from an illness decreases prevalence. Diseases with large incidence rates may have low prevalences if they are rapidly fatal. People may also exit the prevalence pool by recovering from disease or emigrating from the population.

Recall that a stationary population has an equal number of people entering and exiting during any unit of time. Suppose that both the population at risk and the prevalence pool are stationary and that everyone is either at risk or has the disease. Then the number of people entering the prevalence pool in any time period will be balanced by the number exiting from it:

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

People can enter the prevalence pool from the nondiseased population and by immigration from another population. Suppose there is no immigration into or emigration from the prevalence pool, so that no one enters or leaves the pool except by disease onset, death, or recovery. If the size of the population is N and the size of the prevalence pool is P, then the size of the population at risk that "feeds" the prevalence pool will be $N - P$. Also, during any time interval of length Δt, the number of people who enter the prevalence pool will be

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

where I is the incidence rate, and the outflow from the prevalence pool will be

$$I'P\Delta t,$$

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.

Prevalence, Incidence, and Mean Duration

Earlier we mentioned that in the absence of migration the reciprocal of an incidence rate in a stationary population equals the mean time spent in the population before the incident event. Therefore, in the absence of migration, the reciprocal of I' will equal the mean duration of the disease, $\overline{D}$, which is the mean time until death or recovery. It follows that

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

which yields

$$\frac{P}{N - P} = I \cdot \overline{D}.$$

$P/(N - P)$ is the ratio of diseased to nondiseased people in the population or, equivalently, the ratio of the prevalence proportion to its complement (1 – prevalence proportion). (We could call those who are nondiseased healthy except that we mean they do not have a specific illness, which doesn't imply an absence of all illness.) The ratio $P/(N - P)$ is called the *prevalence odds;* it is the odds of having a disease relative to not having the disease. As shown above, the prevalence odds equals the incidence rate times the mean duration of illness. If the prevalence is small, say <0.1, then

$$\text{Prevalence proportion} \doteq I \cdot \overline{D},$$

since the prevalence proportion will approximate the prevalence odds for small values of prevalence. More generally (Freeman and Hutchison, 1980), under the assumption of stationarity and no migration in or out of the prevalence pool,

$$\text{Prevalence proportion} = \frac{I \cdot \overline{D}}{1 + I \cdot \overline{D}},$$

which can be obtained from the above expression for the prevalence odds, $P/(N - P)$.

Like the incidence proportion, the prevalence proportion is dimensionless, with a range of 0 to 1. 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, which has the dimensionality of time, giving a dimensionless product. Furthermore, the product $I \cdot \overline{D}$ has the range of 0 to infinity, which corresponds to the range of prevalence odds, whereas the expression

$$\frac{I \cdot \overline{D}}{1 + I \cdot \overline{D}}$$

is always in the range 0 to 1, corresponding to the range of a proportion.

Unfortunately, the above formulas have limited practical utility because of the no-migration assumption and because they do not apply to age-specific prevalence (Miettinen, 1976a). If we consider the prevalence pool of, say, diabetics ages 60–64, we can see that this pool experiences considerable immigration from younger diabetics aging into the pool, and considerable emigration from members aging out of the pool. Under such conditions, we require more elaborate formulas that give prevalence as a function of age-specific incidence, duration, and other population parameters (Preston, 1987; Keiding, 1991; Alho, 1992).

Utility of Prevalence in Etiologic Research

Seldom is prevalence of direct interest in etiologic applications of epidemiologic research. Since prevalence reflects both the incidence rate and the probability of surviving with disease, studies of prevalence or studies based on prevalent cases yield associations that reflect the determinants of survival with disease just as much as the causes of disease. The study of prevalence can be misleading in the paradoxical situation in which better survival from a disease and therefore a higher prevalence follow from the action of preventive agents that mitigate the disease once it occurs. In such a situation, the preventive agent may be positively associated with the prevalence of disease and so be misconstrued as a causative agent.

Nevertheless, for one class of diseases, namely, congenital malformations, prevalence is usually employed. The proportion of babies born with some malformation is a prevalence proportion, 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. The measure classifies the population of newborns as to their disease status, malformed or not, at the time of birth. This example illustrates that the time reference for a prevalence need not be a common point in calendar time; it can be a point on another time scale, such as an individual's life span.

It would be more useful and desirable to study the incidence than the prevalence of congenital malformations; as already noted, studying prevalence makes it impossible to distinguish the effects of agents that increase the incidence rate from the effects of agents that increase survival with the disease once the disease occurs. Unfortunately, it is seldom possible to measure the incidence rate of malformations, since the population at risk, young embryos, is difficult to ascertain, and learning the occurrence and timing of the malformations among the embryos is equally problematic. Consequently, in this area of research, incident cases are not usually studied, and most investigators settle for the theoretically less desirable but much more practical study of prevalence at birth.

Prevalence is sometimes used to measure the occurrence of nonlethal degenerative diseases with no clear moment of onset. It is also used in seroprevalence studies of the incidence of infection, especially when the infection has a long asymptomatic (silent) phase that can only be detected by serum testing. The human immunodeficiency virus infection is a prime example. In these 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 relevant measure than incidence.

STANDARDIZATION

Suppose we are given a distribution of person-time specific to a series of variables, for example, the person-years at risk experienced within age categories 50–59, 60–69, and 70–74, for males and females in Quebec in 1990. Let $T_1, T_2, \ldots, T_6$ be the person-years in the six age-sex categories in this example. Suppose also that we are given a schedule of six age-sex specific incidence rates $I_1, I_2, \ldots, I_6$ corresponding to the age-sex specific strata; these rates may come from the same population or a different population, or may be purely hypothetical. From this distribution and rate schedule, we can compute a weighted average of the rates with weights from the distribution,

$$I_s = \frac{I_1T_1 + \ldots + I_6T_6}{T_1 + \ldots + T_6} = \frac{\sum_{k=1}^{6} I_kT_k}{\sum_{k=1}^{6} T_k}.$$

The numerator of I_s may be recognized as the number of cases one would see in a population that had the person-time distribution $T_1, T_2, \ldots T_6$ and these stratum-specific rates. The denominator of I_s is the total person-time in such a population. Therefore, I_s is the rate one would see in a population with distribution $T_1, T_2, \ldots T_6$ and specific rates $I_1, I_2, \ldots I_6$.

I_s is traditionally called a *standardized rate,* and $T_1, T_2, \ldots T_6$ is called the *standard distribution* on which I_s is based. I_s represents the overall rate that will be observed in a population whose person-time follows the standard distribution and whose specific rates are $I_1, I_2, \ldots I_6$.

The standardization process can also be conducted with incidence or prevalence proportions. Suppose for example we have a distribution $N_1, N_2, \ldots N_6$ of persons rather than person-time at risk and a corresponding set of stratum-specific incidence proportions $R_1, R_2, \ldots, R_6$. From this distribution and set of proportions, we can compute the weighted average risk

$$R_s = \frac{R_1N_1 + \ldots + R_6N_6}{N_1 + \ldots + N_6} = \frac{\sum_{k=1}^{6} R_kN_k}{\sum_{k=1}^{6} N_k},$$

which is a *standardized risk* based on the standard distribution $N_1, N_2, \ldots, N_6$.

Because the rates that apply to a population can affect the person-time distribution, the standardized rate is not necessarily the rate that would describe what would happen to a population with the standard distribution $T_1, \ldots, T_6$ if the specific rates $I_1, I_2, \ldots, I_6$ were applied to it. This problem will be discussed further in the next chapter. The problem does not arise when considering standardized risks because the initial distribution $N_1, \ldots, N_6$ cannot be affected by the subsequent risks $R_1, \ldots, R_6$.

Standardized rates and proportions will recur at several points in this book, where they will be described in more detail, especially in the chapters on basic statistical techniques (Chapters 14 and 15).