Chs. 3/4

fairly narrow range. Rather, the point in time is the *time of observation* for each child. The fact that each child's disease status was observed only as of one point in time, not monitored over a period of time, is the key feature. The calendar time period and age range over which the examinations were done are relevant as descriptors, along with place and other population-defining characteristics, to put the prevalence estimate in its proper context. Prevalence can be compared across time periods or age groups, just as other disease frequency measures can.

In some studies, the observation times on individuals may indeed be synchronized in calendar time, by age, or on another time scale. Then prevalence also pertains to a specific time point on that time scale.

Example: A study of Crohn's disease and ulcerative colitis in Manitoba, Canada, used administrative data from the province's only health insurer to identify persons who had received care for one of these chronic gastrointestinal diseases (Bernstein et al., 1999). Prevalent cases as of December 31, 1994, were those who had made the requisite number of medical visits for the condition during the preceding two years and who had not died or emigrated from Manitoba before the end of 1994. The resulting prevalence estimate pertained to what can be considered a point in calendar time—December 31, 1994.

Example: The prevalence of HIV infection among inmates entering U.S. correctional facilities was estimated from HIV-1 antibody tests on routine blood samples obtained upon entry to jail or prison (Vlahov et al., 1991). These intake examinations occurred on various calendar dates, but they pertained to the same point on the time scale that chronicled each inmate's incarceration.

INCIDENCE

Incidence measures how frequently susceptible individuals become disease cases as they are observed over time. It is based on disease events, each of which represents a transition from being at risk to being diseased.

Counts

An *incident case* occurs when an individual changes from being susceptible to being diseased, by the study's case definition. The *count of incident cases* is the number of such events that occur in a defined population during a specified time period. Recurrences of disease in the same person may or may not qualify as

incident cases, depending on the study's operational definition of disease, as discussed in Chapter 2.

Simple counts of incident cases can sometimes be sufficient to guide health planning. For example, knowing the number of lower-extremity amputations per year in a certain health plan could be used to project the number of limb prostheses likely to be needed.

Counts may also be adequate for comparing incidence across populations that can safely be assumed to be of similar size.

Example: Phillips and colleagues (1999) found that, over a 16-year period, 113.8 deaths due to substance abuse occurred in the U.S. during the first week of the month for every 100 such deaths in the last week of the month. They hypothesized that the excess may be related to receipt of government benefit payments from Social Security, welfare, or military benefits at the beginning of each month. Because the population at risk would be nearly the same size across different weeks of the month, the study could be based simply on the number of deaths in each one-week period.

Example: In 2000, 702,093 new cases of genital *Chlamydia trachomatis* infection were reported to the U.S. Centers for Disease Control and Prevention (CDC), compared with 358,995 new cases of gonorrhea (Centers for Disease Control and Prevention, 2001b). Assuming similar completeness of reporting for both diseases, these counts by themselves should accurately reflect differences in incidence between these two sexually transmitted diseases. This is because the sizes of the populations at risk for each disease should be about the same (or nearly so, after subtracting prevalent cases).

Cumulative Incidence

Cumulative incidence is the proportion of initially susceptible individuals in a closed population who become incident cases during a specified time period:

$$\text{Cumulative incidence} = \frac{\text{Number of incident cases}}{\text{Number of persons initially at risk}}$$

Cumulative incidence is also sometimes called the *incidence proportion* or *attack rate*. It is the simplest measure of incidence to account explicitly for the size of the population at risk.

Koepsell & Weiss: Epidemiologic Methods 2003.

Example: A jumbo jet full of tourists bound from Tokyo to Copenhagen stopped at Anchorage, Alaska, for refueling and reprovisioning. Upon reaching cruising altitude again, passengers were served breakfast. Somewhere over the polar ice cap, an illness characterized by cramps, vomiting, and diarrhea swept through the plane, and by the time they reached Copenhagen, 196/344 = 57% of passengers had become ill. Epidemiologists who investigated the outbreak used interview data and food service records to calculate the cumulative incidence of illness among those who did and those who did not eat various food items. Eating ham proved to be strongly associated with becoming ill. Among those who ate ham that had been prepared by a particular cook, 86% got sick, compared with none of those who ate ham prepared by a different cook. Microbiological tests found heavy staphylococcal contamination of the suspected ham, which was eventually found to have resulted from improper food handling (Eisenberg et al., 1975).

The time period cumulative incidence refers to is usually fixed, specified, and the same for all members of the study population. For example, the proportion of patients undergoing a surgical procedure who develop deep venous thrombosis during the two weeks after surgery could be termed the "two-week cumulative incidence" of that complication.

In some situations, the time period that cumulative incidence refers to may not be stated and may, in fact, vary among individuals. For example, the cumulative incidence of death before discharge among hospitalized patients is sometimes used as a measure of disease severity or outcome. Because of differences in length of hospital stay, however, the amount of time at risk for death varies among patients.

Cumulative incidence is easy to calculate and to interpret, but unfortunately it can only be measured directly in closed populations (as defined in Chapter 2). In particular, the population cannot gain or lose members during the period of follow-up, except for losses that occur after disease occurs. The reason is that cumulative incidence is designed to estimate the proportion of persons initially at risk who develop disease during follow-up. If gains or losses in the study population took place, the essential correspondence between the case count in the numerator, and the defined population at risk in the denominator, would be broken. For example, if a new member were to join the population partway through follow-up and then become a case, he or she would be added to the numerator, even though she or he had not been counted as a member of the denominator population at risk. If an original member of the denominator population were lost to follow-up, he or she might actually go on to become a case during the study period who would go undetected.

Chapter 4 describes how cumulative incidence can be estimated indirectly under certain assumptions, even when follow-up data on some original population members are incomplete. It also describes methods for obtaining confidence limits.

Incidence Rate

The *incidence rate* is the count of incident cases divided by the amount of at-risk experience from which they arose. Its denominator is usually measured in units of person-time.

$$\text{Incidence rate} = \frac{\text{Number of incident cases}}{\text{Amount of at-risk experience}}$$

Whether disease recurrences are counted in the numerator depends on the study's case definition, as discussed in Chapter 2.

The incidence rate also goes by several other names, including *incidence density* (a term originally suggested by Miettinen, 1976), *person-time incidence rate,* or sometimes simply *incidence.*

Example: Gardner et al. (1999) studied on-the-job back sprains and strains among 31,076 material handlers employed by a large retail merchandise chain. Payroll data for a 21-month period during 1994–1995 were linked with job injury claims, which provided data on the timing of each injury, body part injured, and mechanism of injury. A total of 767 qualifying back injuries occurred during 54,845,247 working hours, yielding an *incidence rate* of 1.40 back injuries per 100,000 worker-hours. Higher incidence was found among males and among employees whose work was more physically demanding.

The work force in this example comprised an open, defined population. Thousands of workers joined or left the company during the study period. Only on-the-job back injuries were of interest, so each worker's at-risk experience consisted of many discontinuous time periods at work, separated by periods away from work. These features of the research situation made an incidence-rate approach to measuring disease frequency attractive and a good match to the available data.

The basic rationale behind the incidence rate is straightforward. Other things being equal, the number of new cases of disease should be proportional to *(1)* the size of the population at risk and *(2)* the amount of time over which susceptible individuals are observed. The denominator simply combines these two elements.

The number of cases and the number of persons at risk are unitless counts, while the time component of the denominator has units, so an incidence rate has units of time^{-1}.

Incidence rates can be used across a wide range of epidemiologic research situations. They can be applied to both closed and open populations, with or without detailed information on the time at risk for each individual, and for both recurrent and non-recurrent disease events—circumstances in which cumulative incidence may be impossible to apply.

Estimating incidence rate with detailed data on individual times at risk

In many epidemiologic studies, detailed information is available on the amount of time at risk for each individual and the timing of each disease event. In the Gardner back-injury study, for example, payroll records furnished each worker's time on the job right down to the hour, and injury claims contained data on the timing of each back injury.

To see how a person-time denominator is calculated from detailed individual data, consider the small population shown in Figure 3–3. It deliberately involves several features that would make cumulative incidence impossible to apply but that can be accommodated easily under an incidence-rate approach. Four cases occur among six individuals during a 30-day period. Some people enter late in the study period, some are observed only intermittently, some drop out early, one (person no. 4) is not at risk for part of the time, and one (person no. 5) has two separate disease events.

Depending on the study purpose, recurrent disease events in the same person might or might not be relevant and qualify for inclusion. In this instance, that decision affects the contributions of several individuals to both the numerator and

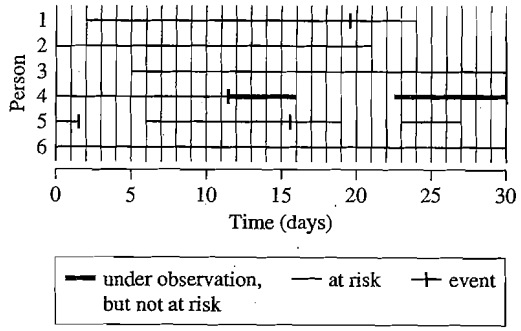

Figure 3–3. Hypothetical Population to Illustrate Incidence Rate Estimation with Detailed Data on Individual Times at Risk.

Table 3–1. Example of Incidence Rate Calculation, Keyed to Figure 3–3

	If All Cases Qualify		If Only First Cases Qualify	
PERSON	CONTRIBUTION TO NO. OF CASES	DAYS AT RISK	CONTRIBUTION TO NO. OF CASES	DAYS AT RISK
1	1	22	1	17.5
2	0	21	0	21
3	0	25	0	25
4	1	11.5	1	11.5
5	2	18.5	1	1.5
6	0	30	0	30
Total cases	4		3	
Total person-days		128		106.5
Incidence rate—per 100 person-days	3.13		2.82	

the denominator of the incidence rate estimate. Table 3–1 shows the calculations both ways.

- If recurrent events qualify, then both of the disease events in person no. 5 are added to the numerator. In addition, anyone who becomes a case may continue thereafter to contribute person-time at risk to the denominator, because he or she remains at risk for recurrence.
- If recurrent events do *not* qualify, then person no. 5 contributes only one event to the numerator. In addition, anyone who becomes a case contributes no further person-time to the denominator thereafter, because he or she is no longer at risk for a first event.

Estimating incidence rate without detailed data on individual times at risk

Often detailed information about each population member's time at risk is unknown and not feasibly obtainable. This problem often arises, for example, when the defined population of interest consists of residents of a geographic area over some time period. The number of incident cases may be readily available, but the challenge is to estimate the total amount of person-time at risk from which those cases arose.

Figure 3–4 provides a graphical example. It shows gradual growth in the size of a true population at risk over an observation period that extends from Time A to Time E. Total person-time at risk corresponds to the area of the shaded region, which could be calculated exactly if moment-by-moment details about the size

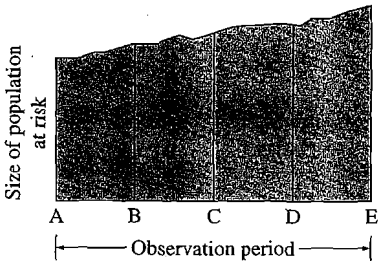

Figure 3–4. Estimating Average Size of the Population at Risk.

of the population at risk were known. Otherwise, the area must be estimated by sampling the size of the population at risk at one or more time points, averaging these population-size estimates, and multiplying by the duration of the observation period. Variations of this approach include:

1. Using estimated population size at mid-period (here, Time C) as an estimate of the average. This method might be suitable when a single population count is made at or near the middle of the observation period. This could apply, for example, to a county or state observed over the four-year period from 1998 to 2001, because 2000 was a census year.
2. Averaging the estimated population size at the start and at the end of the observation period (here, Times A and E). This method might be suitable if, for example, the observation period spans a 10-year period between two decennial censuses.
3. Averaging several population size estimates made periodically during the observation period. For example, government planning agencies in many areas publish year-by-year estimates of population size and composition for geopolitical areas. Average population size over a four-year observation period could then be estimated by averaging four annual estimates.

Example: Some 702,093 new cases of genital *Chlamydia trachomatis* infection were reported in the U.S. in 2000 (Centers for Disease Control and Prevention, 2001b). The U.S. Census Bureau estimates that the population of the U.S. on July 1, 2000 was about 282.1 million. Under method no. 1 above, 282,100,000 can be treated as an estimate of the average size of the population at risk during the one-year period from January 1, 2000, through December 31, 2000, yielding an estimated 282,100,000 *person-years at risk*. The estimated incidence rate of genital *Chlamydia trachomatis* infection would therefore be $702,093/282,100,000 = .00249 = 2.49$ cases per 1000 person-years at risk.

By similar logic, it is sometimes possible to calculate an incidence rate from a published paper even if the results are not reported as such. If a cohort of N persons is described as having been followed for an average of $\overline{T}$ years, then they experienced $N\overline{T}$ person-years in all. If all of this person-time can be considered at-risk time, and if c incident cases occur, then $c/N\overline{T}$ is an estimate of the incidence rate.

For some diseases, the prevalence of disease may be high enough, or a not-at-risk state common enough, that the discrepancy between total population size and size of the true population at risk is too large to be ignored. Corrections may then need to be based on the estimated prevalence of disease or the estimated proportion of the population that is not at risk. For example, the estimated incidence of dementia in the elderly has been found to increase considerably after prevalent cases of dementia were subtracted from the denominator (Rocca et al., 1998). For uterine cancer, higher and almost certainly more accurate incidence estimates have been obtained when the estimated number of women with a prior hysterectomy were subtracted from the denominator (Marrett, 1980).

Denominators other than person-time

In some areas of epidemiologic research, such as the study of injuries, metrics other than person-time are often used to quantify the amount of at-risk experience from which a set of incident cases arose. For example, the incidence of motor-vehicle collision injuries can be expressed as injuries per 100,000 person-years, as injuries per 100,000 licensed-driver-years, or as injuries per million vehicle-miles traveled. The extent to which older adults are a high-risk group for motor-vehicle collision injuries has been shown to depend strongly on which measure of incidence is used (Massie et al., 1995). Relative to younger adults, a smaller percentage of older adults have a valid driver's license, and even those who do have a driver's license drive fewer miles per year than younger drivers. Hence the increase in incidence by age is more marked when the denominator is vehicle-miles traveled.

Comparison of Cumulative Incidence and Incidence Rate

The distinction between cumulative incidence and incidence rate was appreciated by early epidemiologists and health statisticians (Vandenbroucke, 1985). The differences are both conceptual and statistical (Morgenstern et al., 1980; Elandt-Johnson, 1975). Table 3–2 summarizes and contrasts several properties of these two measures of incidence.

Despite the differences, the generic term *incidence* is widely applied to both cumulative incidence and incidence rate throughout the epidemiologic literature. The specific kind of incidence being discussed must often be inferred from the context. To accustom readers to this widespread practice, and for brevity,

Table 3–2. Comparison of Cumulative Incidence and Incidence Rate

CHARACTERISTIC	CUMULATIVE INCIDENCE	INCIDENCE RATE
Units	None	Time^{-1}
Range	0–1	0–infinity
Directly calculable by:	Observing a closed population over time	Observing a closed or open population over time with detailed data on individual times at risk
Indirectly calculable by:	Survival-analysis methods in presence of censoring[a]	Estimating total person-time as (average size of population at risk) × (duration of observation period)
Individual-level counterpart	Risk (probability)	Hazard rate[a]

[a]Discussed in Chapter 4.

this book often simply uses the generic term *incidence* when its meaning seems unambiguous.

Chapter 4 describes how confidence limits for incidence rates can be obtained; how cumulative incidence and the incidence rate are related mathematically and, under certain assumptions, computable from each other; and how incidence rates in a population relate to individual-level hazard rates.

Variants of Incidence

Incidence can actually be thought of as a family of disease-frequency measures. Some members of this family traditionally go by names of their own, but in reality they are just special types of incidence.

Mortality

Mortality is the incidence of fatal cases of a disease in the population at risk for dying of the disease. The denominator includes both prevalent cases of the disease as well as persons who are at risk for developing the disease. Subtypes are *cumulative mortality* and *mortality rate*. *Mortality density* and *death rate* are essentially synonyms for the mortality rate.

Example: Some 8,911 deaths due to AIDS were recorded in the U.S. in 2000 (Centers for Disease Control and Prevention, 2001a). Essentially the entire U.S. population is considered to be at non-zero risk for dying of AIDS, although the level of risk clearly varies greatly from person to person. Hence the denominator for the mortality rate is (estimated average size of the U.S. population during 2000) × (length of observation period) = 282,100,000 × 1 year. The mortality rate for AIDS in 2000 was thus 8,911/282,100,000 = 3.16 deaths per 100,000 person-years.

Fatality

Fatality refers to the incidence of death from a disease *among persons who develop the disease*. The difference between fatality and mortality is in their denominators. Fatality reflects the prognosis of the disease among cases, while mortality reflects the burden of deaths from the disease in the population as a whole.

In principle, *cumulative fatality* and *fatality rate* can be defined as special types of cumulative incidence and incidence rate, respectively, with appropriate restrictions on who counts toward the numerator and denominator. In practice, these terms are rarely used, although the underlying theory still applies.

Instead, *case fatality* is a commonly used measure of fatality. It is:

$$\text{Case fatality} = \frac{\text{Number of fatal cases}}{\text{Total number of cases}} \tag{3.1}$$

Case fatality can be viewed as the cumulative incidence of death due to the disease among those who develop it. As with *attack rate*, a fixed time period after disease onset may or may not be explicitly specified and must often be inferred from the context. As a variant of cumulative incidence, case fatality is most readily applied for diseases of relatively short duration, in which there are few losses to follow-up or deaths from other causes.

Example: The National Highway Traffic Safety Administration (2001) reported that 4,739 deaths occurred in the U.S. during 2000 when a pedestrian was struck and killed by a motor vehicle. They estimate that 78,000 pedestrians were injured in pedestrian/motor-vehicle collisions during that year. Based on these data, the case fatality of pedestrian/motor-vehicle collision injury in 2000 was 4,739/78,000 = 6.1%.

Proxy Measures of Incidence

Sometimes good denominator data for the desired measure of incidence cannot feasibly be obtained. Yet case counts alone are likely to be inadequate for

comparing incidence between populations that differ in size or other key characteristics. Under those circumstances, a proxy denominator may be better than none at all.

Proportional mortality

The *proportional mortality* for a disease is:

$$\text{Proportional mortality} = \frac{\text{Deaths from the disease}}{\text{Deaths from all causes}}$$

As its name indicates, it is simply the proportion of all deaths that are due to a particular cause for a specified population and time period of interest. This proportion can provide useful descriptive information in its own right: for example, the statement that heart disease accounted for 30% of all deaths among Americans in 1999 refers to proportional mortality (National Center for Health Statistics, 2001).

For comparing disease frequency between populations, the main advantage of proportional mortality is that its denominator—total number of deaths—can usually be ascertained from the same source that furnishes its numerator. The count of all deaths serves as a proxy for person-time at risk under the assumption that, other things being equal, one would expect total deaths to vary in proportion to population size and in proportion to the duration of the monitoring period.

A potential limitation of comparing proportional mortality between populations or subpopulations can be illustrated by an example:

Example: Berkel and de Waard (1983) studied mortality among Seventh-day Adventists (SDA) in the Netherlands over a ten-year period. The church proscribes its members from using tobacco or alcoholic beverages and recommends a vegetarian diet. These policies led the investigators to expect a reduced death rate among SDA from cancer (particularly lung cancer, which is strongly related to smoking) and heart disease.

The second column of Table 3–3 shows the observed number of deaths among SDA, and the third column shows the percentage of those deaths due to each cause. For comparison, the fourth column shows the percentage of deaths by cause in a similarly aged sample of the full population of the Netherlands during the same ten years. Based on a comparison of proportional mortality (columns 3 and 4), there seems to be no evidence of a reduced occurrence of death due to lung cancer and only a slight reduction in mortality due to cardiovascular disease.

But in this instance, the investigators also had detailed year-by-year data on the size of the SDA population, from which they could determine the number

Table 3–3. Proportional Mortality and Mortality Rate Analyses of Deaths among Dutch Seventh-Day Adventists (SDA)

CAUSE OF DEATH	OBSERVED DEATHS IN SDA	Proportional Mortality		Expected Deaths in SDA, Based on:	
		SDA	NETHERLANDS	NETHERLANDS PROPORTIONAL MORTALITY	NETHERLANDS MORTALITY RATES
Lung cancer	12	2.5%	2.5%	12	27
Other cancer	103	21.3%	18.9%	91	204
Cardiovascular	227	47.1%	50.8%	245	547
Other causes	130	27.0%	27.7%	134	299
All causes	482	100.0%	100.0%	482	1077

[*Source:* Based on Berkel and de Waard (1983).]

of person-years at risk contributed by SDA during the study period, by age and gender. They obtained the age- and sex-specific mortality rates for the Netherlands as a whole from published sources. By applying these published Dutch mortality rates to the SDA denominator data, they were able to estimate how many deaths would have been expected among the SDA if they had experienced the mortality rates in effect for all Dutch people of similar age and gender.

The rightmost column of Table 3–3 shows these results, and they lead to quite a different conclusion. The observed numbers of lung cancer and cardiovascular disease deaths in SDA were in fact sharply lower than the number of such deaths expected based on rates for all Dutch people of similar age and gender. But deaths from *other* causes were also substantially lower than expected among SDA. Hence the *proportions* of SDA deaths from lung cancer and heart disease differed very little from those in the Netherlands in general. In this example, we would have been led astray if only a proportional mortality analysis had been possible. The total number of deaths was actually a poor proxy for population size because of a major difference in all-causes mortality between populations.

Other proxies for incidence are based on the same basic idea, applied to nonfatal events. For example, hospital admissions for diabetes can be expressed as a proportion of all hospital admissions if no good data are available on the size of the true population at risk for hospitalization. Similarly, incident cases of colon cancer can be expressed as a proportion of all incident cancer cases. The same potential pitfall applies, however: comparisons could be misleading if the overall

hospitalization rate or the overall cancer incidence rate were to differ between populations being compared. Contrasts based on proxy measures must therefore be cautiously interpreted.

Fetal death ratio

In perinatal epidemiology, the frequency of fetal death in a certain population over a specified time period is quantified as:

$$\text{Fetal death ratio} = \frac{\text{Number of fetal deaths}}{\text{Number of live births}}$$

The denominator for a cumulative-incidence measure of fetal death would be the total number of pregnancies. But some pregnant women may undergo spontaneous or elective abortions that can be difficult to ascertain and count. Hence the number of live births is used as a proxy for the total number of pregnancies.

In contrast to proportional mortality, the fetal death ratio and other analogues that do not include the numerator as part of the denominator are not proportions.

OTHER MEASURES OF DISEASE FREQUENCY

Period Prevalence

Earlier, *prevalence* was described as reflecting the frequency of the diseased state at a specified point in time. Especially when *prevalence* refers to a point in calendar time, the term *point prevalence* is often used (Last, 2000). In contrast, *period prevalence* is a hybrid of prevalence and cumulative incidence. Like cumulative incidence, it refers to a period of time, rather than a point in time. Cases counted in its numerator, however, include both *(1)* cases that are extant when the observation period begins, and *(2)* new cases that occur during the period. Referring to Figure 3–5, persons no. 1, no. 3, no. 4, and no. 5 would all count as cases. The denominator includes both *(1)* extant cases when the period starts and *(2)* persons

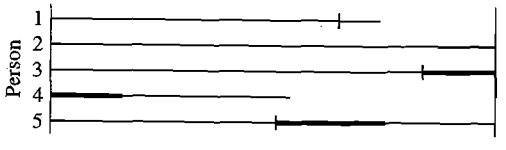

Figure 3–5. Illustration of Period Prevalence.

at risk when the period starts. For Figure 3–5, the period prevalence would thus be $4/5 = 0.8$.

Period prevalence is essentially uninterpretable except in a closed population, for the same reasons that apply to cumulative incidence. For a closed population, if P = point prevalence when the observation period starts, and CI = cumulative incidence among individuals at risk at that time, then period prevalence can be seen to be:

$$\text{Period prevalence} = P + (1 - P) \cdot CI$$

For Figure 3–5, this would be $1/5 + (1 - 1/5) \times 3/4 = 0.8$.

The main limitation of period prevalence is that point prevalence and cumulative incidence convey very different kinds of information about disease frequency. Those distinctions are lost when they are combined in this way, which limits the usefulness of period prevalence as a summary measure. When possible, point prevalence and cumulative incidence are generally better kept separate as two more interpretable components.

Yet sometimes this separation cannot be made from the data available. For example, the U.S. Centers for Disease Control (1998) reported that 25.3 per 1000 U.S. women who delivered a live-born infant during 1993–1995 had diabetes during the pregnancy, according to data on the baby's birth certificate. Some of these mothers had diabetes before becoming pregnant, while others developed diabetes during pregnancy. In any event, all reportedly had diabetes sometime during the period of pregnancy, so 25.3/1000 is probably best regarded as a period prevalence.

Years of Potential Life Lost

As noted earlier, case counts alone can be used to compare the frequency of two or more diseases within the same population. For example, the purpose may be to help guide allocation of resources among different programs aimed at specific diseases. Because of the special importance often attached to fatal cases, and because mortality data are often readily available, such comparisons are often based on the number of deaths from each disease.

Implicitly, these comparisons weight all deaths equally. It has been argued, however, that "premature" deaths—those occurring at younger ages—have greater social and economic impact than do deaths in old age, and that age at death should be considered when comparing diseases (Centers for Disease Control and Prevention, 1986). One measure designed to do this is *years of potential life lost (YPLL)* (Gardner and Sanborn, 1990). One version, used in reporting of national health

Table 3–4. Top Ten Causes of Death by Years of Potential Life Lost Before Age 75 Years and by Total Deaths: United States, 1998

	By Years of Potential Life Lost		By Number of Deaths	
RANK	DISEASE CATEGORY	$YPLL_{75}$ [a]	DISEASE CATEGORY	NUMBER OF DEATHS
1	Cancer	1716	Heart disease	724,915
2	Heart disease	1343	Cancer	549,787
3	Unintentional injuries	1052	Stroke	167,340
4	Suicide	365	COPD [b]	124,153
5	Homicide	301	Unintentional injuries	97,298
6	Stroke	233	Diabetes mellitus	68,379
7	COPD [b]	186	Pneumonia and influenza	63,686
8	HIV infection	177	Alzheimer's disease	44,507
9	Diabetes mellitus	174	Chronic renal disease	35,524
10	Chronic liver disease	159	Septicemia	30,670

[a] Years of potential life lost to age 75 years, per 100,000 persons age <75 years
[b] COPD = Chronic obstructive pulmonary disease
[*Source:* National Center for Health Statistics (2001).]

statistics for the U.S., is:

$$\text{YPLL} = \sum_{a=1}^{X} d_a(X - a)$$

where a denotes age at death (in years), d_a denotes number of deaths at age a, and X denotes a particular cutoff age, often 65 or 75 years. Essentially, YPLL weights each death by the number of years before age X at which the death occurs. Deaths in infancy get the most weight; deaths at or after age X years get zero weight. YPLL can also be expressed per 1000 population (say), but this is not really necessary if all comparisons are made within the same population.

The impact of this weighting by age at death is shown in Table 3–4. For the U.S. in 1998, it shows the top ten disease categories as ranked by YPLL with $X = 75$ and the top ten as ranked by number of deaths. Disease categories such as injuries, which tend to kill people at younger ages, rise higher in the ranking by YPLL.

Criticisms of YPLL include the fact that the choice of a cutoff age X is somewhat arbitrary; rankings by YPLL depend on the age distribution of the population at risk, which also affects comparability of YPLL between populations or over time; and the implicit assumption that persons who died of a certain disease before age X years would otherwise have lived to age X or beyond (Gardner and Sanborn, 1990; Lai and Hardy, 1999). Nonetheless, YPLL is increasingly reported as a measure of disease impact on a population and conveys information that other such measures may not readily capture.

EXERCISES

1. Atrial fibrillation (AF) is a heart rhythm abnormality that can be either chronic or "paroxysmal" (occurring in repeated episodes). AF increases the risk of stroke, but the excess risk can be reduced by taking anticoagulants.

   To estimate the prevalence of AF among older adults in a certain region of England, 4843 persons were sampled at random from a list of all persons aged 65 years or older who were registered with a National Health Service primary care physician. Of the 3678 who participated and had an electrocardiogram, 207 were found to have AF.

   To check for participation bias, medical records were also reviewed for a sample of participants and for a sample of nonparticipants. A diagnosis of AF was found somewhere in the medical record for 139/1413 in the participant sample and for 40/382 in nonparticipants.

   (a) Based on these results, what is your best estimate of the prevalence of AF among older adults in the region?

   (b) Do the results from medical record review for a subsample of participants and nonparticipants suggest that persons with AF were any more or less likely to be surveyed?

   (c) Why do you think the percentage of patients with AF in the medical record substudy was so much higher than the percentage found to have AF in the survey?

2. The so-called "sex ratio" is usually calculated as the number of male cases of a condition divided by the number of female cases.

   (a) You are studying patterns of disease occurrence in your community using data on hospital discharges. The sex ratio in 80 cases of pyloric stenosis, which is almost always diagnosed during the first year of life, is found to be 3:1. (Duplicate hospitalizations by the same patients have been eliminated.) Does this finding suggest that male babies are at higher risk for pyloric stenosis than are female babies in your community? Why or why not?

   (b) Below age 75, the sex ratio for myocardial infarction is found to be 2:1. Above age 75, it is about 1:2. Does this imply that men in the area are more prone to heart attacks below age 75, but that women are more prone after that age? Why or why not?

3. Lenaway et al. (1992) described epidemiologic characteristics of school-related injuries among 5,518 students in nine schools in the Boulder, Colorado, area during a particular school year. During this period, 509 injuries were reported, which occurred at the following times:

TIME	PERCENT
Before school	2%
Morning	41%
Lunch	27%
Afternoon	16%
After school	14%
Total	100%

From this information, can you conclude that the risk of injuries was highest during the morning hours? Why or why not?

4. If a hen and a half lay an egg and a half in a day and a half, how many eggs would one hen lay in three days?
5. Vancouver, British Columbia, and Seattle, Washington, are geographically near each other and are quite similar with regard to population size and several measures of socioeconomic status. Over a seven-year period, the following data were obtained from the respective police departments concerning homicides, according to the weapon used.

Percentage of Homicides Committed Using Each Weapon Type

TYPE OF WEAPON	SEATTLE	VANCOUVER
Firearm	42.5%	14.3%
Knife	27.4%	50.0%
Other	30.1%	35.7%

A newspaper reporter is sitting beside you when these data are shown at a press conference. He voices his conclusion that a Seattle resident may be more likely than a Vancouver resident to be shot to death by someone else, but that Seattleites can at least take comfort in knowing that they are less likely to be stabbed to death or killed by other weapons than are Vancouver residents. Do you agree? Why or why not?

ANSWERS

1. *(a)* Prevalence = 207/3678 = .056.
   (b) AF was found for 139/1413 = 9.8% of participants and 40/382 = 10.5% of nonparticipants, suggesting little participation bias.
   (c) The kind of prevalence measured in the community survey was *point prevalence* as of the time the electrocardiogram was taken for each participant. The kind of prevalence measured in the medical record review is better considered *period prevalence*. It referred not to the proportion of patients who had AF at a particular *point* in time, but over the *period* of time during which patients had received care from the clinic whose medical record was reviewed.
2. *(a)* In this instance, yes. At least in most societies, it would be safe to assume that there are about equal numbers of male and female babies at risk during the first year of life, even though the exact numbers at risk may be unknown.
   (b) Not necessarily. The shift in the sex ratio with advancing age might be largely due to differences in the gender composition of the population at risk, with women outnumbering men at the older ages because they generally live longer.
3. No. We can convert the percentages back to the number of cases that occurred during each time period to get a set of numerators for some kind of incidence measure. We could also probably assume that the number of students at risk during each of the time periods shown was about the same. But the *duration* of each time period, while not specified, undoubtedly differed among the time periods. The lunch period, for example, probably lasted only an hour or less, while morning could have spanned three or four hours. Clearly, the longer the time period, the more injuries we would expect to see in the period, even if the intrinsic risk to students per unit of time were the same.

   A good incidence measure here would be the incidence rate, computed using a person-time denominator. We cannot calculate it from the data given for lack of the time component of the denominator.
4. This familiar riddle is actually an incidence-rate problem. The number of eggs laid should be proportional to the number of hens and to the amount of time spent waiting for eggs. The "incidence rate" of egg-laying is 1.5 eggs/(1.5 hens × 1.5 days) = 2/3 eggs/hen-day. One hen on the job for three days amounts to 3 hen-days, so we would expect 3 × 2/3 = 2 eggs.
5. The table concerns only "numerator data" on the distribution of homicides by weapon type. It does not show whether the incidence of homicides, overall or of any type, is higher in one city than in the other.

Here are the actual homicide incidence rates from the two cities during 1980–1986 (Sloan et al., 1988):

Incidence of Homicide per 100,000 Person-Years by Weapon Type

TYPE OF WEAPON	SEATTLE	VANCOUVER
Firearm	4.8	1.0
Knife	3.1	3.5
Other	3.4	2.5
All types	11.3	7.0

The overall incidence of homicide was higher in Seattle, and the difference in rates for firearms accounted for most of the excess. The incidence of homicide carried out with knives was slightly higher in Vancouver, but the incidence of murder involving other weapons was actually higher in Seattle than in Vancouver.

REFERENCES

Ast DB, Schlesinger ER. The conclusion of a ten-year study of water fluoridation. Am J Public Health 1956; 46:265–71.

Babbage C. Letter to Alfred Lord Tennyson. Quoted in: Newman JR (ed.). The world of mathematics. Volume 3, p. 1487. New York: Simon and Schuster, 1956.

Berkel J, de Waard F. Mortality pattern and life expectancy of Seventh-day Adventists in the Netherlands. Int J Epidemiol 1983; 12:455–59.

Bernstein CN, Blanchard JF, Rawsthorne P, Wajda A. Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study. Am J Epidemiol 1999; 149:916–24.

Centers for Disease Control and Prevention. Premature mortality in the United States: public health issues in the use of years of potential life lost. MMWR 1986; 35:1S–11S.

Centers for Disease Control and Prevention. Diabetes during pregnancy—United States, 1993–1995. MMWR 1998; 47:408–14.

Centers for Disease Control and Prevention. HIV/AIDS Surveillance Report 12 (No. 2). Atlanta, Ga.: Centers for Disease Control and Prevention, 2001a.

Centers for Disease Control and Prevention. Sexually transmitted disease surveillance, 2000. Atlanta, Ga.: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2001b.

Eisenberg MS, Gaarslev K, Brown W, Horwitz M, Hill D. Staphylococcal food poisoning aboard a commercial aircraft. Lancet 1975; 2:595–99.

Elandt-Johnson RC. Definition of rates: some remarks on their use and misuse. Am J Epidemiol 1975; 102:267–71.

Gardner JW, Sanborn JS. Years of potential life lost (YPLL)–what does it measure? Epidemiology 1990; 1:322–29.

Gardner LI, Landsittel DP, Nelson NA. Risk factors for back injury in 31,076 retail merchandise store workers. Am J Epidemiol 1999; 150:825–33.

Lai D, Hardy RJ. Potential gains in life expectancy or years of potential life lost: impact of competing risks of death. Int J Epidemiol 1999; 28:894–98.

Last JM (ed.) A dictionary of epidemiology (4th edition). New York: Oxford, 2000.

Lenaway DD, Ambler AG, Beaudoin DE. The epidemiology of school-related injuries: new perspectives. Am J Prev Med 1992; 8:193–98.

Marrett LD. Estimates of the true population at risk of uterine disease and an application to incidence data for cancer of the uterine corpus in Connecticut. Am J Epidemiol 1980; 111:373–78.

Massie DL, Campbell KL, Williams AF. Traffic accident involvement rates by driver age and gender. Accid Anal Prev 1995; 27:73–87.

Miettinen O. Estimability and estimation in case-referent studies. Am J Epidemiol 1976; 103:226–35.

Morgenstern H, Kleinbaum DG, Kupper LL. Measures of disease incidence used in epidemiologic research. Int J Epidemiol 1980; 9:97–104.

National Center for Health Statistics. Health, United States, 2001, with urban and rural health chartbook. Hyattsville, Md.: National Center for Health Statistics, 2001.

National Highway Traffic Safety Administration. 2000 Motor vehicle traffic crashes, injury and fatality estimates, early assessment. Washington, D.C.: U.S. Department of Transportation, 2001.

Phillips DP, Christenfeld N, Ryan NM. An increase in the number of deaths in the United States in the first week of the month. An association with substance abuse and other causes of death. N Engl J Med 1999; 341:93–98.

Rocca WA, Cha RH, Waring SC, Kokmen E. Incidence of dementia and Alzheimer's disease. A reanalysis of data from Rochester, Minnesota, 1975–1984. Am J Epidemiol 1998; 148:51–62.

Sloan JH, Kellermann AL, Reay DT, Ferris JA, Koepsell T, Rivara FP, et al. Handgun regulations, crime, assaults, and homicide. A tale of two cities. N Engl J Med 1988; 319:1256–62.

Vandenbroucke JP. On the rediscovery of a distinction. Am J Epidemiol 1985; 121:627–28.

Vlahov D, Brewer TF, Castro KG, Narkunas JP, Salive ME, Ullrich J, et al. Prevalence of antibody to HIV-1 among entrants to U.S. correctional facilities. JAMA 1991; 265:1129–32.

4

DISEASE FREQUENCY: ADVANCED

Chapter 3 offered an overview of ways to measure disease frequency in populations. In this chapter we return to take a closer look at several of the main techniques, highlighting properties and relationships among them that may not be apparent on a first encounter.

People embark on the study of epidemiology from varying backgrounds and with varying amounts of statistical training. Readers without much prior statistical training may find parts of this chapter challenging but are encouraged to try to follow the basic reasoning and conclusions without getting too bogged down in mathematical details. Those with more statistical experience should find a few helpful connections between new terminology and familiar concepts.

PREVALENCE

Prevalence and Length-Biased Sampling

Not all cases of a disease necessarily have an equal chance of being included in a set of prevalent cases, which are counted in the numerator of a prevalence estimate. The reason is that the time course of many diseases is quite variable from person to person, and an individual's chance of being a case at the time of a prevalence survey depends on how much time he or she spends in the diseased state.

Coronary heart disease, for example, can take several forms, including chronic cardiac chest pain (angina pectoris), acute myocardial infarction, or sudden cardiac death. Figure 4–1 shows the time course of coronary heart disease for three hypothetical cases in a workforce population of middle-aged men during a one-year period. At mid-year, case no. 1 develops chronic angina pectoris, which lasts through the end of the year and beyond. Case no. 2 experiences a myocardial infarction early in the year and dies a week later. Case no. 3 remains disease-free until late in the year, when he suddenly collapses with ventricular fibrillation and soon dies of sudden cardiac death.

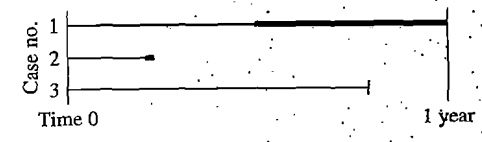

Figure 4–1. Three Hypothetical Cases of Coronary Heart Disease.

Now suppose that an employee health survey seeks to measure the prevalence of coronary heart disease by enumerating all prevalent cases in the workforce. For simplicity, say that a questionnaire is sent to all employees simultaneously and that all eligible cases are identified. This protocol is tantamount to drawing a vertical line somewhere in Figure 4–1, at a horizontal position reflecting the date of the survey, and counting all active cases crossed by that line. If a survey date is chosen at random, case no. 1 has about a 50% chance of being included as a prevalent case. Case no. 2 has about a 1/52 chance of being included, because only a few potential survey dates would fall within the week when he is an active case. Case no. 3 has an infinitesimal chance of being included—the survey would have to reach him between the onset of ventricular fibrillation and when he dies a few minutes later.

Other things being equal, a person's probability of being captured as a prevalent case is proportional to the duration of his or her disease. A set of prevalent cases thus tends to be skewed toward cases with more chronic forms of the disease. This principle has important implications for the design of some kinds of epidemiologic research—particularly case-control studies, to be discussed in Chapter 15. For example, a set of prevalent cases may not be ideal for use in a case-control study of etiologic risk factors, because the frequency of any risk factor that is also associated with chronicity of the disease may be distorted among such cases (Wang et al., 1999). The same principle arises in evaluating the effects of disease screening programs: screening is like a prevalence survey, and cases detected by screening tend to be skewed toward more slowly progressive forms of pre-symptomatic disease (Morrison, 1992).

Confidence Limits

Prevalence is a proportion. Methods of obtaining confidence limits for an estimate of a proportion based on a simple random sample are described in Appendix 4A.

Example: In the Newburgh, New York, dental-decay survey described in Chapter 3, 116 first-graders were found to meet the case definition for dental decay, out

of 184 first-graders examined. The estimated prevalence was $116/184 = .63$. The 95% confidence limits for this estimate may be calculated as follows.

$$c = \text{number of cases} = 116$$

$$n = \text{number of examinees} = 184$$

$$\hat{p} = \text{point estimate of prevalence} = c/n = .63$$

$$se(\hat{p}) = \text{standard error of } \hat{p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.63(1-.63)}{184}} = .0356$$

$$Z_\alpha = \text{standard normal deviate for desired confidence level} = 1.96 \text{ (for 95\% confidence limits)}$$

The desired 95% confidence limits for prevalence are:

$$\hat{p} \pm Z_\alpha \times se(\hat{p}) = .63 \pm 1.96 \times .0356 = (.56 \text{ to } .70)$$

When an observed prevalence is based on a complete enumeration of all cases in the population, an argument can be made that no sampling error is involved and that confidence limits are unnecessary. Even in this situation, however, an observed prevalence is ordinarily treated as an estimate of the true prevalence in a larger source population from which the study population has been sampled at random. The study sample may also be regarded as a random sample in time.

CUMULATIVE INCIDENCE

Estimating Cumulative Incidence in the Presence of Censoring

Sometimes we would like to estimate cumulative incidence but cannot do so directly because some persons drop out during the observation period, even though they had not become a case before dropping out. Disease occurrence information on such subjects is termed *censored*. Censoring can occur for many reasons,

including voluntary withdrawal, departure from the disease-surveillance system's coverage area, death from some unrelated disease, or the scheduled end of study data collection. *Survival analysis* encompasses a family of biostatistical methods that can allow the epidemiologist to estimate cumulative incidence in the presence of censoring, under certain assumptions (Kalbfleisch and Prentice, 1980; Hosmer and Lemeshow, 1999; Kleinbaum, 1996). A simple and widely used method that requires relatively few assumptions is described here: the *Kaplan-Meier* or *product-limit* method (Kaplan and Meier, 1958).

To see how the method works, a specific context will be helpful. An abdominal aortic aneurysm is a balloon-like expansion of the abdominal aorta caused by weakening of the aortic wall. Theory predicts that the larger the aneurysm grows, the weaker the vessel wall becomes and the greater the chance of still further expansion and potentially catastrophic rupture. But surgical repair of an unruptured aneurysm involves significant risk, pain, and cost in its own right. To help decide between early surgery and watchful waiting, doctors and patients need to know the risk of rupture and how it varies over time.

As a "thought experiment," we could imagine monitoring a group of newly diagnosed aneurysm patients over time, without censoring. The cumulative incidence of rupture would rise over time. How high and how quickly the risk rises would help determine the urgency of elective surgery.

In practice, however, aneurysm patients would be diagnosed on widely varying calendar dates, and it would be almost impossible to follow them all until rupture occurred. Censoring could happen due to death from other causes, elective surgical repair, or the scheduled end of data collection.

Example: A study by Nevitt and colleagues (1989) involved tracking the experience of 176 residents of Rochester, Minnesota, who were first diagnosed with an unruptured abdominal aortic aneurysm between 1951 and 1984. Among them, 11 ruptures were identified within eight years after diagnosis. However, even by five years after diagnosis, only 76 of the original patients were actually still at risk for rupture and being followed. Had all 176 patients been tracked for a full eight years without censoring, the cumulative incidence of rupture no doubt would have exceeded 11/176, possibly by a large amount.

Figure 4–2 portrays five hypothetical patients who are diagnosed with an aneurysm at different times during a five-year study period. In the top panel, each patient's experience is shown as a horizontal line that begins at diagnosis and terminates either with rupture (a bold vertical bar) or with censoring (a vanishing line).

In the second panel, the time scale is changed to "time since diagnosis," which bears more directly on the research question at hand. To make this conversion, the

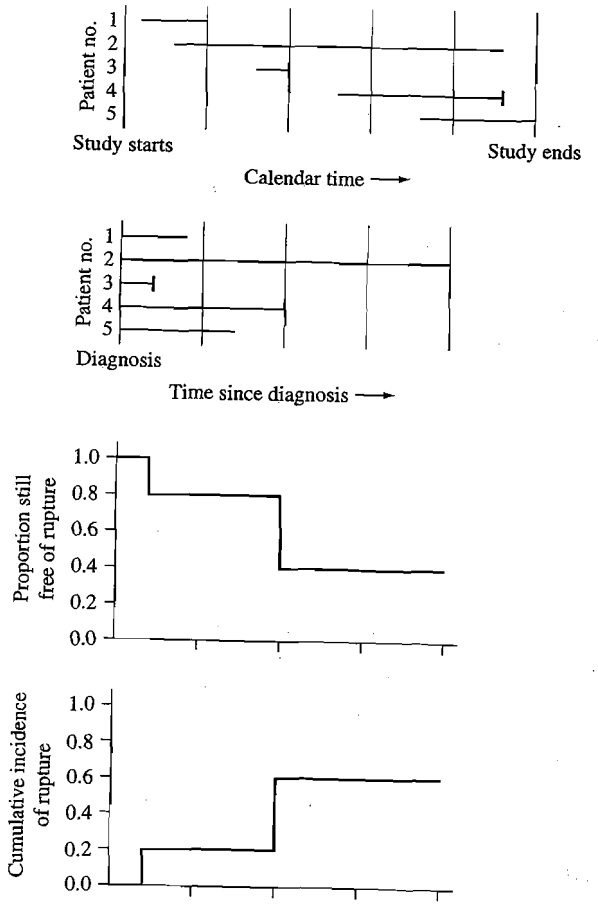

Figure 4–2. Application of Kaplan-Meier Method to Estimate Cumulative Incidence for Five Hypothetical Subjects.

line for each person is moved leftward to align its start with the vertical axis, keeping its length and manner of termination unchanged.

The third panel is the Kaplan-Meier survival curve derived from these data. (Such "curves" normally have a stair-step shape like this one.) Although we are mainly interested in cumulative incidence, it is mathematically more convenient to focus first on "survival"—here, the probability of *not* having had a rupture. To construct the curve, we proceed from left to right. By definition, all patients with a newly discovered unruptured aneurysm are free of rupture at diagnosis (Time 0), so the survival curve begins at height 1.0. Here, no ruptures occurred until halfway

through the first year. Of the five patients then under surveillance, four remained rupture-free immediately after patient no. 3's rupture. Hence the curve drops at that point to $1.0 \times (4/5) = 0.8$. The next rupture (patient no. 4) occurred two years after diagnosis. At that time, one of the two patients still at risk and under surveillance remained rupture-free after patient no. 4's rupture. Hence the curve drops to $0.8 \times (1/2) = 0.4$ at that time. Between the two ruptures, patients no. 1 and no. 5 dropped out due to censoring. Those losses have no effect on the height of the survival curve at the times they occurred, but they increase the size of the drop at the time of patient no. 4's later rupture. After patient no. 4's rupture, no further ruptures occurred through year 4, when the last person under surveillance dropped out.

The bottom panel of Figure 4–2 shows the desired plot of cumulative incidence over time, obtained by calculating cumulative incidence = 1 − (proportion still free of rupture) at each follow-up time. In effect, it is the third panel turned upside-down. Note that the estimated cumulative incidence of rupture at four years is 0.6, not $2/5 = 0.4$, as might have been guessed naïvely without accounting for censoring. A more generic description of the Kaplan-Meier method and related methods can be found in references on survival analysis (Kaplan and Meier, 1958; Kalbfleisch and Prentice, 1980; Kleinbaum, 1996).

When there is no censoring, the Kaplan-Meier method yields the same cumulative incidence estimate as the simpler direct method described earlier. When censoring is present, the method uses the experience of those remaining at risk and under follow-up to estimate the shape of the curve. The validity of the resulting curve and cumulative incidence estimates depends on an assumption that censoring is unrelated to risk. In other words, it is assumed that, had they been observed to completion, the survival curve for persons with censored data would look the same as the curve for everyone else, aside from sampling variability. This assumption is usually not empirically testable, but a judgment about its plausibility can often be made by considering the reasons for censoring. In our aneurysm example, censoring due to the arbitrary end of the study period might well be unrelated to risk and create no bias. But censoring due to surgical repair might be triggered by the onset of symptoms or by evidence of rapid aneurysm growth, so that the surgeon's hand may have been forced by an impending rupture. To the extent that censoring for that reason is common, we might suspect that the cumulative incidence of rupture without surgical intervention could be underestimated by the Kaplan-Meier method.

Survival analysis includes several other conceptually similar but computationally more complex methods for estimating cumulative incidence when adjustment must be made for subject characteristics (covariates) that may differ across comparison groups (Kalbfleisch and Prentice, 1980; Hosmer and Lemeshow, 1999; Kleinbaum, 1996).

Confidence Limits

Cumulative incidence, like prevalence, is a proportion. When it is estimated directly from data on a closed population, methods described in Appendix 4A can be used to obtain confidence limits. When the estimate is obtained by the Kaplan-Meier method, confidence limits must be obtained by more complex techniques, as described in several statistical texts (Kalbfleisch and Prentice, 1980; Hosmer and Lemeshow, 1999; Kleinbaum, 1996; Rosner, 1995).

INCIDENCE RATE

Population-Level and Individual-Level Perspectives

Measures of disease frequency in populations have a dual interpretation. First, they estimate the burden of disease on a population as a whole. This perspective bears directly on such public health activities as detection and tracking of epidemics, health planning and resource allocation, and evaluation of policies and programs, which focus on the population as a unit.

Second, disease frequency in a population is also used to estimate disease *risk* in individuals. From this perspective, the population is viewed as a collection of individuals who have certain characteristics in common. The population is a set of replicate observations. Viewed this way, data on disease frequency in the population provide input for inductive reasoning: predictions about the likely fate of one individual can be based on the observed experience of others.

For example, the percentage of newborn babies weighing less than 2500 grams at birth has been found to be higher among babies of mothers who smoked cigarettes during pregnancy than among babies of non-smoking mothers. For any particular mother, there is no way to know for sure whether she will or will not have a low-birth-weight baby. Yet based on the experience of other mothers, we infer that the *risk* or *probability* that she will have a low-birth-weight baby is greater if she smokes than if she does not. This view of a population as a set of replicate observations also underlies statistical theory for obtaining confidence limits for measures of disease frequency.

Duality of perspectives applies to many disease-frequency measures, not just incidence rates, and epidemiologists are used to moving freely between them. But there are situations in which population-level disease frequency does not necessarily translate directly into individual-level risk estimates, and then it becomes important to distinguish between the perspectives. This issue has special relevance to incidence rates.

Incidence Rate and Hazard Rate

A key feature of the person-time incidence rate is that its denominator is a "lump sum." Person-time is regarded as a freely interchangeable commodity, and all that matters in the final calculation is the total amount. Observing one susceptible person for 12 years, 12 people for one year, or 144 people for a month all result in adding 12 person-years to the denominator. This property was assumed in the calculations shown in Table 3–1.

This feature of the incidence rate can be both a strength and a limitation. The examples in Figure 3–3 and Table 3–1 illustrated that the ability to combine person-time across people and over time makes the incidence rate a much more broadly applicable measure of incidence than cumulative incidence. But to appreciate the implications of this pooling, it is helpful to consider a model that breaks down the population's disease experience into smaller building blocks.

Looking again at Figure 3–3, the vertical lines divide the total observation period into 30 one-day periods. Taken one day at a time, many of the original complicating factors that interfered with direct calculation of cumulative incidence—censoring, recurrent cases, periods not at risk, multiple observation periods per person—become less problematic. On any given day, recurrent events in the same person and gains and losses to the population at risk are rare or non-existent.

Furthermore, we can imagine that if we had detailed data on the timing of events, we could extend this divide-and-conquer strategy still further, splitting days into hours, hours into minutes, and so on, rather like viewing a movie one frame at a time. No two disease events occur in the same person at exactly the same time on a sufficiently fine time scale, so in principle it is always possible to choose a time increment short enough that the chance of multiple cases occurring within the same increment is negligible.

In addition, there is a fixed number of instances when some population member joined or left the population, or moved into or out of the susceptible state. But again, there is no limit on how short a time increment we could select. Hence the *proportion* of intervals involving censoring of this sort can be made as small as we wish, and ultimately rare enough to be negligible.

So suppose that we specify a certain short time increment, such as a minute or a second—short enough to eliminate recurrent cases within a single interval and short enough to allow censoring and susceptibility changes to be ignored. Call this increment Δt. The entire study period of interest is then split into a series of intervals, each of duration Δt. The population's disease experience over time could now be represented as a very large matrix, as illustrated in Figure 4–3. Each row refers to a different individual who belongs to the population for at least part of the study period. Consecutive columns refer to consecutive short time intervals throughout the observation period. Each cell corresponds to a tiny piece

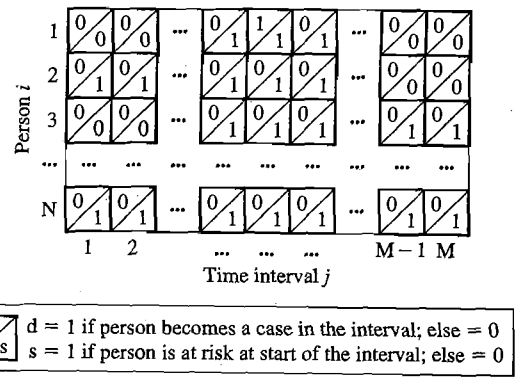

Figure 4–3. Matrix Representation of Population Disease Experience.

of person-time. Rows (individuals) are numbered 1 through N, while columns (time intervals) are numbered 1 through M. Any particular cell can be referred to by its row number (i) and column number (denoted j).

This matrix can be thought of as a "digitized" version of the kind of population line diagram shown in Figure 3–3. Each row of the line diagram in Figure 3–3 maps to a row of the matrix in Figure 4–3. The horizontal resolution is adjustable by the choice of Δt, with smaller values of Δt yielding finer resolution.

Each cell in Figure 4–3 contains two numbers. The upper left number in cell ij is 1 if the person in row i became a case during time interval j; otherwise it is 0. In other words, this number is the value of a *disease* indicator, d_{ij}. The lower right number in that cell is 1 if person i was at risk and under observation at the start of time interval j; otherwise it is 0. It is the value of a *susceptibility* indicator, s_{ij}.

Cells for which $s = 0$ are of little interest, corresponding to persons who were non-susceptible, not under observation, or already ill at the time. We know in advance that $d = 0$ for those cells—there is no uncertainty, and the observed value of d provides no real information. But every cell for which $s = 1$ corresponds to a brief "experiment of nature," or binomial trial, as considered earlier for cumulative incidence. "Chance" (as a euphemism for our incomplete knowledge about disease causes) determines whether d is 1 or 0. Each of these *trial-cells* captures the experience of a single at-risk individual over a short, fixed time period: person i either develops disease during time interval j or does not.

Associated with each trial is an underlying probability p_{ij} that person i would become a case during interval j. Because Δt was already made short enough to let us ignore the possibility of recurrent disease events within one interval, p_{ij} can also be interpreted as the *expected number* of disease events in that trial-cell. The true values of these p's are not observable. A trial-cell, however, contains either 0/1 or

1/1, either of which can be read as the observed cumulative incidence in a certain one-person population over a certain brief time interval. It is a crude estimate—the crudest possible estimate, in fact—of the corresponding p. Symbolically, $\widehat{p_{ij}} = d_{ij}/s_{ij} = d_{ij}$, because the denominator is always 1 for all trial-cells.

The total number of trial-cells, and the probability that a case occurs in one of them, clearly depend on the choice of Δt. The binomial-trial model fits better and better as Δt gets shorter, so it is worth considering what happens as Δt approaches 0. The smaller Δt is, the more trial-cells there are, and the less likely it is that any given one of them contains a case. Moreover, as Δt gets near enough to 0, p_{ij} becomes *proportional* to Δt. This is because a fixed number of cases is being spread over an increasingly large number of cells. Over a short enough period, p remains essentially constant for the same person from one instant to the next. Dividing p expected cases among, say, k still shorter sub-intervals puts p/k expected cases into each sub-interval of duration $\Delta t/k$. The k's cancel out, and the *ratio* of expected events to interval duration remains unchanged.

This ratio is variously termed the *hazard* or *hazard rate* (our preferred terms), *event rate, probability rate, force of morbidity,* or *instantaneous probability*. It is usually denoted with the Greek lambda (λ):

$$\lambda_{ij} = \frac{p_{ij}}{\Delta t}$$

For a given individual, a value of λ can be associated with each *point* in time by taking the limiting value of λ as $\Delta t \to 0$ for the interval that includes that time point.

The hazard rate merits some contemplation. It is the expected number of disease events per unit of time for a certain person at a certain moment (which explains why *event rate* is one of its aliases). Computed as the ratio of a unitless probability to an amount of time, the hazard rate has units of time^{-1}, just as does the incidence rate. But we must remember that incidence is an observable population measure of disease occurrence over a period of time, while the hazard rate is an unobservable individual measure of disease risk per time unit, evaluated at a moment in time.

We can now re-aggregate the data in the many trial-cells back toward what we as epidemiologists observe at the population level. The observed number of cases, c, is the total number of trial-cells for which $d = 1$, summed over all rows and columns:

$$c = \sum_i \sum_j d_{ij}$$

The total time at risk (call it T with no subscript), summed for all population members, is the number of trials for which $s = 1$, times the duration of each trial:

$$T = \sum_i \sum_j s_{ij} \cdot \Delta t$$

The total number of trials is:

$$N = T/\Delta t$$

Finally, the *mean hazard rate* across all persons and time intervals at risk is:

$$\bar{\lambda} = \frac{\sum_i \sum_j \frac{p_{ij}}{\Delta t}}{N}$$

$$= \frac{\sum_i \sum_j \frac{p_{ij}}{\Delta t}}{T/\Delta t}$$

$$= \frac{\sum_i \sum_j p_{ij}}{T}$$

We noted earlier that p_{ij} is interpretable as the expected number of cases in cell ij, so the numerator of the last expression is the total expected number of cases for the population. The *expected* total number of cases is not directly observable, but in a particular set of data, the *observed* number of cases c, is an estimate of it. Hence:

$$\text{Incidence rate (IR)} = \frac{c}{T} \text{ estimates } \bar{\lambda}$$

In words, the incidence rate is an estimate of the mean hazard rate over all person-time at risk contributed by population members during the study period. This interpretation holds true regardless of how variable the hazard rate may be within individuals and over time. It also does not require that the p_{ij} be independent of one another.

Incidence rate as a weighted average

As noted earlier, one of the main uses of incidence data in a population is to infer disease risk in individual population members. Later in this chapter, we will examine a "weighted-average rule," showing that the rate for a population is always a weighted average of rates for the population's component parts, which are usually subgroups of the population. The weight for each subgroup is its size—here, how much person-time it contributes. In the present context, the weighted-average rule says that the overall incidence rate will be most heavily influenced by hazard rates that apply over the largest amounts of person-time.

- Suppose the hazard rate is constant for all individuals and over time. Then the incidence rate estimates a weighted average of this constant, which is just the constant itself.
- Suppose each individual has his or her own possibly unique hazard rate, λ_i, which remains constant over time. Then the incidence rate estimates a weighted average of the λ_i. The weight for each λ_i is the proportion of the total person-time at risk contributed by person i. Hazard rates that apply to persons who are at risk and under observation longer are thus weighted more heavily. This can be important if censoring is more or less common among individuals with relatively high hazard rates. The incidence rate will tend to be skewed toward hazard rates among those persons who are least subject to censoring.
- Suppose the hazard rate changes over time but is similar for all at-risk individuals who are under observation at a given time. This situation might apply, for example, in a cohort of similar individuals who are tracked for occurrence of new cases over a long time. As cohort members age, their hazard rates may change. Concurrently, members of the cohort may be lost to attrition, so that more person-time comes from earlier in the study period than from later. The incidence rate will be skewed toward hazard rates in effect during early parts of the observation period, when more people were being observed.

Confidence Limits

As noted earlier, few assumptions are needed in order to interpret the incidence rate as an estimate of the mean hazard rate. To obtain confidence limits for an incidence rate estimate, however, additional assumptions must be made. Three models are discussed here and some research situations to which each might correspond.

Constant hazard

The simplest and probably most widely used assumption is that the hazard rate is constant across individuals and over time. Under this assumption, all person-time within the observation period is freely interchangeable. An analogy from physics is decay of a radioactive element: the hazard rate of fissioning in a certain atom at a certain moment is thought to be constant across all atoms of the same isotope and over time (Armitage and Berry, 1994).

In the human health arena, there are probably not many exact counterparts—perhaps the hazard of being struck by a giant meteor from outer space—but some

situations come closer than others. The constant-hazard assumption is most plausible for a relatively homogeneous population observed over a relatively short time period. "Relatively homogeneous" and "relatively short" need to be interpreted in the context of the disease in question. In occupational epidemiology, for example, cases and person-time at risk are routinely partitioned into categories defined by job title, age range, gender, and calendar year, with incidence being estimated separately for each of the resulting categories (Checkoway et al., 1989).

When this constant-hazard assumption is met, methods based on the Poisson distribution can be used to obtain confidence limits for an incidence rate estimate (Mendenhall et al., 1986; Armitage and Berry, 1994; Breslow and Day, 1987). Specifically, say that c cases are observed in T person-time. T is considered a fixed quantity, not subject to sampling error, and confidence limits for the rate are based on confidence limits for c alone. If $c > 100$, the following expression (based on the normal approximation to the Poisson distribution) is reasonably accurate (Armitage and Berry, 1994, p. 142):

$$\text{Confidence limits} = \frac{c \pm Z_\alpha \cdot \sqrt{c}}{T}$$

where Z_α is the standard normal deviate for the desired confidence level ($Z_\alpha = 1.96$ for two-sided 95% confidence limits). The second term in the numerator is subtracted to get the lower confidence limit and added to get the upper limit.

If $c \leq 100$, more accurate confidence limits can be obtained by basing them directly on the Poisson distribution. Table 4–3 in Appendix 4B can be used to obtain a lower and an upper multiplier for c, the observed case count. Multiplying each of these by the observed point estimate of incidence yields the desired lower and upper confidence limits.

Example: In the study of back injuries by Gardner et al. (1999) described in Chapter 3, nine back injuries were reported in 322,193 working hours by female department managers who had been employed for less than eight months, for a rate of 2.79 cases per 100,000 worker-hours. Using the table in Appendix 4B, the upper and lower multipliers needed to obtain Poisson 95% confidence limits for an incidence rate that is based on nine cases are 0.457 and 1.898. The desired confidence limits therefore extend from $2.79 \times 0.457 = 1.28$ to $2.79 \times 1.898 = 5.30$ cases per 100,000 worker-hours.

Hazard varying randomly among individuals

In some situations, theory or available data suggest that the hazard rate varies among individuals, even after accounting for measured differences in exposure

to risk factors and other personal characteristics. This variation could arise, for example, from differences in genetic susceptibility, differences in exposure to unmeasured risk factors, or just biological variation. In the biostatistical literature, random inter-individual differences in hazard rates are called differences in *frailty* (Aalen, 1994; Clayton, 1994).

This model seems particularly applicable to studies of recurrent illness (Glynn et al., 1993; Glynn and Buring, 1996; Cumming et al., 1990). Under the constant-hazard assumption considered earlier, someone who has had one disease event is no more or less likely than anyone else to have another event in the future. But for many diseases, evidence suggests that future risk is often elevated among persons who have already experienced an initial event. For example, victims of assault have been found to be at greatly increased risk of being assaulted again (Dowd et al., 1996). Children treated for an unintentional injury are more likely than are other children to experience a future unintentional injury (Johnston et al., 2000). Postmenopausal women who experience a vertebral fracture are at high risk of having an additional fracture within the next year (Lindsay et al., 2001). Possible mechanisms include continued exposure to a hazardous environment, existence of a chronic underlying health condition that predisposes to recurrent complications, or effects of the initial illness event itself, as might occur if an assault victim confronted his or her attacker. Whatever the reason, an initial event may serve as a marker for a subpopulation with a systematically higher hazard rate. Statistically, the problem is known as *extra-Poisson variation*. When it is present, confidence limits based on the Poisson distribution, which assumes constant hazard, are too narrow (Glynn and Buring, 1996; Clayton, 1994).

Several statistical approaches have been proposed to deal with this problem (Glynn and Buring, 1996; Clayton, 1994; Sturmer et al., 2000). One involves computing an individual event rate for each population member based on his or her observed number of events and person-time at risk, and basing confidence limits for the overall incidence rate on the observed variance in those event rates across persons (Glynn and Buring, 1996). When follow-up times are unequal among individuals, the individual rates can be weighted by amount of time at risk (Stukel et al., 1994). More complex multivariate methods include logistic regression using generalized estimating equations, Poisson regression with correction for overdispersion, or adaptations of proportional-hazards survival analysis (Sturmer et al., 2000).

Variation in hazard rates among individuals can have another effect in the context of non-recurrent disease in closed populations. It can affect the degree to which changes in incidence rates in the population over time reflect corresponding changes in individual risk (Aalen, 1994, 1988). For example, suppose that a population under surveillance initially consists of a 50:50 mixture of a high-risk subgroup and a low-risk subgroup. The earliest cases arise mainly from the high-risk

subgroup. But as those early cases occur, they also preferentially deplete the high-risk subgroup. Over time, the original 50:50 mixture thus shifts toward an increasing predominance of low-risk individuals, which in turn yields a decline over time in incidence for the population as a whole. The observed decline in incidence could be misinterpreted as implying a decline in risk to individual population members, either as they age or as calendar time passes, when in fact the decline would be due at least in part to changes in the composition of the population at risk.

Hazard varying over time

As noted earlier, the simple and common model of constant hazard rates can be expected to hold best over relatively short observation periods. As time passes, changes in such factors as exposure to environmental causes, diagnostic methods, and disease classification commonly occur and can affect disease frequency. Closed populations also age. To reduce variation in hazard over time in the face of these factors, a long period of observation is often subdivided into shorter sub-periods or "time bands" for analysis. Other analytic strategies include modeling the effects of time itself on incidence—for example, by including time as a predictor in the kinds of multivariate models to be described later (Chapter 11).

Often, however, changes in hazard rates over time are not of main interest and are instead just a potential source of bias when making comparisons among subgroups or populations. This viewpoint has helped make the proportional-hazards model popular in epidemiology (Cox, 1972; Kalbfleisch and Prentice, 1980; Kleinbaum, 1996). Briefly, under this model, the "baseline" hazard may change over time in an arbitrary way, and these changes are assumed to apply to all individuals. But at any given moment, an individual's hazard relative to that of other individuals then at risk is assumed to depend on his or her measured personal characteristics, at least one of which is exposure to a potential risk factor of main interest.

Incidence Rate and Mean Time to Disease Onset

The incidence rate has units of time^{-1}. Under somewhat idealized circumstances, the *reciprocal* of incidence, which is in units of time, can be interpreted as the mean time to disease onset. Although this odd fact is perhaps of more theoretical than practical importance in epidemiology, its basis is explained here. It will soon play a role in linking prevalence, incidence, and disease duration.

Consider a hypothetical population of N susceptible individuals who are followed indefinitely for development of a non-recurrent disease. Say that incidence rate remains constant at some value IR throughout the follow-up period. If there are no competing risks, and if the population is followed long enough, then everyone in it must eventually develop the disease. Before becoming a case, person i contributes a certain amount of person-time at risk, T_i. Total person-time, $T = \sum_i T_i$, stops increasing when the last case occurs. At that time, N cases would have occurred in T person-time. By definition, the incidence rate was constant throughout follow-up, so $IR = N/T$.

Now suppose we are interested in how much time goes by, on average, until a susceptible person becomes a case. This would be $\sum_i T_i/N = T/N = 1/IR$. In other words, the reciprocal of the incidence rate estimates the mean time to disease onset under the circumstances described. Although a proof is beyond the scope of this text, this property also holds for recurrent diseases.

Example: Say that upper respiratory infections occur at the (very high) incidence of three per person-year in a population. The average time to the next upper respiratory infection for a person at risk would be $1/(3\ \text{year}^{-1})$, or 4 months.

For lower-incidence diseases, the required assumption of no competing risks will rarely be satisfied, in which case the resulting numerical estimate of mean time to disease onset may not be very meaningful. But the algebraic rule itself will prove useful below.

RELATIONSHIPS AMONG MEASURES OF DISEASE FREQUENCY

Populations and Their Subpopulations

Many commonly used measures of disease frequency take the form of fractions. The numerator is usually a case count, and the denominator is a measure of population size or of the amount of person-time in which those cases occurred. Prevalence, cumulative incidence, person-time incidence rate, mortality rate, case fatality, and various other measures all fit this description. It will be convenient to call all such measures "rates" for now, recognizing that "rate" has a narrower meaning in other contexts.

A simple and very useful algebraic relationship connects the value of any such rate in the whole population to its value in subpopulations formed from the whole.

Suppose that, for a certain study population, an overall rate of disease, r, is calculated by dividing the total number of cases, c, by an appropriate denominator, n, so that $r = c/n$. The population can be divided in various ways into a set of *mutually exclusive and collectively exhaustive* subgroups—for example, by gender, by age category, or by exposure to some environmental factor. A separate "local" rate can then be calculated for each subgroup, simply by restricting both the numerator and the denominator to members of that subgroup.

Imagine that the population is separated into males and females, and that the gender-specific rates are $r_m = c_m/n_m$ for males and $r_f = c_f/n_f$ for females. The rate in the full population is:

$$r = \frac{c_m + c_f}{n_m + n_f}$$

Some algebra shows that:

$$r = \frac{c_m}{n_m + n_f} + \frac{c_f}{n_m + n_f}$$

$$= \frac{c_m}{n_m} \cdot \frac{n_m}{n_m + n_f} + \frac{c_f}{n_f} \cdot \frac{n_f}{n_m + n_f}$$

$$= r_m \cdot w_m + r_f \cdot w_f$$

where $w_m = n_m/(n_m + n_f)$ is the proportion of the overall denominator contributed by males, and w_f is the proportion contributed by females. Note also that $w_m + w_f = 1$. The w's can be interpreted as *weights* for the corresponding gender-specific rates.

Example: In a population of 700 women and 300 men, there are 35 prevalent cases of diabetes among the women and 30 cases among the men. The overall prevalence of diabetes is thus $(35 + 30)/(700 + 300) = 65/1000 = 6.5\%$, while the gender-specific prevalences are $35/700 = 5.0\%$ in women and $30/300 = 10\%$ in men. The overall prevalence is a 700:300 weighted average of 5% and 10%: $6.5\% = (5\%)(0.7) + (10\%)(0.3)$.

The algebra above can be extended to cover any number of mutually exclusive and collectively exhaustive subgroups. The general rule is:

> The overall disease rate in a population is a weighted average of the rates in its subpopulations. The weight for each subpopulation rate is the proportion of the overall rate's denominator contributed by that subpopulation.

This property applies to all disease-frequency measures that are fractions. Among other uses, the rule underlies direct and indirect standardization of rates (discussed in Chapter 11)—techniques that enable valid comparison of rates across populations that differ with regard to sociodemographic or other characteristics.

Cumulative Incidence and Incidence Rate

A useful relationship between the two main measures of incidence can be developed for non-recurrent diseases. Consider again what would happen over time in a closed, susceptible population in which the incidence rate of a certain

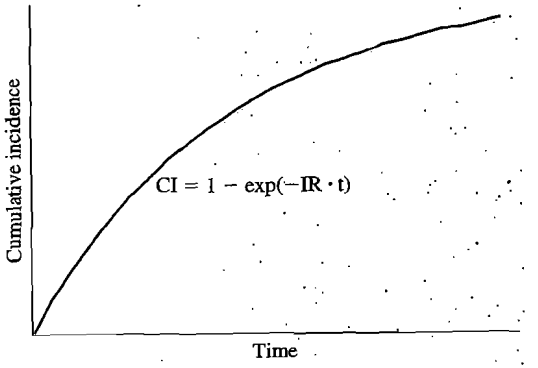

Figure 4–4. Cumulative Incidence over Time When Incidence Rate Is Constant.

non-recurrent disease remains constant. As new cases occur, they are subtracted from the population at risk. It can be shown with calculus that the population at risk declines *exponentially* over time—a process analogous to the exponential decay of a radioactive element.

Say that N_0 is the number of persons originally at risk, N_t is the number remaining at risk at time t, IR is the incidence rate (assumed to be constant), and e is the base of natural logarithms, approximately 2.71828. Then:

$$N_t = N_0 \cdot e^{(-IR \cdot t)} = N_0 \cdot \exp(-IR \cdot t) \qquad (4.1)$$

Because of the steadily declining population at risk, a smaller and smaller *number* of new cases occurs per unit of time. Cumulative incidence (CI) continues to rise, but with decreasing slope. Specifically:

$$CI = 1 - \exp(-IR \cdot t) \qquad (4.2)$$

Figure 4–4 illustrates this relationship.

Relation (4.2) can be handy, for example, when comparing results from two or more studies that used different incidence measures. As we have seen, the incidence rate applies to a broader range of populations and disease types, while cumulative incidence is more easily interpretable in terms of disease probability or risk. Relation (4.2) provides one way to use incidence rate data to address "What if . . .?" questions involving cumulative incidence.

Example: Table 4–1 is based on results from a study by Morris and colleagues (1953) of coronary heart disease incidence among London bus drivers and conductors. It was among the first studies to suggest a link between regular physical

Table 4–1. Using Incidence Rate to Project Cumulative Incidence of Coronary Heart Disease in London Bus Drivers and Conductors

	Results from Morris Study			Projected Experience of 1000 Hypothetical Workers		
JOB/AGE	CASES	PERSON-YEARS	INCIDENCE RATE[a]	AGE	STILL AT RISK	CUMULATIVE INCIDENCE SINCE AGE 35[b]
Drivers						
35–44	8	12,360	0.6	35	1000	—
45–54	29	11,698	2.5	45	994	0.6%
55–64	43	6668	6.4	55	969	3.1%
				65	909	9.1%
Conductors						
35–44	0	9622	0.0	35	1000	—
45–54	11	5522	2.0	45	1000	0.0%
55–64	20	4022	5.0	55	980	2.0%
				65	933	6.7%

[a]Per 1000 person-years

[b]In the absence of competing risks

[*Source:* based on Morris et al. (1953).]

activity and lower heart disease risk. The conductors moved around the bus all day collecting fares, climbing up and down stairs. Meanwhile, drivers remained relatively sedentary while seated driving the bus. Because the work force was an open population, the results were reported as coronary heart disease cases per 1000 worker-years for each job type and for each of three 10-year age categories.

In each age group, the incidence rate of heart disease was higher in drivers than in conductors, consistent with the hypothesis that physical activity lowers the risk of heart disease. The implications of these results in terms of individual risk can be clarified by using the incidence rate data to estimate the cumulative incidence of coronary heart disease in two hypothetical cohorts of 1000 drivers and 1000 conductors from age 35 to 65 years.

On the right side of Table 4–1, relation (4.1) was applied to each age decade in turn. The number of cohort members still at risk for incident coronary heart disease was estimated for the end of each decade, based on how many were at risk at the start of the decade and on the incidence rate for that decade from the Morris study. Cumulative incidence follows directly from the projected number still at risk. For example, among 1000 bus drivers at risk starting at age 35, the number expected *not* to develop coronary heart disease in the next decade would be:

$$1000 \times \exp(-.0006 \cdot 10) = 994$$

Thus, the projected cumulative incidence of coronary heart disease over the 35–44 age decade would be $(1000 - 994)/1000 = .006 = 0.6\%$. For the 45–54 age decade, the same formula is applied again, substituting 994 for 1000 and .0025 for .0006, and so on down the table. We estimate that, in the absence of competing risks, a man who began working as a London bus driver at age 35 would stand a 9.1% chance of developing coronary heart disease within the next 30 years. The corresponding 30-year risk in a conductor would be 6.7%.

In many situations, the disease is rare enough and the observation period short enough that very little reduction in the size of the population at risk takes place during observation. Expressed with like denominators, incidence rate and cumulative incidence may thus appear to be numerically very close. For example, the incidence rate of coronary heart disease among London bus drivers aged 35–44 years was 0.6 cases per 1000 person-years. Based on (4.2), the projected one-year cumulative incidence among 1,000 such drivers would be $1 - \exp(-0.6/1000 \cdot 1) = 0.0005998 = 0.5998$ per 1000 drivers.

Prevalence, Incidence, and Duration

Under certain circumstances, prevalence and incidence can be easily related to each other. Consider a closed population in which a recurrent disease state occurs, such as urinary tract infection, the common cold, or depression. Assume that all individuals who do not have the disease are susceptible (that is, there is no third not-at-risk state), and that the incidence rate is constant at some value *IR* among all susceptibles and over time. Under this simple two-state model, people move back and forth between the states over time (Fig. 4–5).

The flow along the disease-onset path, measured in the number of events per unit of time, depends on *(1)* the size of the susceptible pool and *(2)* the incidence rate. Because the disease is recurrent, there is also a counter-flow of individuals from diseased back to susceptible, which we may call *recovery*. The number of recovery events per unit of time depends on *(1)* the size of the diseased pool, and *(2)* what we may call a *recovery rate*, which is just like the incidence rate but operates in the opposite direction. We shall further assume that this recovery rate is constant over time and is the same for all diseased individuals at some value *RR*.

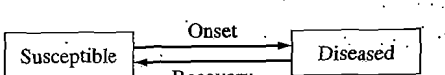

Figure 4–5. Two-State Model to Illustrate the Relationship Among Prevalence, Incidence, and Duration.

As demonstrated earlier, $1/IR$ can be interpreted as the mean time to disease onset among susceptibles. By similar logic, $1/RR$ can be interpreted as mean time to recovery among diseased individuals. In other words, $1/RR$ is the *mean duration* of disease, which it will be convenient to call $\bar{d}$.

Now suppose we begin with all individuals in the susceptible state and imagine what happens over time. People start to develop the disease at a rate determined by IR. In the process, they start emptying the susceptible compartment and start filling the diseased compartment. As the diseased compartment accumulates prevalent cases, recoveries begin to occur. The more prevalent cases accumulate, the more recoveries occur, which tends to empty the diseased compartment and to refill the susceptible compartment. As long as the two opposing flows are unequal, one compartment will grow and the other shrink, which in turn will act to equalize the two flows. Eventually an equilibrium is reached, in which the flow of incident cases is exactly balanced by the flow of recoveries. Because the flows into and out of each compartment are equal at that point, both compartments maintain a stable size.

More formally, suppose that the flows and compartment sizes at equilibrium are labelled as follows:

S = number of susceptible persons

D = number of diseased persons (prevalent cases)

i = number of incident cases per time unit Δt

r = number of recoveries per time unit Δt

At equilibrium,

$$\begin{aligned} i &= r \\ IR \cdot S &= RR \cdot D \\ \frac{D}{S} &= IR \cdot \frac{1}{RR} \\ &= IR \cdot \bar{d} \end{aligned} \tag{4.3}$$

Relation (4.3) says that, under the specified assumptions and at equilibrium, D/S is the product of incidence rate and mean duration of disease. D/S is not quite the prevalence, which would be $D/(S+D)$. Rather, D/S is the *prevalence odds*, which expresses the relative frequency of the diseased state as an odds rather than as a proportion. For many realistic situations, however, the prevalence of disease is low, so that $S \gg D$, and therefore $D/S \approx D/(S+D)$. The final result is then:

$$\text{Prevalence} \approx \text{(Incidence rate)} \times \text{(Mean disease duration)} \tag{4.4}$$

Relation (4.4) links two key measures of disease frequency. It is a time-honored rule of thumb in epidemiology. Nonetheless, it is probably best regarded as a conceptual aid rather than as a relation that can be expected to hold true consistently in real data. The main reason is that the assumptions behind the hypothetical model are often poorly or only approximately met under real-world conditions. For example, the incidence rate and recovery rate may not remain constant long enough for an equilibrium to be achieved, because of changes in environmental or behavioral exposures, disease-control activities, diagnostic technologies, disease treatments, and so on. Also, the population of interest may not be closed, so that in- and out-migration of prevalent cases is possible. Nonetheless, relation (4.4) is quite useful for understanding the two main determinants of disease prevalence and for predicting how prevalence may change as a result of changes in incidence or disease duration.

Example: To illustrate how relation (4.4) works, imagine a population of married women aged 15–45 years in whom the incidence and prevalence of pregnancy are studied. Table 4–2 shows three scenarios. In the "base case," the incidence rate of pregnancy is 8 per 100 woman-years. Full-term pregnancies last 9 months, or 0.75 years. If all pregnancies go to term, and if incidence has been stable long enough for equilibrium to be reached, then a prevalence survey would be expected to find about $8/100 \times 0.75 = 6/100$ of women pregnant on a random survey date.

Now suppose that highly effective oral contraceptives become available for the first time, and a random 50% of women choose to use them. No other changes in reproductive practices occur. Use of the "pill" should reduce the incidence of pregnancy by half to 4 per 100 woman-years, but it should not affect the duration of pregnancies that do occur. Once a new equilibrium is achieved, we would expect another prevalence survey to find about 3/100 women pregnant on a random date.

Table 4–2. Incidence, Duration, and Prevalence of Pregnancy in a Hypothetical Population of Women of Reproductive Age

	Pregnancy		
BIRTH CONTROL USE	INCIDENCE RATE[a]	DURATION (YEARS)	PREDICTED PREVALENCE (APPROX.)
None	8	0.75	6%
50% use "pill"	4	0.75	3%
50% have abortion at 3 months	8	0.50	4%

[a]Pregnancies per 100 woman-years

Starting over from the "base case" without oral contraceptives, suppose instead that elective abortions become available. A random 50% of women who become pregnant decide to terminate the pregnancy at three months, while the rest carry the child to term. Abortion would have no effect on the rate at which women become pregnant, so incidence would remain at 8 per 100 woman-years. But abortions would reduce the average duration of a pregnancy from 9 months to $(0.5)(3) + (0.5)(9) = 6$ months. Hence we would expect fewer women—about 4/100—to be in the pregnant state on a random survey date.

Mortality, Incidence, and Case Fatality

The following relations follow directly from the definitions of mortality, incidence, and case fatality:

$$\text{Mortality rate} \approx \text{incidence rate} \times \text{case fatality}$$

$$\text{Cumulative mortality} \approx \text{cumulative incidence} \times \text{case fatality}$$

Intuitively, we can think of the risk of dying of a disease as (the risk of getting the disease) × (the risk of dying of it if you get it). These relations are shown as approximations because *(1)* the denominators of mortality and incidence differ slightly with regard to inclusion of prevalent cases; and *(2)* some time must pass between disease onset and death from the disease. The incidence rate when disease-related deaths occur may differ from the rate in effect when those cases arose.

For the pedestrian/motor-vehicle collision injury example described in Chapter 3, neither of these caveats would be of serious concern. The United States' population in 2000 was about 282,100,000, so:

$$\begin{aligned}\text{Mortality rate} &\approx \frac{4739 \text{ deaths}}{282{,}100{,}000 \text{ person-years}} \\ &\approx \frac{78{,}000 \text{ cases}}{282{,}100{,}000 \text{ person-years}} \times \frac{4739 \text{ deaths}}{78{,}000 \text{ cases}} \\ &\approx (\text{incidence rate}) \times (\text{case fatality})\end{aligned}$$

APPENDIX 4A

Confidence Limits for a Proportion

Several measures of disease frequency are *proportions*, including prevalence, cumulative incidence, and proportional mortality. Let c be the number of cases in the numerator and n the number of persons in the denominator. If $c \geq 10$ and $n - c \geq 10$, the normal approximation to the binomial distribution gives reasonably accurate confidence limits (Armitage and Berry, 1994, p. 122) based on the estimated standard error (s.e.) of $\hat{p}$:

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Confidence limits for } p = \hat{p} \pm Z_\alpha \times \text{s.e.}(\hat{p})$$

where Z_α is the standard normal deviate for the desired confidence level: $Z_\alpha = 1.96$ for two-sided 95% confidence limits.

If $c < 10$ or $n - c < 10$, confidence limits for p are more accurate if based directly on the binomial distribution and can be obtained by any of the following methods:

1. Using standard statistical software that calculates exact binomial confidence limits.
2. Calculating and summing tail probabilities for the binomial distribution, according to the algorithm described in Rosner, 1995, pp. 176–7.
3. Consulting published statistical tables or figures, as in Rosner, 1995, or Ciba-Geigy, 1982.
4. Using tables of the F-distribution as follows (Armitage and Berry, 1994, p. 121):
   - Set $A = F_{\alpha/2,[2(n-c+1),2c]}$
   - Set $B = 1/F_{\alpha/2,[2(c+1),2(n-c)]}$
   - Calculate $p_{lower} = \frac{c}{c+(n-c+1)\times A}$
   - Calculate $p_{upper} = \frac{c+1}{c+1+(n-c)\times B}$

Confidence limits for proportions estimated from complex probability samples, such as those used in several national health surveys, require special statistical methods beyond the scope of this text—see Levy and Lemeshow, 1991; Korn and Graubard, 1991, 1999.

APPENDIX 4B

Poisson-Based Confidence Limits for Incidence Rate Estimates Based on 100 or Fewer Cases

Table 4–3 provides multipliers that can be used to estimate confidence limits for an incidence rate, based on the Poisson distribution. Select the row that corresponds

Table 4–3. Rate Multipliers to Obtain 95% Poisson Confidence Limits for an Incidence Rate That Is Based on 100 or Fewer Cases

	Multipliers			Multipliers			Multipliers	
COUNT	LOWER	UPPER	COUNT	LOWER	UPPER	COUNT	LOWER	UPPER
1	0.025	5.572	35	0.697	1.391	68	0.777	1.268
2	0.121	3.611	36	0.701	1.384	69	0.778	1.266
3	0.206	2.922	37	0.704	1.378	70	0.780	1.263
4	0.273	2.560	38	0.708	1.373	71	0.781	1.261
5	0.325	2.333	39	0.711	1.367	72	0.782	1.259
6	0.367	2.176	40	0.715	1.362	73	0.784	1.257
7	0.402	2.060	41	0.718	1.357	74	0.785	1.255
8	0.432	1.970	42	0.721	1.352	75	0.787	1.254
9	0.457	1.898	43	0.724	1.347	76	0.788	1.252
10	0.480	1.839	44	0.727	1.342	77	0.789	1.250
11	0.499	1.789	45	0.730	1.338	78	0.790	1.248
12	0.517	1.747	46	0.732	1.334	79	0.792	1.246
13	0.533	1.710	47	0.735	1.330	80	0.793	1.245
14	0.547	1.678	48	0.737	1.326	81	0.794	1.243
15	0.560	1.649	49	0.740	1.322	82	0.795	1.241
16	0.572	1.624	50	0.742	1.318	83	0.796	1.240
17	0.583	1.601	51	0.745	1.315	84	0.798	1.238
18	0.593	1.580	52	0.747	1.311	85	0.799	1.237
19	0.602	1.562	53	0.749	1.308	86	0.800	1.235
20	0.611	1.544	54	0.751	1.305	87	0.801	1.233
21	0.619	1.529	55	0.753	1.302	88	0.802	1.232
22	0.627	1.514	56	0.755	1.299	89	0.803	1.231
23	0.634	1.500	57	0.757	1.296	90	0.804	1.229
24	0.641	1.488	58	0.759	1.293	91	0.805	1.228
25	0.647	1.476	59	0.761	1.290	92	0.806	1.226
26	0.653	1.465	60	0.763	1.287	93	0.807	1.225
27	0.659	1.455	61	0.765	1.285	94	0.808	1.224
28	0.665	1.445	62	0.767	1.282	95	0.809	1.222
29	0.670	1.436	63	0.768	1.279	96	0.810	1.221
30	0.675	1.428	64	0.770	1.277	97	0.811	1.220
31	0.680	1.419	65	0.772	1.275	98	0.812	1.219
32	0.684	1.412	66	0.773	1.272	99	0.813	1.217
33	0.689	1.404	67	0.775	1.270	100	0.814	1.216
34	0.693	1.397						

to the number of cases counted in the numerator of the rate. Then multiply the observed rate by the "lower" multiplier to get the lower 95% confidence limit, then by the "upper" multiplier to get the upper 95% confidence limit. Using Poisson-based confidence limits assumes constant hazard in the base of experience from which the cases arose.

Confidence limits for incidence rate estimates derived from multi-stage samples, such as those used in several national health surveys, require special statistical methods (Levy and Lemeshow, 1991; Korn and Graubard, 1999).

EXERCISES

1. Table 4–4 shows some fictitious data describing the frequency of hepatitis among high school students in a particular school district. Which of the following explanations could be compatible with the time trends seen in these data? (There may be more than one.)
   (a) More aggressive treatment, resulting in earlier and more frequent cures.
   (b) Adoption of a new treatment that, though it diminishes the severity of hepatitis symptoms, suppresses the immune response and thereby prolongs the clinical course of the disease.
   (c) Success of efforts to prevent new cases of hepatitis.
   (d) A shift toward the occurrence of more aggressive disease, leading to earlier and more frequent deaths among afflicted students.

Table 4–4. Hypothetical Data Showing Incidence and Prevalence of Hepatitis by Year in a Certain School District

YEAR	INCIDENCE[a]	PREVALENCE[b]
1985	24.5	41.8
1986	24.9	41.2
1987	23.8	40.9
1988	24.6	40.1
1989	24.1	38.4
1990	24.7	37.9
1991	24.2	35.3
1992	23.9	33.2
1993	25.1	29.8
1994	24.5	27.2

[a] Cases per 100,000 person-years
[b] Cases per 100,000 persons

Table 4–5. Prevalence of Low Birth Weight, by Town and Mother's Race

	% of Babies Weighing <2500 Grams at Birth	
RACE	TOWN 1	TOWN 2
Black	12%	18%
White	6%	9%
All babies	8%	15%

2. In the early years of the AIDS epidemic, it was generally accepted that AIDS was almost always fatal. According to the Centers for Disease Control, the incidence of AIDS in the U.S. in 1990 was 17.2 cases per 100,000 person-years, yet the mortality rate in that year was only 12.4 deaths per 100,000 person-years. Can you reconcile these apparently conflicting data?
3. The local health department where you work has received funding to set up one new prenatal clinic in some needy area of the county. You and your colleagues decide that the birth prevalence of low birth weight will be used as the primary indicator of need. You are helping the department decide whether to put the clinic in Allenville or Bakertown. Allenville is predominantly African American, while Bakertown is predominantly white, and neither community contains any significant number of residents of other races. At your request, a data technician has compiled some statistics from birth certificate data for babies born in each town over the last two years. Unfortunately, in his haste, he forgot to write down which town was which. He shows you the results in Table 4–5.

   He apologizes and is about to set off to re-do his analysis and identify the towns. Instead, you ponder the data carefully, then thank him for giving you all the information you need to determine that the needier community is Allenville. How did you reach that conclusion?
4. You are a hospital epidemiologist working with intensive-care specialists to evaluate a new type of indwelling urinary catheter. The clinical team needs to know how the cumulative incidence of urinary tract infection (UTI) increases in relation to how long the catheter has been in place.

   During the month of September, 10 patients received the new catheter. Daily urine cultures were done on all patients. All patients were monitored until they developed a UTI, no longer needed an indwelling urinary catheter, were discharged from the intensive-care unit, or died, whichever came first. Their experience is summarized in Table 4–6. Based on these early data, what is your best estimate of the one-week cumulative incidence of UTI among patients who receive the new catheter?

Table 4–6. Use of New Urinary Catheter Among Ten Patients on an Intensive Care Unit

PATIENT NO.	CATHETER INSERTED ON	LAST URINE CULTURE ON	REASON FOR ENDING FOLLOW-UP
1	9/1	9/8	Discharged
2	9/2	9/5	Developed UTI
3	9/3	9/9	Died
4	9/5	9/8	Discharged
5	9/5	9/7	Developed UTI
6	9/8	9/13	Catheter no longer needed
7	9/9	9/13	Developed UTI
8	9/11	9/19	Died
9	9/13	9/18	Catheter no longer needed
10	9/14	9/20	Developed UTI

ANSWERS

1. Inspection of the data shows that the prevalence of hepatitis was declining, while its incidence was stable. The approximate relation $P \approx ID \cdot \bar{d}$ tells us that the decline in prevalence could be accounted for by a decline in disease duration, however. Explanations *(a)* and *(d)* are compatible with this assertion. Explanation *(b)* is not, since it implies an increase in duration of disease, not a decrease. Explanation *(c)* suggests that incidence should have been dropping, which it was not.
2. In order for the approximate relation (mortality) $\approx$ (incidence) $\times$ (case fatality) to hold true numerically, incidence and case-fatality rates must be stable over a period of time so that a steady-state situation can develop. This requirement is clearly not met for AIDS: in 1990, its incidence was still rising sharply. Moreover, although AIDS was usually fatal at that time, death did not occur immediately after diagnosis but occurred months or years later. Deaths occurring in 1990 might thus have consisted primarily of patients diagnosed in, say, 1988. The rise in incidence over that period tells us that there were fewer new AIDS patients in 1988 than in 1990.
3. You knew that the overall rate (here, the overall prevalence of weighing under 2500 grams at birth) in a population is always a weighted average of subgroup-specific rates within that population, and that the weights are the proportion of the population in each subgroup. Because most of Allenville's pregnant mothers were African American, the prevalence for "All babies" in Allenville must lie closer to the prevalence for African American mothers than to the prevalence for white mothers. For Town no. 1, the overall prevalence of 8% is closer to the 6% for whites than it is to the 12% for African Americans, so Town no. 1 cannot

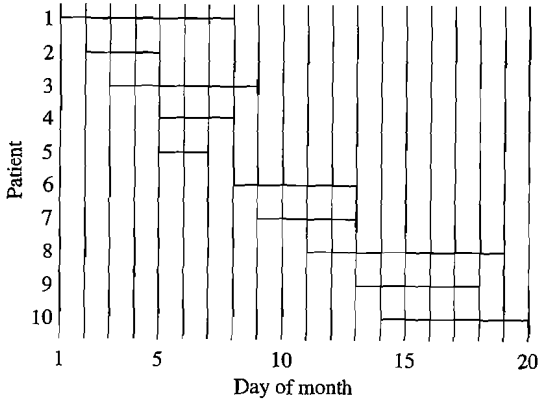

Figure 4–6. Catheter Use by Day of Month in Patients on an Intensive Care Unit.

be Allenville; it must be Bakertown. For Town no. 2, the overall prevalence of 15% is closer to the 18% prevalence for African Americans than to the 9% prevalence in whites, which makes sense if it is, in fact, Allenville.

Having identified which town is which, it is easy to decide which is needier. The race-specific and overall prevalences of low birth weight are all higher in Allenville, so it gets the clinic.

4. Four cases of UTI occurred among 10 patients. But most patients were under observation for less than a full week. Discharge, death, and discontinuation of the catheter were all forms of *censoring*. This is a job for the Kaplan-Meier method. Diagrammatically, the experience of these 10 patients was as shown in Figure 4–6.

The day of the month on which each patient entered and left the study are not really relevant, however. Instead, we are interested in the cumulative incidence of UTI in relation to *time since catheter insertion*. We can change the time scale by aligning the leftmost end of each patient's line along the vertical axis in a new plot. Arithmetically, this is done by just calculating how many days transpired between catheter insertion and the last urine culture for each patient and letting this be the length of that patient's line in the new figure, shown in Figure 4–7.

From these data, we can estimate the proportion "surviving" (*not* having developed a UTI) on each day since catheter insertion, as shown in Table 4–7. The estimated seven-day cumulative incidence is $(1 - .514) = .486$. In other words, our best estimate is that 48.6% of patients with the new catheter develop a UTI within seven days after its insertion.

Table 4–7. Proportion of Patients Remaining Free of Urinary Tract Infection, by Days Since Catheter Insertion

DAYS SINCE INSERTION	NO. STILL UNDER OBSERVATION	NO. OF CASES	PROPORTION "SURVIVING"
0	10	0	1.000
1	10	0	1.000
2	10	1	$1.000 \times 9/10 = .900$
3	9	1	$.900 \times 8/9 = .800$
4	7	1	$.800 \times 6/7 = .686$
5	6	0	.686
6	4	1	$.686 \times 3/4 = .514$
7	2	0	.514

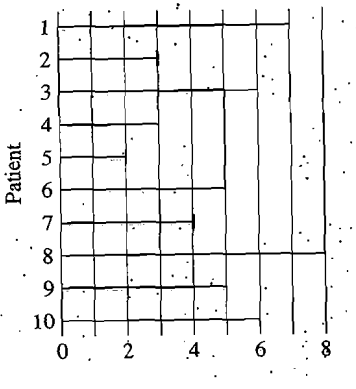

Figure 4–7. Duration of Catheter Use by Time Since Insertion in Patients on an Intensive Care Unit.

REFERENCES

Aalen OO. Heterogeneity in survival analysis. Stat Med 1988; 7:1121–37.

Aalen OO. Effects of frailty in survival analysis. Stat Methods Med Res 1994; 3:227–43.

Armitage P, Berry G. Statistical methods in medical research (3rd edition). London: Blackwell, 1994.

Breslow NE, Day NE. Statistical methods in cancer research. Vol. II—The design and analysis of cohort studies. Lyon, France: International Agency for Research on Cancer, 1987.

Checkoway H, Pearce NE, Crawford-Brown DJ. Research methods in occupational epidemiology. New York: Oxford, 1989.

Ciba-Geigy. Geigy scientific tables. Vol. 2. Introduction to statistics. Statistical tables. Mathematical formulae (8th ed.). Basel, Switzerland: Ciba-Geigy, 1982.

Clayton D. Some approaches to the analysis of recurrent event data. Stat Methods Med Res 1994; 3:244–62.
Cox DR. Regression models and life tables (with discussion). J R Stat Soc B 1972; 34:187–220.
Cumming RG, Kelsey JL, Nevitt MC. Methodologic issues in the study of frequent and recurrent health problems. Falls in the elderly. Ann Epidemiol 1990; 1:49–56.
Dowd MD, Langley J, Koepsell T, Soderberg R, Rivara FP. Hospitalizations for injury in New Zealand: prior injury as a risk factor for assaultive injury. Am J Public Health 1996; 86:929–34.
Gardner LI, Landsittel DP, Nelson NA. Risk factors for back injury in 31,076 retail merchandise store workers. Am J Epidemiol 1999; 150:825–33.
Glynn RJ, Buring JE. Ways of measuring rates of recurrent events. BMJ 1996; 312:364–67.
Glynn RJ, Stukel TA, Sharp SM, Bubolz TA, Freeman JL, Fisher ES. Estimating the variance of standardized rates of recurrent events, with application to hospitalizations among the elderly in New England. Am J Epidemiol 1993; 137:776–86.
Hosmer DW Jr, Lemeshow S. Applied survival analysis: regression modeling of time to event data. New York: Wiley and Sons, 1999.
Johnston BD, Grossman DC, Connell FA, Koepsell TD. High-risk periods for childhood injury among siblings. Pediatrics 2000; 105:562–68.
Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley and Sons, 1980.
Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53:457–81.
Kleinbaum DG. Survival analysis: a self-learning text. New York: Springer, 1996.
Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. Am J Public Health 1991; 81:1166–73.
Korn EL, Graubard BI. Analysis of health surveys. New York: Wiley, 1999.
Levy PS, Lemeshow S. Sampling of populations: methods and applications. New York: Wiley and Sons, 1991.
Lindsay R, Silverman SL, Cooper C, Hanley DA, Barton I, Broy SB, et al. Risk of new vertebral fracture in the year following a fracture. JAMA 2001; 285:320–23.
Mendenhall W, Scheaffer RL, Wackerly DD. Mathematical statistics with applications (3rd edition). Boston: Duxbury Press, 1986.
Morris JN, Heady JA, Raffle PAB, Roberts CG, Parks JW. Coronary heart-disease and physical activity of work. Lancet 1953; 2:1053–57.
Morrison AS. Screening in chronic disease (2nd edition). New York: Oxford, 1992.
Nevitt MP, Ballard DJ, Hallett JW Jr. Prognosis of abdominal aortic aneurysms: a population-based study. N Engl J Med 1989; 321:1009–14.
Rosner B. Fundamentals of biostatistics (4th edition). New York: Duxbury Press, 1995.
Stukel TA, Glynn RJ, Fisher ES, Sharp SM, Lu-Yao G, Wennberg JE. Standardized rates of recurrent outcomes. Stat Med 1994; 13:1781–91.
Sturmer T, Glynn RJ, Kliebsch U, Brenner H. Analytic strategies for recurrent events in epidemiologic studies: background and application to hospitalization risk in the elderly. J Clin Epidemiol 2000; 53:57–64.
Wang HX, Fratiglioni L, Frisoni GB, Viitanen M, Winblad B. Smoking and the occurrence of Alzheimer's disease: cross-sectional and longitudinal data in a population-based study. Am J Epidemiol 1999; 149:640–44.

5

OVERVIEW OF STUDY DESIGNS

> We're all of us guinea pigs in the laboratory of God.
>
> Tennessee Williams

An epidemiologic study generally begins with a question. Once the research question has been specified, the next step in trying to answer it is to choose a study design.

A study design is a plan for selecting study subjects and for obtaining data about them. Study subjects in epidemiology are typically individual people, but at times they can be other kinds of observation units, such as social groups, places, time periods, or even published articles. Information on study subjects can come from pre-existing sources or can be gathered anew by various methods, including direct observation, interviews, examinations, or physiological measurements.

In principle, the number of possible study designs is infinite. But in practice, a few standard designs account for most epidemiologic research. Collectively, these standard designs offer enough flexibility to address a wide range of research questions. Knowledge of their pros and cons can usually guide the investigator to a study design that is well matched to a particular research question. This chapter seeks to provide a broad overview by introducing several standard designs and the terms that are commonly used to describe them and distinguish them from each other. Later chapters cover specific designs in more depth.

DESIGN TREE

Just as there are many possible study designs, there are many possible ways to classify them, depending on which features are highlighted. Figure 5–1 is a tree diagram that organizes designs according to important distinguishing features. Major branches of this tree include:

- *Descriptive* studies are undertaken without a specific hypothesis. They are often among the earliest studies done on a new disease, in order to