PART 2

Measures of Disease Occurrence and Association

CHAPTER 2

Measuring Disease Occurrence

2.1 Introduction

The outcomes of epidemiologic research have been traditionally defined in terms of disease, although the growing application of epidemiology to public health and preventive medicine increasingly requires the use of outcomes measuring health in general (e.g., outcome measures of functional status in epidemiologic studies related to aging). Outcomes can be expressed as either discrete (e.g., disease occurrence or severity) or continuous variables.

Continuous variables, such as blood pressure and glucose levels, are commonly used as outcomes in epidemiology. The main statistical tools used to analyze correlates or predictors of these types of outcomes are the correlation coefficients, analysis of variance, and linear regression analysis, which are discussed in numerous statistical textbooks. Linear regression is briefly reviewed in Chapter 7, Section 7.4.1, as a background for the introduction to multivariate regression analysis techniques in epidemiology. Other methodological issues regarding the analysis of continuous variables in epidemiology, specifically as they relate to quality control and reliability measures, are covered in Chapter 8.

Most of this chapter deals with *categorical* dichotomous outcome variables, which are the most often used in epidemiologic studies. The frequency of this type of outcome can be generically defined as the number of individuals with the outcome (the numerator) divided by the number of individuals at risk for that outcome (the denominator). Depending on the time frame of reference, there are two types of absolute measures of outcome frequency: incidence and prevalence (**TABLE 2-1**).

The term *incidence* has been traditionally used to indicate a proportion of newly developed (incident) cases of a disease over a specific time period. However, this measure can be used to characterize the frequency of any new health- or disease-related event, including death, recurrent disease among patients, disease remission, menopause, and so forth. Incidence is a particularly important measure for analytical epidemiologic research, as it allows the estimation of risk necessary to assess causal associations (Chapter 10, Section 10.2.4).

TABLE 2-1 Absolute measures of disease frequency.

Measure	Expresses	Types of events
Incidence	Frequency of a *new* event over time	Newly developed disease Death in the total population at risk (mortality) Death in patients (case fatality) Recurrence of a disease Development of a side effect of a drug Disease remission
Prevalence	Frequency of an *existing* event	*Point* prevalence: cases at a given point in time *Period* prevalence: cases during a given period (e.g., 1 year) *Cumulative* (lifetime) prevalence: cases at any time in the past (up to the present time)

Prevalence, on the other hand, measures the frequency of an existing outcome either at one point in time (point prevalence) or during a given period (period prevalence). A special type of period prevalence is lifetime prevalence, which measures the cumulative lifetime frequency of an outcome to the present time (i.e., the proportion of people who have had the event at any time in the past).

For both prevalence and incidence, it is necessary to have a clear definition of the outcome as an *event* (a "noteworthy happening," as defined in an English dictionary[1]). In epidemiology, an event is typically defined as the occurrence of any disease or health phenomenon that can be discretely characterized. For incidence (see Section 2.2), this characterization needs to include a precise definition of the time of occurrence of the event in question. Some events are easily defined and located in time, such as "birth," "death," "surgery," and "trauma." Others are not easily defined and require a relatively arbitrary operational definition for study, such as "menopause," "recovery," "dementia," or cytomegalovirus (CMV) disease (**TABLE 2-2**). An example of the complexity of defining certain clinical events is given by the widely adopted definition of a case of AIDS, which uses a number of clinical and laboratory criteria.[2]

The next two sections of this chapter describe the different alternatives for the calculation of incidence and prevalence. The last section describes the *odds*, another measure of disease frequency that is the basis for a measure of association often used in epidemiology, particularly in case-based case-control studies (Chapter 1, Section 1.4.2), namely, the odds ratio (Chapter 3, Section 3.4.1).

2.2 Measures of Incidence

Incidence is best understood in the context of prospective (cohort) studies (Chapter 1, Section 1.4.1). The basic structure of any incidence indicator is represented by the number of events occurring in a defined population over a specified period of time (numerator) divided by the population at risk for that event over that time (denominator). There are two types of measures of incidence defined by the type of denominator: (1) incidence based on persons at risk and (2) incidence based on person-time units at risk.

TABLE 2-2 Examples of operational definitions of events in epidemiologic studies.

Event	Definition	Reference
Natural menopause	Date of last menstrual period after a woman has stopped menstruating for 12 months	Bromberger et al. 1997*
Remission of diarrhea	At least 2 days free of diarrhea (diarrhea = passage of ≥ 3 liquid or semisolid stools in a day)	Mirza et al. 1997†
Dementia	A hospital discharge, institutionalization, or admission to a day-care center in a nursing home or psychiatric hospital with a diagnosis of dementia (ICD-9-CM codes 290.0–290.4, 294.0, 294.1, 331.0–331.2)	Breteler et al. 1995‡
CMV disease	Evidence of CMV infection (CMV antigen on white blood cells, CMV culture, or seroconversion) accompanied by otherwise unexplained spiking fever over 48 hours and either malaise or a fall in neutrophil count over 3 consecutive days.	Gane et al. 1997§

*Data from Bromberger JT, Matthews KA, Kuller LH, Wing RR, Meilahn EN, Plantinga P. Prospective study of the determinants of age at menopause. *Am J Epidemiol.* 1997;145:124-133.

†Data from Mirza NM, Caulfield LE, Black RE, Macharia WM. Risk factors for diarrheal duration. *Am J Epidemiol.* 1997;146:776-785.

‡Data from Breteler MM, de Groot RR, van Romunde LK, Hofman A. Risk of dementia in patients with Parkinson's disease, epilepsy, and severe head trauma: a register-based follow-up study. *Am J Epidemiol.* 1995;142:1300-1305.

§Data from Gane E, Saliba F, Valdecasas JC, et al. Randomised trial of efficacy and safety or oral ganciclovir in the prevention of cytomegalovirus disease in liver-transplant recipients. *Lancet.* 1997;350:1729-1733.

2.2.1 Incidence Based on Individuals at Risk

This is an index defined in terms of the probability of the event, also known as *cumulative incidence* (or *incidence proportion*[3]), which is the basis for the statistical techniques collectively known as *survival analysis*.

If follow-up is complete on every individual in the cohort, the estimation of the cumulative incidence is simply the number of events occurring during the follow-up time divided by the initial population. In epidemiologic studies, however, the follow-up is almost always incomplete for many individuals in the study. In a typical cohort study, there are individuals lost to follow-up, those dying from causes other than the outcome of interest, and those whose follow-up is shorter because they are recruited later in the accrual period for the study. All these losses to follow-up are called *censored observations*, and they require special analytical approaches.*

The traditional techniques for the estimation of cumulative incidence (or its complement, *cumulative survival* or *survival function*) in the presence of censored observations are the life table of the actuarial type (interchangeably referred to in this chapter as the *classic*, *actuarial*, or *interval-based life table*) and the Kaplan–Meier method.[4]

*Duration of follow-up may be regarded as a confounder when it differs between the groups under comparison, as it is usually related to the outcome (e.g., death) (see Chapter 5, Section 5.2).

As an example, **FIGURE 2-1** provides a schematic representation of a study for which the outcome of interest is death, in which 10 individuals are followed for up to 2 years (2015–2016). Each horizontal line in the figure represents the follow-up time of a single individual. Follow-up can be terminated either by the event (D) or by a loss (withdrawal) from the study, also referred to as censored observation (denoted in the figure as an arrow ending at the time when follow-up ended). Individuals are recruited at different points in time and also leave the study (because of either death or censoring) at different times. For example, individual 1 is recruited in November 2015 and dies in December 2015 after only 1 month of follow-up, and individual 5 lives throughout the entire follow-up period (2 years). **FIGURE 2-2** shows a reorganization of the data from Figure 2-1, where the time scale has been changed to reflect follow-up time rather than calendar time. Thus, time 0 now represents the beginning of the follow-up for each individual (regardless of the actual date of the start of follow-up). Much of the discussion of incidence indexes that follows is based on Figure 2-2.

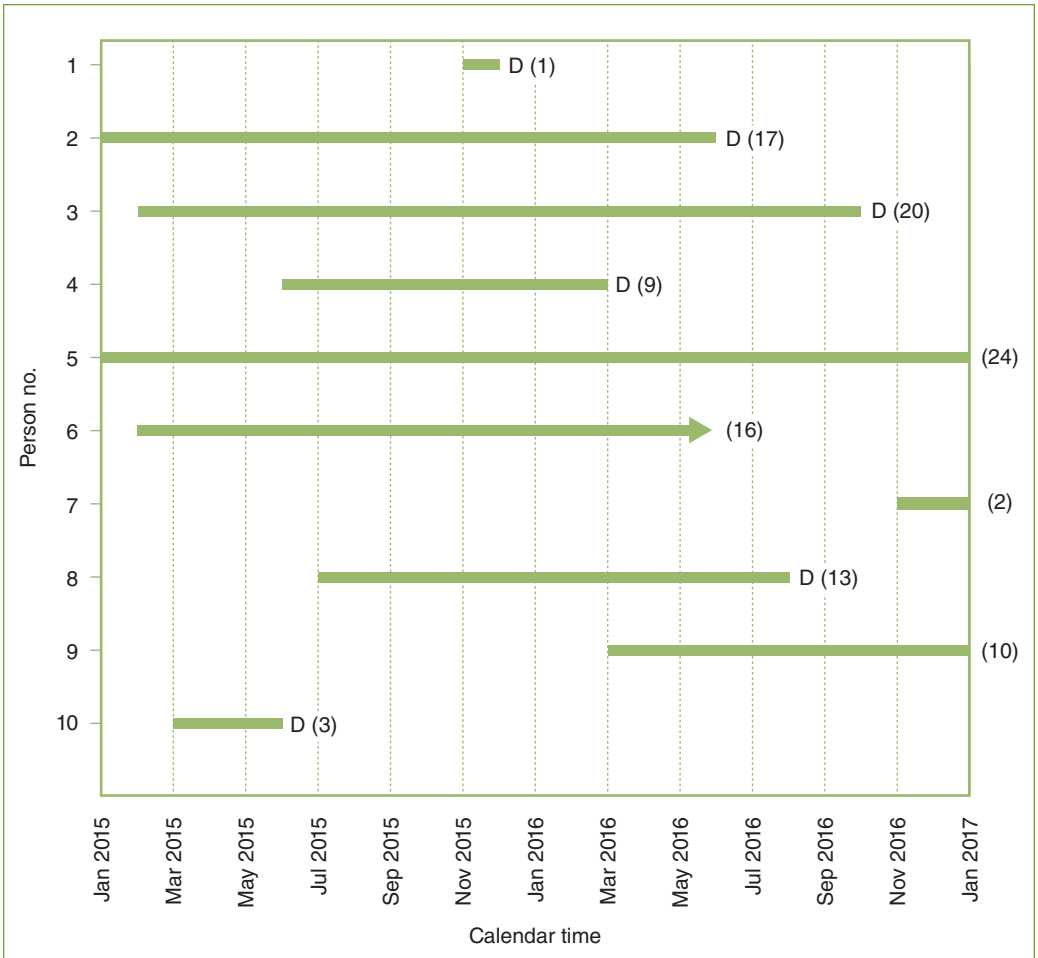

FIGURE 2-1 Hypothetical cohort of 10 persons followed for up to 24 months from January 2015 through December 2016. D, death; arrow, censored observation; ( ), duration of follow-up in months (all assumed to be exact whole numbers).

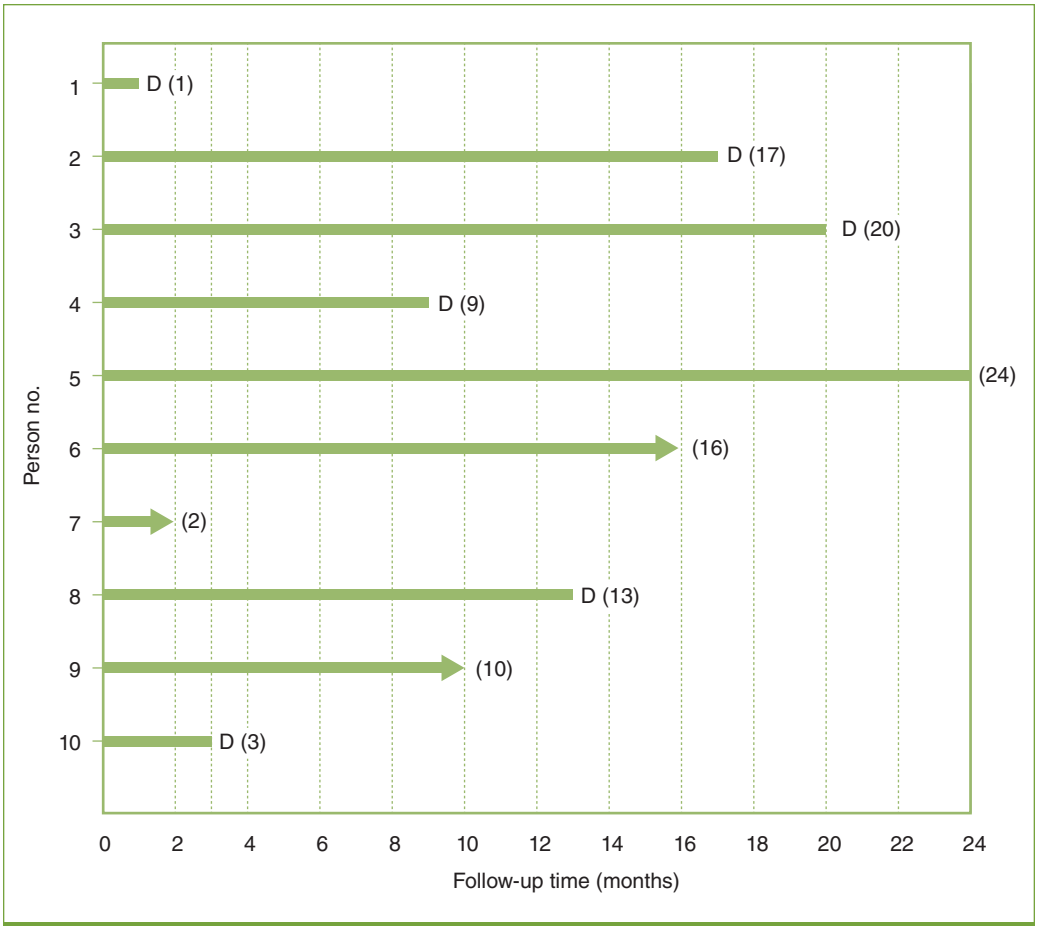

FIGURE 2-2 Same cohort as in Figure 2-1, with person-time represented according to time since the beginning of the study. D, death; arrow, censored observation; ( ), duration of follow-up in months (all assumed to be exact whole numbers).

Cumulative Incidence Based on the Life-Table Interval Approach (Actuarial Life Table)

The cumulative probability of the event during a given interval (lasting m units of time and beginning at time x) is the proportion of new events during that period of time (with events noted as ${}_m d_x$) in which the denominator is the initial population (l_x) corrected for losses (${}_m c_x$). In the classic life table, this measure corresponds to the interval-based probability of the event ${}_m q_x$.[5] Its calculation is straightforward. As seen in Figure 2-2, six deaths occurred among the 10 individuals who were alive at the beginning of the follow-up. If no individual had been lost to observation, ${}_2 q_0$ (with time specified in years) would simply be the number of deaths over this 2-year interval (${}_2 d_0$) divided by the number of individuals at the beginning of the interval (l_0); that is, $6 \div 10 = 0.60$, or 60%. Because the three individuals lost to observation (censored, ${}_2 c_0$) were not at risk during the entire duration of the follow-up, however, their limited participation must be accounted for in the

denominator of the cumulative probability. By convention, half of these individuals are subtracted from the denominator, and the probability estimate is then calculated as follows:

$$ {}_2q_0 = \frac{{}_2d_0}{l_0 - 0.5 \times {}_2c_0} = \frac{6}{10 - 0.5 \times 3} = 0.71 \qquad \text{(Eq. 2.1)} $$

The conventional approach of subtracting one-half of the total number of censored observations from the denominator is based on the assumption that censoring occurred uniformly throughout that period and thus, on average, these individuals were at risk for only one-half of the follow-up period.

The complement of this cumulative probability of the event (q) is the cumulative probability of survival (p); that is,

$$ {}_2p_0 = 1 - {}_2q_0 = 0.29 $$

It is important to note that the cumulative probability of an event (or the cumulative survival) has no time period intrinsically attached to it: Time must be specified. Thus, in this example, one has to describe q as the "2-year cumulative probability of death."

Usually, the classic life table uses multiple intervals, for example, five intervals of 2 years for a total follow-up of 10 years. Within each interval, the probability of survival is calculated using as the denominator the number of individuals under observation at the beginning of the interval corrected for losses during the interval as described previously. To be part of the denominator for the calculation of the survival probability in the second interval (${}_2p_2$), for example, one has to survive through the first interval; likewise, the survival probability for the third interval (starting at year 4, or ${}_2p_4$) is calculated among only those who survived both the first and second time intervals. This is the reason these interval-specific probabilities are technically called "conditional probabilities"; that is, survival in a given interval is conditional upon survival in the previous intervals. A cumulative probability of survival over more than one interval—for example, the full 10-year follow-up with five 2-year intervals—is obtained by multiplying the conditional survival probabilities over all the intervals:

$$ {}_{10}p_0 = {}_2p_0 \times {}_2p_2 \times {}_2p_4 \times {}_2p_6 \times {}_2p_8 $$

The cumulative probability of having the event is the complement of this joint probability of survival:

$$ {}_{10}q_0 = 1 - {}_{10}p_0 = 1 - ({}_2p_0 \times {}_2p_2 \times {}_2p_4 \times {}_2p_6 \times {}_2p_8) \qquad \text{(Eq. 2.2)} $$

This is analogous to the calculation of the cumulative survival function using the Kaplan–Meier method illustrated in the following section.

It is not necessary that the intervals in a classic (interval-based) life table be of the same duration. The length of the interval should be determined by the pace at which incidence changes over time so that, within any given interval, events and withdrawals occur at an approximately uniform rate (discussed later). For example, to study survival after an acute myocardial infarction, the intervals should be very short soon after onset of symptoms when the probability of death is high and rapidly changing. Subsequent intervals could be longer, however, as the probability of a recurrent event and death tends to stabilize. If the sample size allows it and the pace of the event change is not known, it is reasonable to use shorter intervals when first analyzing survival; subsequently, intervals can be changed according to the shape of the survival curve.

Examples of the use of the actuarial life-table method can be found in reports from classic epidemiologic studies (e.g., Pooling Project Research Group[6]). More details and additional examples can be found in other epidemiology textbooks (e.g., Gordis[7] and Kahn and Sempos[8]).

Cumulative Incidence Based on the Kaplan–Meier (Exact Event Times) Approach

The Kaplan–Meier approach involves the calculation of the probability of each event at the time it occurs. The denominator for this calculation is the population at risk at the time of each event's occurrence.[4] As for the actuarial life table, the probability of each event is a "conditional probability"; in other words, it is conditioned on being at risk (alive and not censored) at the event time. If each event (first, second, etc.) is designated by its time of occurrence i, then the formula for the conditional probability is simply as follows:

$$q_i = \frac{d_i}{n_i}$$

where d_i is the number of deaths (or other type of event) occurring at time i, and n_i is the number of individuals still under observation (i.e., at risk of the event) at time i. (Usually, $d_i = 1$ unless more than one event is occurring simultaneously—something that will occur only when nonexact discrete measures of time are used.)

To facilitate the calculations, **FIGURE 2-3** shows the same data as in Figures 2-1 and 2-2 but with the individuals' follow-up times arranged from shortest to longest. When the first death occurs exactly at the end of the first month (person 1), there are 10 individuals at risk; the conditional probability is then as follows:

$$q_1 = \frac{1}{10}$$

When the second death occurs after 3 months of follow-up (person 10), there are only eight persons at risk; this is because in addition to the one previous death (D), one individual had been lost to observation after 2 months (person 7) and therefore was not at risk when the second death occurred. Thus, the conditional probability at the time of the second death is estimated as follows:

$$q_3 = \frac{1}{8} = 0.125$$

These calculations are repeated for all the event times. For example, for the fifth event, when person 2 died at 17 months of follow-up, there were three individuals still under observation (Figure 2-3), and thus, the conditional probability of the death at month 17 can be estimated as follows:

$$q_{17} = \frac{1}{3} = 0.333$$

TABLE 2-3 (column 4) shows the calculation of these conditional probabilities for each of the six event times in this example. The censored observations are skipped in these calculations, as they do not represent an identified event. Censored observations, however, are included in the denominator for the computation of conditional probabilities corresponding to events occurring up to the time when the censoring occurs. This represents the most efficient use of the available information.[4]

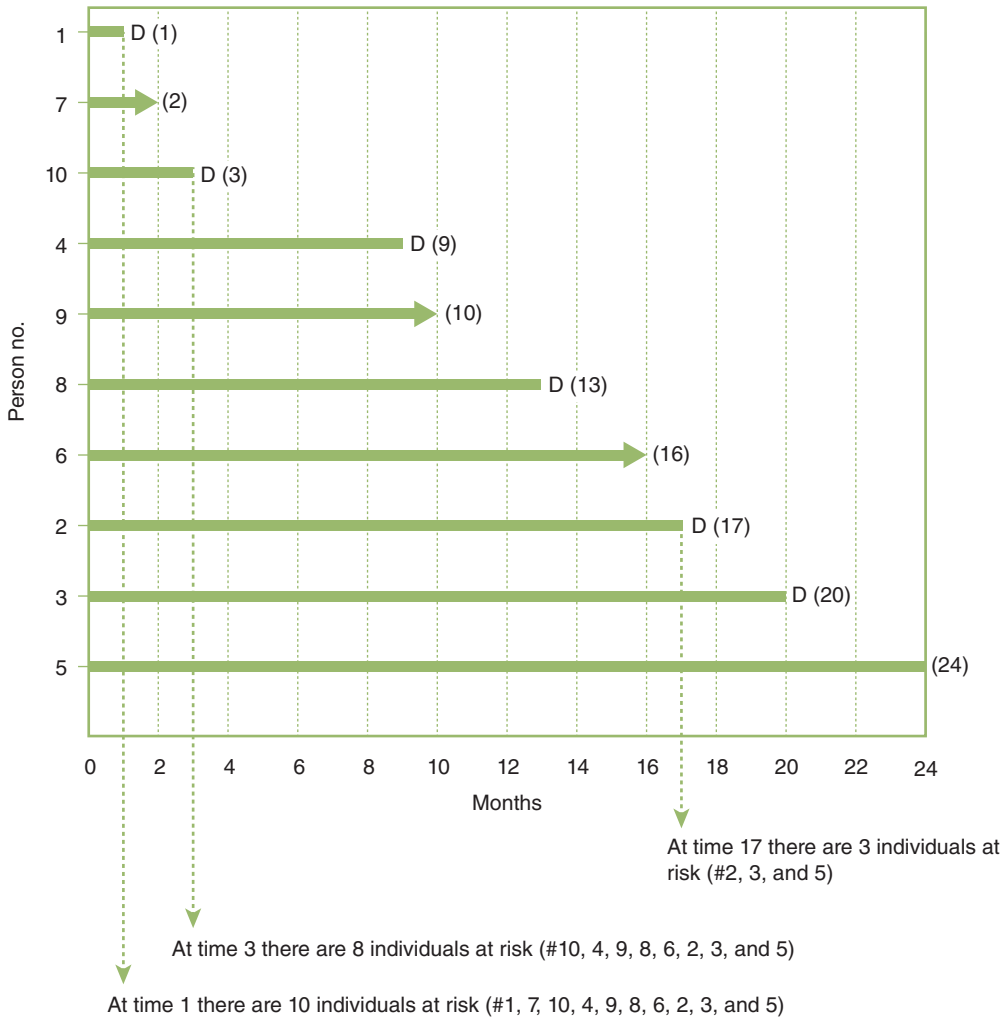

FIGURE 2-3 Same cohort as in Figures 2-1 and 2-2, with individuals sorted according to follow-up time from shortest to longest. D, death; arrow, censored observation; ( ), duration of follow-up in months (all assumed to be exact whole numbers). As examples, the vertical arrows mark the individuals who were at risk for the calculations of the conditional probabilities of death at three of the event times: 1 month, 3 months, and 17 months (see text).

Column 5 in Table 2-3 shows the complements of the conditional probabilities of the event at each time, that is, the conditional probabilities of survival (p_i), which, as in the classic life-table method, represent the probability of surviving beyond time i among those who were still under observation at that time (i.e., conditioned on having survived up to time i). Column 6, also shown graphically in **FIGURE 2-4**, presents the cumulative probabilities of survival, that is, the so-called Kaplan–Meier survival function (usually notated as S_i). This represents the probability of surviving beyond time

TABLE 2-3 Calculation of Kaplan–Meier survival estimates for the example in Figure 2-3.

Time (months) (1) i	Number of individuals at risk (2) n_i	Number of events (3) d_i	Conditional probability of the event (4) $q_i = d_i/n_i$	Conditional probability of survival (5) $p_i = 1 - q_i$	Cumulative probability of survival* (6) S_i
1	10†	1	1/10 = 0.100	9/10 = 0.900	0.900
3	8†	1	1/8 = 0.125	7/8 = 0.875	0.788
9	7	1	1/7 = 0.143	6/7 = 0.857	0.675
13	5	1	1/5 = 0.200	4/5 = 0.800	0.540
17	3†	1	1/3 = 0.333	2/3 = 0.667	0.360
20	2	1	1/2 = 0.500	1/2 = 0.500	0.180

*Obtained by multiplying the conditional probabilities in column (5)—see text.
†Examples of how to determine how many individuals were at risk at three of the event times (1, 3, and 17 months) are shown with vertical arrows in Figure 2-3.

i for all those present at the beginning of follow-up, calculated as the product of all conditional survival probabilities up to time i. In the example, the cumulative probability of surviving beyond the end of the follow-up period of 2 years (S_i, where $i = 24$ months) is as follows:

$$S_{24} = \frac{9}{10} \times \frac{7}{8} \times \frac{6}{7} \times \frac{4}{5} \times \frac{2}{3} \times \frac{1}{2} = 0.18$$

Thus, the estimate of the cumulative probability of the event ($1 - S_i$) is as follows:

$$1 - S_{24} = 1 - 0.18 = 0.82$$

As for the cumulative probability based on the actuarial life-table approach (Equation 2.2), the time interval for the cumulative probability using the Kaplan–Meier approach also needs to be specified (in this example, 24 months, or 2 years). For a method to calculate confidence limits for a cumulative survival probability estimate, see Appendix A, Section A.1.

Regardless of the method used in the calculation (actuarial or Kaplan–Meier), the cumulative incidence is a proportion in the strict sense of the term. It is unitless, and its value can range from 0 to 1 (or 100%).

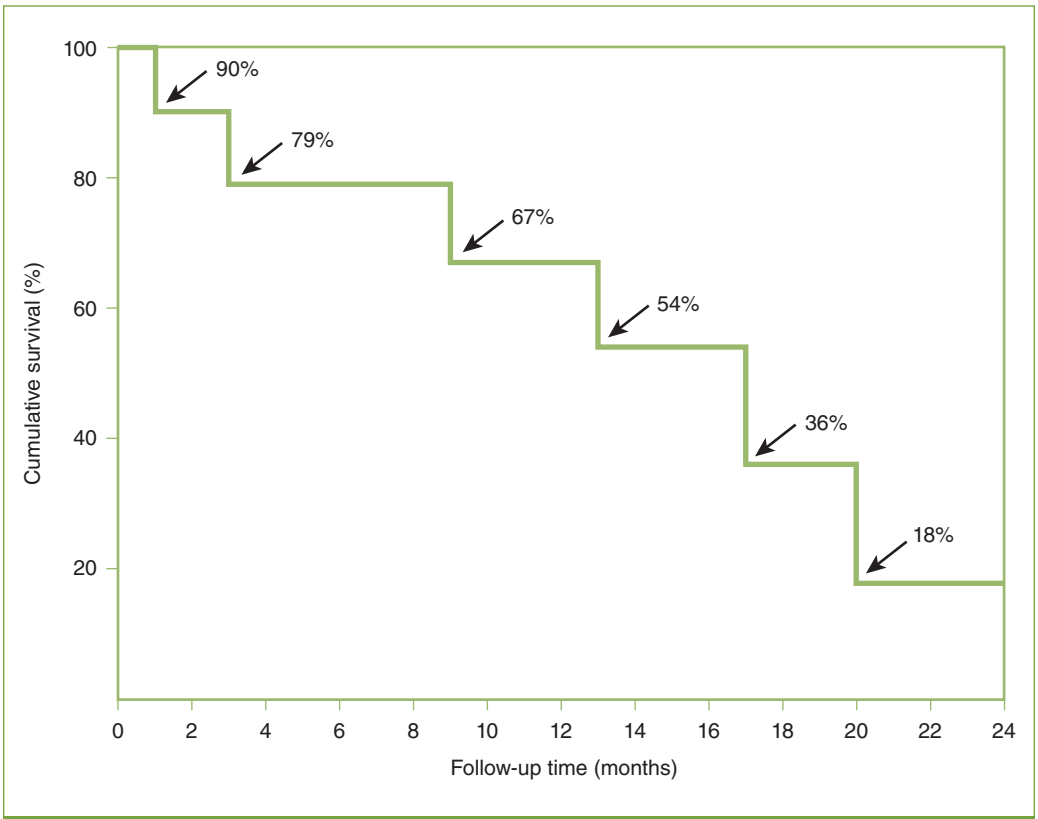

FIGURE 2-4 Kaplan–Meier curve corresponding to the data in Table 2-3, column 6.

Kaplan–Meier curves can be used to *compare* survival in different groups (e.g., exposed or unexposed or different degrees of exposure). These curves provide a visual tool to assess associations between an exposure and the risk of an outcome—complementary to the quantitative measures that are discussed in Chapters 3 and 7. For example, **FIGURE 2-5** shows the survival curves in groups defined according to the presence/severity of sleep among participants in the Wisconsin Sleep Cohort Study.[9] The figure shows that survival was highest among participants without sleep apnea and gradually decreased with increasing severity of the condition—being lowest among those with "severe" sleep apnea.

Instead of the survival curve, investigators might choose to plot its complement, in other words, a curve depicting the cumulative probability of the event ($1 - S_i$) as the follow-up progresses. As an example, **FIGURE 2-6** displays the cumulative incidence of hypertension among individuals in different categories of alcohol consumption among participants in the Coronary Artery Risk Development in Young Adults (CARDIA) Study.[10] This figure is like a mirror image of the survival curves shown in Figures 2-4 and 2-5, with an ordinate scale starting at zero at the beginning of the follow-up—as opposed to starting at 100% as in the typical survival curve. The figure shows how the incidence of hypertension is highest among "former" alcohol drinkers. Note also that in Figure 2-6, all events appear to be occurring at discrete follow-up times, coinciding with the examination times. This is

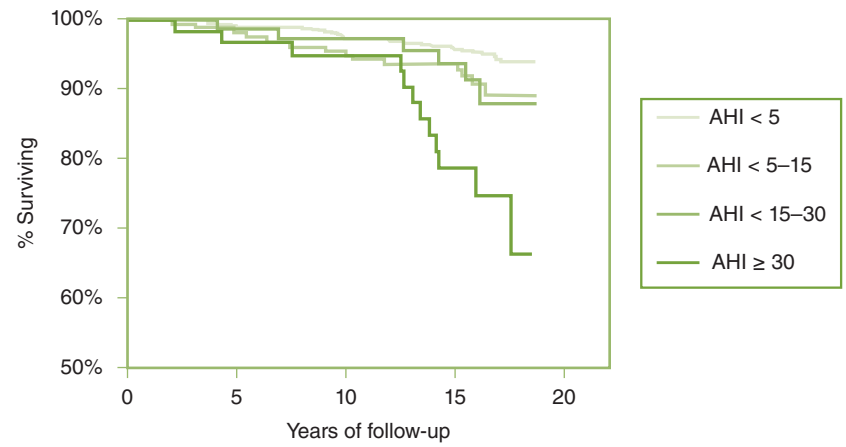

FIGURE 2-5 Kaplan–Meier estimates of survival probability according to sleep apnea severity as defined by the apnea-hypopnea index (AHI) (shown with the *y*-axis truncated at 50% survival). None (AHI $<$ 5), mild (AHI $>$ 5, $<$ 15), moderate (AHI $>$ 15, $<$ 30), and severe (AHI $\geq$ 30), total sample ($n = 1522$); AHI is the mean number of apnea and hypopnea episodes/hour of sleep.

Reprinted with permission from Young T, Finn L, Peppard PE, et al. Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin Sleep Cohort. *Sleep.* 2008;31:1071-1078.[9] By permission of Oxford Press University.

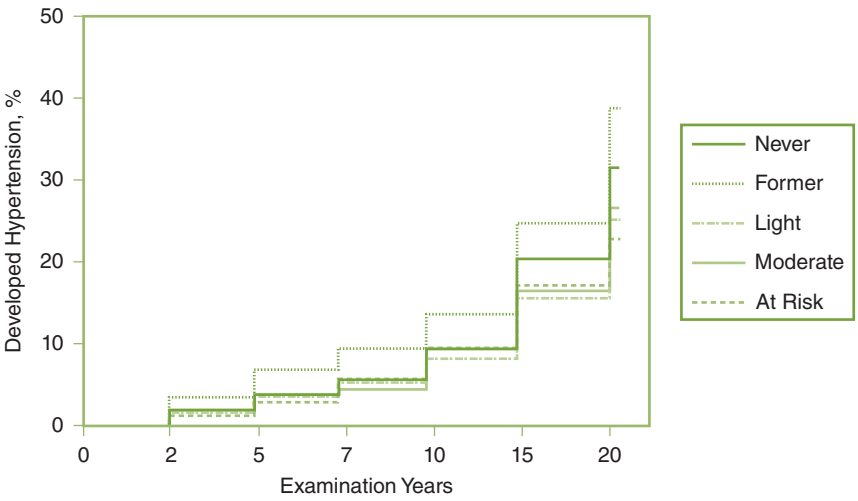

FIGURE 2-6 Kaplan–Meier estimates of time to incident hypertension by drinking category, the Coronary Artery Risk Development in Young Adults (CARDIA) Study, 1985–2006.

Reprinted with permission from Halanych JH, Safford MM, Kertesz SG, et al. Alcohol consumption in young adults and incident hypertension: 20-year follow-up from the Coronary Artery Risk Development in Young Adults Study. *Am J Epidemiol.* 2010;171:532-539.[10] By permission of Oxford University Press.

because "incident hypertension" is determined at the time of each follow-up exam (i.e., participants with high blood pressure or taking anti-hypertensive medications among those who have no evidence of hypertension in previous examinations). The *exact* time when hypertension first started in the interim time between exams (the actual *event date*) is unknown.

Assumptions in the Estimation of Cumulative Incidence Based on Survival Analysis

The following assumptions must be met when conducting survival analysis:

Uniformity of Events and Losses Within Each Interval (Classic Life Table). Implicit in the classic life-table calculation (discussed previously) is the generic assumption that events and losses are approximately uniform during each defined interval. If risk changes rapidly within a given interval, then calculating a cumulative risk over the interval is not very informative. The rationale underlying the method to correct for losses—that is, subtracting one-half of the losses from the denominator (Equation 2.1)—also depends on the assumption that losses occur uniformly. The assumption of uniformity of events and losses within a given interval is entirely related to the way the life table is defined and can be met by adjusting the interval definitions to appropriately uniform risk intervals (e.g., by shortening them). Furthermore, this assumption does not apply to the Kaplan–Meier calculation, where intervals are not defined *a priori*.

Whereas this interval-based assumption applies only to classic life-table estimates, the following two assumptions apply to both classic life-table and Kaplan–Meier estimates and are key to survival analysis techniques and analyses of cohort data in general: (1) independence between censoring and survival and (2) lack of secular trends during the study's accrual period.

Independence Between Censoring and Survival. For the calculations of the conditional and cumulative incidences using the previously described methods, censored individuals are included in the denominator during the entire time they are under observation; after being censored, they are ignored in subsequent calculations. Thus, if one wants to infer that the estimated overall cumulative survival (e.g., $S_{24} = 18\%$, as in Figure 2-4) is generalizable to the entire population present at the study's outset (at time 0), one needs to assume that *the censored observations have the same probability of the event after censoring as those remaining under observation*—a phenomenon also known as "uninformative censoring." In other words, censoring needs to be independent of survival; otherwise, bias will ensue. For example, if the risk were higher for censored than for noncensored observations (e.g., study subjects withdrew because they were sicker), then over time, the study population would include a progressively greater proportion of lower risk subjects; as a result, the (true) overall cumulative incidence would be underestimated (i.e., survival would be overestimated). The opposite bias would occur if censored observations tended to include healthier individuals. The likely direction of the bias according to the reason censoring occurred is summarized in **TABLE 2-4**. With regard to censored observations caused by *death from other causes* in cause-specific outcome studies, if the disease of interest shares strong risk factors with other diseases that are associated with mortality, censoring may not be independent of survival. An example is a study in which the outcome of interest is coronary heart disease death and participants dying from other causes, including respiratory diseases (such as lung cancer and emphysema), are censored at the time of their death (as they are no longer at risk of dying from coronary heart disease). Because coronary heart disease and respiratory diseases share an important risk factor (smoking) and, in addition, respiratory disease deaths are very common in smokers, individuals dying from respiratory diseases may have had a higher risk of coronary heart disease if they had not died from respiratory diseases, resulting in a violation of the assumption of independence between censoring and survival.

TABLE 2-4 Relationship between reason for censoring and the assumption of independence between censoring and survival in survival analysis.

Type of censoring	May violate assumption of independence of censoring/ survival	If assumption is violated, likely direction of bias on the cumulative incidence estimate
Deaths from other causes when there are common risk factors*	Yes	Underestimation
Participants' refusal or inability to allow follow-up contacts	Yes	Underestimation
Migration	Yes	Variable
Administrative censoring	Unlikely†	Variable

*In cause-specific incidence or mortality studies.
†More likely in studies with a prolonged accrual period in the presence of secular trends.

Other frequent reasons for censoring include *refusal* or *inability* of study participants to allow subsequent follow-up contacts (in a study where assessment of the outcome events depends on such contacts) and loss of contact with participants due to *migration* out of the study area. Individuals who refuse or are unable to continue having follow-up contacts may have a less healthy lifestyle than individuals who agree to continuing participation in a prospective study; if that were the case, censoring for this reason may lead to an underestimation of the cumulative incidence. The direction of the bias resulting from the loss to follow-up of individuals because of *migration* is a function of the sociodemographic context in which the migration occurs, for example, whether the individuals who migrate are of higher or lower socioeconomic status (SES). If losses occurred mainly among those in the upper SES, who tend to be healthier, those remaining in the study would tend to have poorer survival. On the other hand, if the losses occurred primarily among individuals with a lower SES, and thus poorer health, the survival of those remaining in the study would be overestimated. Generally, censoring that results in bias is referred to as "informative censoring." For the so-called *administrative* losses, defined as those that occur because the follow-up ends (e.g., persons 7 and 9 in Figure 2-1), the assumption of independence between censoring and survival is regarded as more justified, as these losses are usually thought to be independent of the characteristics of the individuals *per se*. (Administrative losses are, however, amenable to temporal changes occurring during the accrual period; see the following section "Lack of Secular Trends.")

In summary, whether the key *assumption* of independence between censoring and survival for the calculation of cumulative incidence/survival estimates is met depends on the reasons censoring occurred (Table 2-4). This assumption is particularly relevant when the magnitude of the absolute incidence estimate is the focus of the study; it may be less important if the investigator is primarily interested in a relative estimate (e.g., when comparing incidence/survival in two groups defined by

exposure levels in a cohort study), provided that biases resulting from losses are reasonably similar in the groups being compared. (For a discussion of a related bias, the so-called *compensating bias*, see "Selection Bias" in Chapter 4, Section 4.2.) Finally, this assumption can often be verified. For example, it is usually possible to compare baseline characteristics related to the outcome of interest between individuals lost and those not lost to observation. In addition, if relevant study participant identifying information is available, linkage to the National Death Index can be used to compare the mortality experience of those lost and those not lost to follow-up.

Lack of Secular Trends. In studies in which the accrual of study participants occurs over an extended time period, the decision to pool all individuals at time 0 (as in Figure 2-2) assumes a lack of secular trends with regard to the characteristics of these individuals that affect the outcome of interest. This, however, may not be the case in the presence of *birth cohort* and *period* (calendar time) effects (see Chapter 1, Section 1.2). Changes over time in the characteristics of recruited participants as well as significant secular changes in relevant exposures and/or treatments may introduce bias in the cumulative incidence/survival estimates, the direction and magnitude of which depend on the characteristics of these cohort or period effects. Thus, for example, it would not have been appropriate to estimate survival from diagnosis of all patients identified with insulin-dependent diabetes from 1915 through 1935 as a single group, as this extended accrual period would inappropriately combine two very heterogeneous patient cohorts: those diagnosed before and those diagnosed after the introduction of insulin. Similarly, it would not be appropriate to carry out a survival analysis pooling at time 0 all HIV-seropositive individuals recruited into a cohort accrued between 1995 and 1999, that is, both before and after a new effective treatment (protease inhibitors) became available.

2.2.2 Incidence Rate Based on Person-Time

Rather than individuals, the denominator for the incidence *rate* is formed by time units (t) contributed to the follow-up period by the individuals at risk (n). For example, consider a hypothetical cohort in which 12 events occur and the total amount of follow-up time for all individuals is 500 days. The incidence rate in this example is $12 \div 500 = 0.024$ per person-day or 2.4 per 100 person-days. Notice that this rate was calculated even though the actual number of individuals included in this follow-up study was not provided; thus, the "person-time" estimate in the example could have originated from 50 individuals seen during 10 days each (50×10), 5 individuals observed for 100 days (5×100), and so on.

Incidence rates are *not* proportions. They are obtained by dividing the number of events by the amount of time at risk (pooling all study participants) and are measured in units of time^{-1}. As a result, a rate can range from 0 to infinity, depending on the unit of time being used. For example, the previously mentioned incidence rate could be expressed in a number of ways: $12 \div 500$ person-days $= 12 \div 1.37$ person-years $= 8.76$ per person-year (or 876 per 100 person-years). The latter value exceeds 1 (or 100%) only because of the arbitrary choice of the time unit used in the denominator. If a person has the event of interest after a follow-up of 6 months and the investigator chooses to express the rate per person-years, the rate will be $1 \div 0.5$, or 200 per 100 person-years.

The time unit used is at the discretion of the investigator and is usually selected on the basis of the frequency of the event under study. The main reason many epidemiologic studies use person-*years* as the unit of analysis is that it is a convenient way to express rare events. On the other hand, when one is studying relatively frequent health or disease events, it may be more convenient to use some other unit of time (**TABLE 2-5**). The choice is entirely arbitrary and will not affect the inferences derived from the study.

TABLE 2-5 Examples of person-time units according to the frequency of events under investigation.

Population	Event studied	Person-time unit typically used
General	Incident breast cancer	Person-years
General	Incident myocardial infarction	Person-years
Malnourished children	Incident diarrhea	Person-months
Pancreatic cancer cases	Death	Person-months
Influenza epidemic	Incident influenza	Person-weeks
Infants with acute diarrhea	Recovery	Person-days

Rather than a unitless proportion of individuals who develop the event among those at risk (see cumulative incidence described previously in this chapter), incidence based on person-time expresses the "rate" at which the events occur in the population at risk at any given point in time. This type of rate is also called *incidence density*, a concept analogous to that of velocity: the instantaneous rate of change or the "speed" at which individuals develop the event (disease, death, etc.) in the population. This concept is the basis for some of the mathematic modeling techniques used for the analysis of incidence rates (e.g., Poisson regression models) (see Chapter 7, Section 7.4.5). Because the instantaneous rate for each individual cannot be directly calculated, however, the average incidence over a period of time for a population is usually used as a proxy. The average incidence can be calculated based on individual or aggregate follow-up data, as is discussed later in this chapter. Epidemiologists often use the terms *rate* and *density* interchangeably; however, in the discussion that follows, the term *rate* will be primarily used in the context of grouped data, whereas *density* will denote a rate based on data obtained from each individual in the study.

Incidence Rate Based on Aggregate Data

This type of incidence is typically obtained for a geographic location by using the average population estimated for a certain time period as the denominator. Provided that this period is not excessively long and that the population and its demographic composition in the area of interest are relatively stable, the average population can be estimated as the population at the middle of the period (e.g., July 1 for a 1-year period). In a cohort study, the average of the population at the beginning and the end of the period can be obtained for a given follow-up interval. Thus, for a given time interval,

$$\text{Incidence rate} = \frac{\text{Number of events}}{\text{Average population}}$$

In Figure 2-2, for example, 10 individuals are alive and present in the study at the beginning of the follow-up period ("point zero"). Only one person is alive and present in the study when the 2-year follow-up ends (person 5). Thus, the average population (n) for the total 2-year follow-up period is as follows:

$$n = \frac{10 + 1}{2} = 5.5$$

The average population can also be calculated by subtracting one-half of the events (d) and losses (c) from the initial population:

$$n = 10 - \frac{1}{2}(6 + 3) = 5.5$$

As for all mean values, the underlying assumption when using this approach is that, on average, there were 5.5 persons for the duration of the study (2 years). For this assumption to be met, events and withdrawals must occur uniformly throughout the follow-up period. The rate of new events in relationship to the average population is then calculated as follows:

$$\text{Incidence rate} = \frac{6}{5.5} = 1.09 \text{ per person}-2 \text{ years}$$

In this example, the rate is based on a time unit of 2 years and not on the number of individuals. The assumption underlying the use of the average population is that the same rate would have been obtained if 5.5 individuals had been followed for the entire 2-year period, during which six events were observed. This example again highlights the fact that this type of incidence is not a proportion and, thus, is not bound to be 1 (100%) or less. In this instance, the seemingly counterintuitive rate of 109 per 100 person-time obviously resulted from the fact that "2 years" is being used as the time unit; if a "person-year" unit had been used instead, the rate would have been 1.09 ÷ 2 years = 0.545 per person-year (or 54.5 per 100 person-years).

This example illustrates the estimation of the incidence rate using the average population of a defined cohort (i.e., the hypothetical cohort represented in Figure 2-2); however, this is not its usual application. Instead, the calculation of incidence based on grouped data is typically used to estimate mortality based on vital statistics information or incidence of newly diagnosed disease obtained from population-based registries (e.g., cancer registries), in other words, when incidence needs to be estimated for a population or an *aggregate* defined by residence in a given geographic area over some time period. These aggregates are called *open* or *dynamic cohorts* because they include individuals who are added or withdrawn from the pool of the population at risk as they migrate in or out of the area (i.e., a situation more clearly represented by the diagram in Figure 2-1 than that in Figure 2-2).

Incidence Density Based on Individual Data

When relatively precise data on the timing of events or losses are available for each individual from a defined cohort, it is possible to estimate *incidence density*. The total person-time for the study period is simply the sum of the person-time contributed by each individual. The average incidence density is then calculated as follows:

$$\text{Incidence density} = \frac{\text{Number of events}}{\text{Total person-time}}$$

For each individual in the example shown in Figures 2-1, 2-2, and 2-3, the length of the horizontal line represents the length of time between the beginning of the follow-up and the point when the individual either had the event, which in this hypothetical example is death (D), or was lost to observation. For example, for individual 1, death occurred after exactly 1 month. Thus, this individual's contribution to the total number of person-years in the first follow-up year (see Figure 2-3) would be $1 \div 12 = 0.083$; obviously, this person made no contribution to the follow-up during the second year. On the other hand, individual 2 died after remaining in the study for 17 months, or 1 year and 5 months. Thus, his or her contribution to the first follow-up year was $12 \div 12$ and to the second year was $5 \div 12$, for a total of 1.417 person-years.

The contribution of censored individuals is calculated in an identical fashion. For example, the contribution of individual 6 to the total number of person-years was equivalent to 16 months, or 1 full person-year in the first year and $4 \div 12$ person-years in the second year, for a total of 1.333 person-years. The calculation of person-years for all 10 study participants is shown in **TABLE 2-6**. In this example, the incidence density applicable to the total follow-up period is, therefore, $6 \div 9.583 = 0.63$ per person-year (or 63 per 100 person-years). Alternatively, the incidence density could be expressed as $6 \div (9.583 \times 12 \text{ months}) = 0.052$ per person-month (or 5.2 per 100 person-months). For a method to estimate confidence limits of incidence rates, see Appendix A, Section A.2.

Assumptions in the Estimation of Incidence Based on Person-Time

The assumptions of independence between censoring and survival and of lack of secular trends discussed in Section 2.2.1 are also relevant in the context of person-time analysis. The former assumption relates to absence of selection bias resulting from losses to follow-up. Both assumptions apply to any type of cohort study analysis. Furthermore, as for incidence based on the actuarial life table (Equation 2.1), an important assumption when using the person-time approach is that the risk of the event remains approximately constant over time during the interval of interest, or, in other words, the estimated rate should apply equally to *any point in time within the interval*. This means that n persons followed during t units of time are equivalent to t persons observed during n units of time; for example, the risk of an individual living five units of time within the interval is equivalent to that of five individuals living one unit each (**FIGURE 2-7**). When individuals are exposed to a given risk factor, another interpretation of this assumption is that the effect resulting from the exposure is not cumulative within the follow-up interval of interest. Often, this assumption is difficult to accept, as, for example, when doing studies of chronic respiratory disease in smokers: the risk of chronic bronchitis for 1 smoker followed for 30 years is certainly not the same as that of 30 smokers followed for 1 year in view of the strong cumulative effect of smoking and the latency period needed for disease initiation. To decrease the dependency of the person-time approach on this assumption, the follow-up period can be divided into smaller intervals and incidence densities calculated for each interval. For example, using data from Table 2-6 and Figure 2-3, it is possible to calculate densities separately for the first and second years of follow-up as follows:

First follow-up year: $3 \div 7.083 = 42.4$ per 100 person-years
(or $3 \div 85 = 3.5$ per 100 person-months)

Second follow-up year: $3 \div 2.500 = 120$ per 100 person-years
(or $3 \div 30 = 10$ per 100 person-months)

TABLE 2-6 Calculation of the number of person-years based on Figure 2-2.

Person no.	Total follow-up (in months)	Contribution to the total number of person-years by participants in:		
		1st Year of follow-up	2nd Year of follow-up	Total follow-up period
1	1	1/12 = 0.083	0	0.083
2	17	12/12 = 1.000	5/12 = 0.417	1.417
3	20	12/12 = 1.000	8/12 = 0.667	1.667
4	9	9/12 = 0.750	0	0.750
5	24	12/12 = 1.000	12/12 = 1.000	2.000
6	16	12/12 = 1.000	4/12 = 0.333	1.333
7	2	2/12 = 0.167	0	0.167
8	13	12/12 = 1.000	1/12 = 0.083	1.083
9	10	10/12 = 0.833	0	0.833
10	3	3/12 = 0.250	0	0.250
Total	115 months	7.083 years	2.500 years	9.583 years

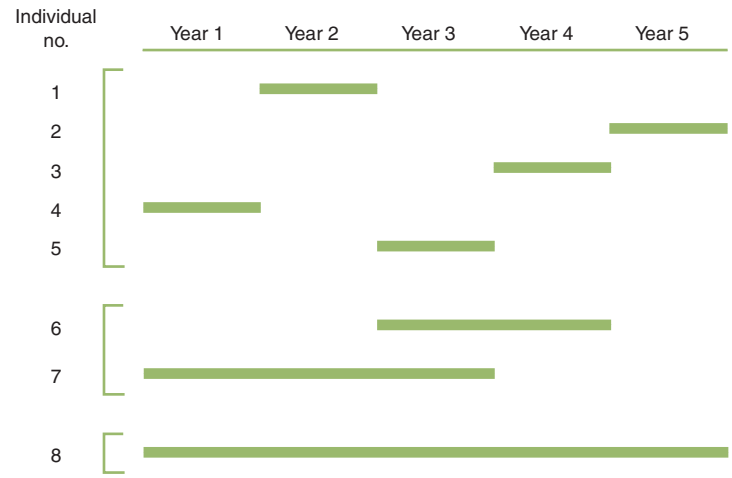

FIGURE 2-7 Follow-up time for eight individuals in a hypothetical study. It is assumed that the sum of the person-time units for individuals nos. 1 to 5 (with a short follow-up time of 1 year each) is equivalent to the sum for individuals nos. 6 and 7 (with follow-up times of 2 and 3 years, respectively) and to the total time for individual no. 8 (who has the longest follow-up time, 5 years). For each group of individuals (nos. 1–5, 6 and 7, and 8), the total number of person-years of observation is 5.

The fact that the densities differ markedly between the first and second follow-up years in this example strongly implies that it would not be reasonable to estimate an incidence density for the overall 2-year period.

Relationship Between Density (Based on Individual Data) and Rate (Based on Grouped Data)

It is of practical interest that when withdrawals (and additions in an open population or dynamic cohort) and events occur uniformly, rate (based on grouped data) and density (based on individual data) are virtually the same. The following equation demonstrates the equivalence between the rate per average population and the density (per person-time), when the former is averaged with regard to the corresponding time unit (e.g., yearly average).

$$\text{Rate} = \frac{\text{No. events } (x)}{\dfrac{\text{average population } (n)}{\text{time } (t)}} = \frac{x}{n \times t} = \text{Density}$$

This idea can be understood intuitively. For a given time unit, such as 1 year, the denominator of the rate (the average population) is analogous to the total number of time units lived by all the individuals in the population in that given time period. An example is given in **TABLE 2-7**, based on data for four persons followed for a maximum of 2 years. One individual is lost to follow-up (censored) after 1 year; two individuals die, one after 0.5 year and the other after 1.5 years; and the

TABLE 2-7 Hypothetical data for four individuals followed for a maximum of 2 years.

Individual no.	Outcome	Timing of event/loss	No. of person-years
1	Death	At 6 months	0.5
2	Loss to observation	At 1 year	1.0
3	Death	At 18 months	1.5
4	Administrative censoring	At 2 years	2.0
		Total no. of person-years:	5.0

fourth individual survives through the end of the study. There is, therefore, perfect symmetry in the distribution of withdrawals or events, which occurred after 0.5, 1, 1.5, and 2 years after the onset of the study. Summing the contribution to the follow-up time made by each participant yields a total of 5 person-years. Density is thus two deaths per 5 person-years, or 0.40.

The average population (n) in this example can be estimated as [(initial population + final population) ÷ 2], or [(4 + 1) ÷ 2 = 2.5]. The rate for the total time ($t = 2$ years) is then 2 ÷ 2.5. The average *yearly* rate is thus equivalent to the density using person-time as the denominator:

$$\text{Yearly rate} = \frac{\frac{x}{n}}{t} = \frac{x}{n \times t} = \frac{x}{\text{person-years}} = \text{Density} = \frac{2}{2.5 \times 2} = \frac{2}{5} = 0.40$$

On the other hand, when losses and events do not occur in an approximate uniform fashion, the incidence rate based on the average study population and the incidence density for a given population and time period may be discrepant. For example, based on the hypothetical data in Figure 2-3, the estimate of the mean yearly incidence based on the average population was 54.5/100 person-years, whereas that based on the incidence density was 63/100 person-years. In real life, when the sample size is large and provided that the time interval is reasonably short, the assumption of uniformity of events/losses is likely to be met.

The notion that the average population is equivalent to the total number of person-time when events and withdrawals are uniform is analogous to the assumption regarding uniformity of events and withdrawals in the actuarial life table (see Section 2.2.1, Equation 2.1). When it is not known exactly when the events occurred in a given time period, each person in whom the event occurs or who enters or withdraws from the study is assumed to contribute one-half of the follow-up time of the interval. (It is expected that this will be the average across a large number of individuals entering/exiting at different times throughout each time period for which person-time is estimated.)

The correspondence between rate (based on grouped data) and density (based on individual person-time) is conceptually appealing, as it allows the comparison of an average yearly rate based on an average population—which, in vital statistics, is usually the midpoint, or July 1, population estimate—with a density based on person-years. It is, for example, a common practice in occupational

epidemiology studies to obtain an expected number of events needed for the calculation of the standardized mortality ratio by applying population vital statistics age-specific rates to the age-specific number of person-years accumulated by an exposed cohort (see Chapter 7, Section 7.3.2).

Stratifying Person-Time and Rates According to Follow-up Time and Covariates

The calculation of person-time contributed by a given population or group is simply the sum of the person-time contributed by each individual in the group during the follow-up period. In most analytical prospective studies relevant to epidemiology, the risk of the event changes with time. For example, the incidence of fatal or nonfatal events may increase with time, as when healthy individuals are followed up as they age. In other situations, risk diminishes as follow-up progresses, as in a study of complications after surgery or of case fatality after an acute myocardial infarction. Because calculating an overall average rate over a long time period when the incidence is not uniform violates the assumptions discussed previously in this chapter (and does not make much sense), it is necessary to estimate the event rate for time intervals within which homogeneity of risk can be assumed. Thus, it is often important to stratify the follow-up time and calculate the incidence rate for each time stratum (as seen in the example based on the data in Table 2-6). Furthermore, in a cohort study, one may additionally wish to control for potentially confounding variables (see Chapter 5). Time and other confounders can be taken into account by stratifying the follow-up time for each individual according to other time variables (e.g., age) and categories of the confounder(s) and then summing up the person-time within each stratum.

The following examples illustrate the calculation of person-time and the corresponding incidence rates based on the data shown in **TABLE 2-8**, from a hypothetical study of four postmenopausal

TABLE 2-8 Hypothetical data for four postmenopausal women followed for mortality after breast cancer surgery.

	Woman no. 1	Woman no. 2	Woman no. 3	Woman no. 4
Date of surgery	2003	2005	2000	2002
Age at surgery	58	50	48	54
Age at menopause	54	46	47	48
Smoking at time of surgery	Yes	No	Yes	No
Change in smoking status (year)	Quits (2006)	No	No	Starts (2003)
Type of event	Death	Loss	Withdrawal alive	Death
Date of event	2009	2008	2010	2004

women followed for mortality after breast cancer surgery (2000 to 2010). Table 2-8 provides the dates of surgery ("entry"), the date of the event (death or censoring), ages at surgery and at menopause, and smoking status.

One Time Scale. Based on the data from Table 2-8, the follow-up of these four women is displayed in **FIGURE 2-8**. The top panel of Figure 2-8 displays the follow-up according to calendar time for each of the four women; the bottom panel displays the follow-up time after surgery. In Figure 2-8 (top), because the precise dates of surgery and events are not known, it is assumed that they occur in the middle of the corresponding year (discussed previously in this chapter).

If it could be assumed that the risk of the event was approximately uniform within 5-year intervals, it would be justified to calculate person-time separately for the first and second 5-year calendar periods (Figure 2-8, top); the calculation of the rates is shown in **TABLE 2-9**. Individuals whose follow-up starts or ends sometime during a given year are assigned one-half of a person-year. For example, a contribution of 0.5 person-year is made by woman 1 in 2003, as her surgery was

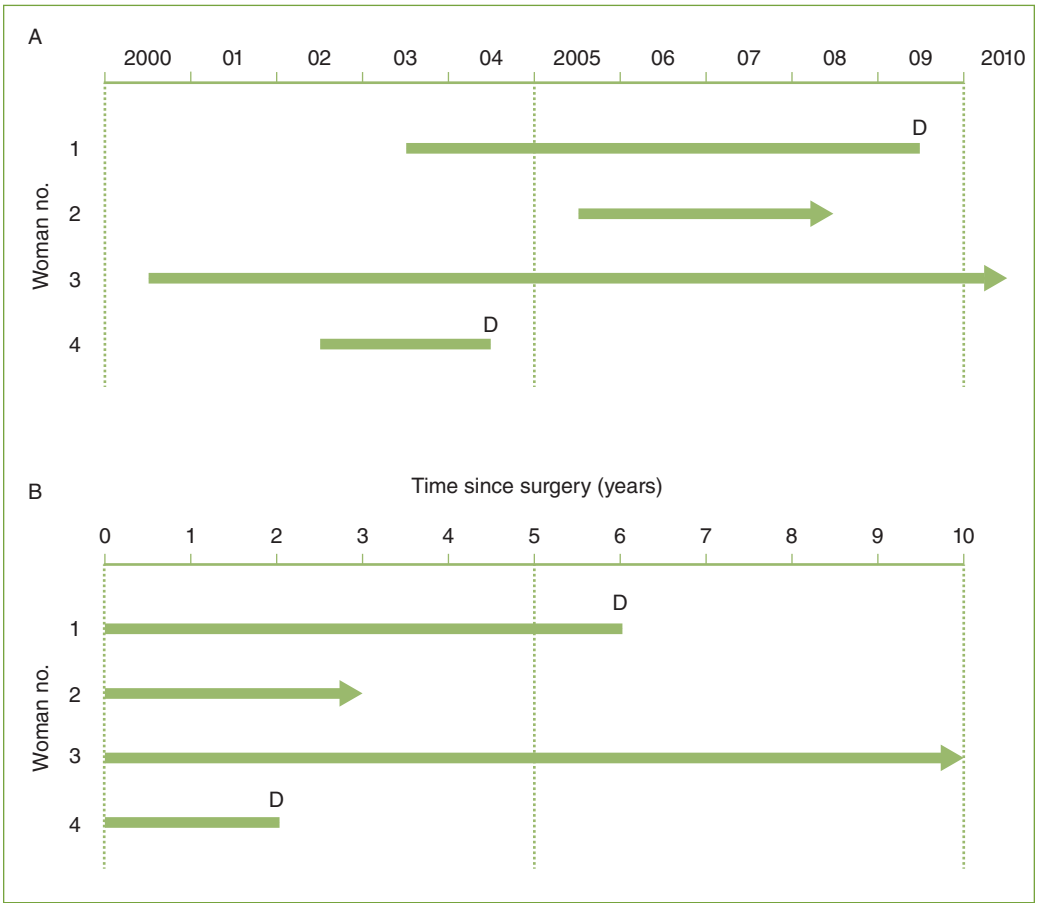

FIGURE 2-8 Schematic representation of person-time for the hypothetical data in Table 2-8, according to one time scale: calendar time (A) and follow-up time (B). Censored observations are shown with arrows; D = death.

TABLE 2-9 Stratification of person-time and rates according to calendar time, based on Table 2-8 and Figure 2-8A.

Calendar time	Person-years	Events	Incidence rate
2000–2004	8	1	0.125
2005–2009	12.5	1	0.080
(2010–2014)	(0.5)	(0)	(0)

TABLE 2-10 Stratification of person-time and rates according to follow-up time (time since surgery), based on Table 2-8 and Figure 2-8B.

Time since surgery	Person-years	Events	Rate
0–4 years	15	1	0.0667
5–9 years	6	1	0.1667

carried out at some time in 2003. Thus, the total person-time for the period 2000 to 2004 is 1.5 years (woman 1) + 4.5 years (woman 3) + 2 years (woman 4) = 8 person-years.

Alternatively, one might be interested in examining the rates in this study according to follow-up time, as shown in **TABLE 2-10**. For example, because woman 1 was followed from 2003.5 to 2009.5, she can be assumed to have a full 6-year follow-up (Figure 2-8, bottom).

Two or More Time Scales. Epidemiologic cohorts are often constituted by free-living individuals who interact and are the subject of multiple and varying biological and environmental circumstances. Thus, it is frequently important to take into consideration more than one time scale; the choice of time scales used for the stratification of follow-up time varies according to the characteristics and goals of the study (**TABLE 2-11**).

In **FIGURE 2-9**, the person-time and outcomes of the four women from Table 2-8 are represented according to two time scales, age and calendar time. In this type of graphic representation (known as a Lexis diagram), each individual's trajectory is represented by diagonals across the two time scales. As in the previous example and given that the time data are approximate (in whole years), it is assumed that entry, censoring, and event of interest occur at the midpoint of the 1-year interval.

TABLE 2-12 shows the corresponding estimates of total person-time in each two-dimensional stratum. They are obtained by adding up the total time lived by the individuals in the study in each age/calendar time stratum represented by eight squares in Figure 2-9; for example, for those 55 to 59 years old between 2000 and 2004, it is the sum of the 1.5 years lived by woman 4 (between age 55 and age 56.5 years, the assumed age of her death) and the 1.5 years lived by woman 1 in that stratum. The relevant events (deaths) are also assigned to each corresponding stratum, thus allowing the calculation of calendar- and age-specific incidence rates, also shown in Table 2-12.

TABLE 2-11 Examples of time scales frequently relevant in the context of cohort studies.

Time scale	Type of study
Follow-up time (time since recruitment)	All studies
Age	All studies
Calendar time	All studies (especially if recruitment is done over an extended period)
Time since employment	Occupational studies
Time since menarche	Studies of reproductive outcomes
Time since seroconversion	Follow-up of patients with HIV infection
Time since diagnosis	Prognostic studies of cancer patients

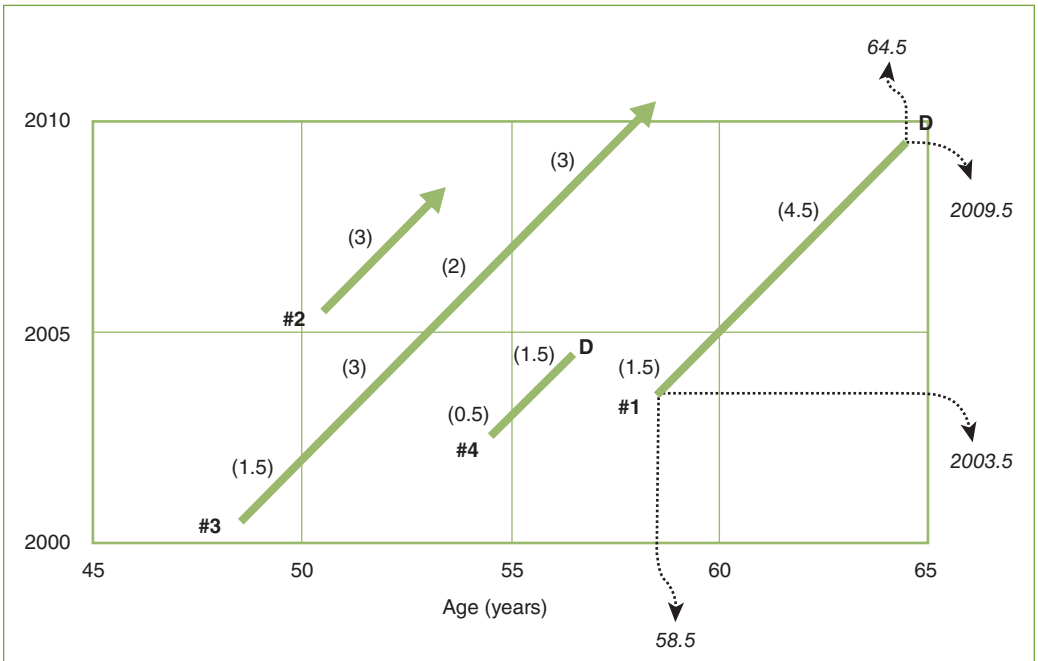

FIGURE 2-9 Schematic representation of person-time for the four women in Table 2-8 according to two time scales, age and calendar time, categorized in 5-year intervals. Because time data are given in whole years, entry, events, and withdrawals are assumed to occur exactly at the middle of the year. Censored observations are represented by an arrow. The total time within each time stratum for all four women is shown in parentheses. The entry and exit times for woman no. 1 are given in italics; D = death.

TABLE 2-12 Stratification of person-time and rates according to calendar time and age (see Figure 2-9).

Calendar time	Age (years)	Person-years	Events	Rate
2000–2004	45–49	1.5	0	0
	50–54	3.5	0	0
	55–59	3	1	0.3333
	60–64	0	—	—
2005–2009	45–49	0	—	—
	50–54	5	0	0
	55–59	3	0	0
	60–64	4.5	1	0.2222

Other time scales may be of interest. In occupational epidemiology, for example, it may be of interest to obtain incidence rates of certain outcomes taking into account three time scales simultaneously, for example, age (if the incidence depends on age), calendar time (if there have been secular changes in exposure doses), and time since employment (so as to consider a possible cumulative effect). For this situation, one could conceive a tridimensional analogue to Figure 2-9: cubes defined by strata of the three time scales and each individual's person-time displayed across the tridimensional diagonals.

The layout for the calculation of person-time and associated incidence rates described in this section can be used for the internal or external comparison of stratified rates by means of standardized mortality or incidence ratios (see Chapter 7, Section 7.3.2).

Time and Fixed or Time-Dependent Covariates. Stratification according to other variables, in addition to time, may be necessary in certain situations. For example, the data in Table 2-8 could be further stratified according to an additional time scale (e.g., time since menopause) and additional covariates (e.g., smoking status at the time of surgery). Thus, instead of eight strata, as in Figure 2-9 and Table 2-12, one would need to stratify the person-time into 32 strata defined by the combination of all four variables (calendar time, age, time since menopause, and smoking). Individuals under observation would shift from stratum to stratum as their status changes. For example, woman 1 (Table 2-8) is a smoker who enters the study in 2003 at the age of 58 years (that is assumed to be 2003.5 and 58.5 years, respectively, as discussed previously), 4 years after menopause. Thus, as she enters the study in the stratum "2000–2004/55–59 years of age/smoker-at-baseline/menopause $<$ 5 years," after contributing 1 person-year to this stratum (i.e., in 2004.5 at age 59.5 years), she becomes "$\geq$ 5 years after menopause" and thus shifts to a new stratum: "2000–2004/55–59 years of age/smoker-at-baseline/menopause $\geq$ 5 years." Half a year later, she turns 60 and enters a new stratum ("2005–2009/60–64 years of age/smoker-at-baseline/menopause $\geq$ 5 years") to contribute her last 4.5 person-years of observation before her death in 2009.5.

In the preceding example, smoking is treated as a fixed covariate, as only baseline status is considered; however, information on smoking status change is available in the hypothetical study data shown in Table 2-8. Changes in exposure status for certain covariates can easily be taken into account when using the person-time strategy. For example, using the information at baseline and change in smoking status shown in Table 2-8, assignment of person-time according to smoking as a *time-dependent covariate* can be represented as illustrated in **FIGURE 2-10**. Using this approach, each relevant event is assigned to the exposure status at the time of the event. Thus, woman 1's death is assigned to the "nonsmoking" group, whereas that of woman 4 is assigned to the "smoking" group. (The latter assignments are opposite to those based on smoking status at baseline described in the preceding paragraph.)

To use the person-time approach to take into account changing exposures involves an assumption akin to that used in crossover clinical trials, that after the exposure status changes, so does the associated risk. This is merely another way of stating that there is no *accumulation* of risk and thus that the effect of a given exposure is "instantaneous." Whether this assumption is valid depends on the specific exposure or outcome being considered. For example, for smoking, the assumption may be reasonable when studying thromboembolic events likely to result from the acute effects of smoking (e.g., those leading to sudden cardiac death). On the other hand, given the well-known latency and cumulative effects leading to smoking-related lung cancer, the assumption of an instantaneous effect would be unwarranted if lung cancer were the outcome of interest. (The cumulative effect of smoking on lung cancer risk can be easily inferred from the fact that the risk in smokers who quit decreases yet never becomes the same as that in people who have never smoked.) If there is a cumulative effect, the approach illustrated in Figure 2-10 (e.g., assigning the event in woman 1 to the nonsmoking category) will result in misclassification of exposure status (see Chapter 4, Section 4.3).

The cumulative effects of exposure can be taken into account with more complex exposure definitions; for example, total pack-years of smoking could be considered even among former smokers. Moreover, lag or latency times could also be introduced in the definition of person-time in relation to events, a frequent practice in occupational or environmental epidemiology studies.[11(pp150-155)]

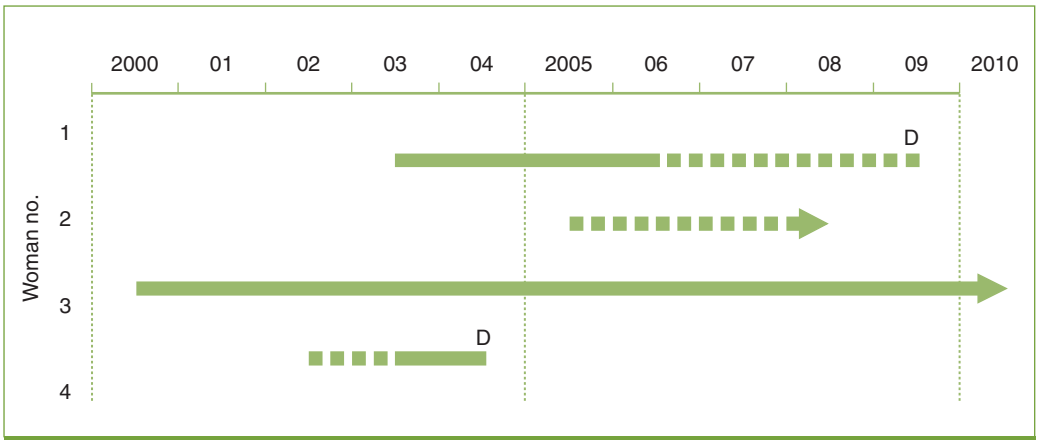

FIGURE 2-10 Schematic representation of person-time for the four women in Table 2-8 according to time-dependent smoking status. Solid lines represent smoking status; broken lines represent nonsmoking status.

Obviously, when the study requires stratification according to more than one time scale and several covariates, person-time and rates will need to be calculated for dozens or even hundreds of multidimensional strata, which will require the use of computer programs.[12-14]

2.2.3 Comparison Between Measures of Incidence

For the estimation of the different incidence measures, the numerator (number of deaths) is constant; the differentiation between the measures is given by the way the denominator is calculated. The main features that distinguish cumulative probability on the one hand and density or rate on the other are shown in **EXHIBIT 2-1**. As discussed previously, the upper limit of values for a rate or a density may exceed 100%, whereas values for probabilities cannot be greater than 100%.

Rates are often calculated as *yearly average rates* or, for example, as *rates per 1000 person-years*. The latter implies an average rate per 1000 persons per year, which underscores the correspondence between a vital statistics–derived rate and a rate per person-time, as discussed previously. On the other hand, no time unit whatsoever is attached to a cumulative incidence (a probability), thus requiring that the relevant time period always be specified (e.g., "the cumulative probability for the initial 3 years of follow-up").

With regard to their numerical value, a cumulative incidence and a rate can be compared only if they are based on the same time unit (e.g., cumulative incidence over a 1-year period and rate per

EXHIBIT 2-1 Comparing measures of incidence: cumulative incidence vs incidence rate.

	Cumulative incidence		Incidence density/rate	
	If follow-up is complete	*If follow-up is incomplete*	*Individual data (cohort)*	*Grouped data (area)*
Numerator	Number of cases	Classic life table Kaplan–Meier	Number of cases	Number of cases
Denominator	Initial population		Person-time	Average population*
Units	Unitless		$Time^{-1}$	
Range	0 to 1		0 to infinity	
Synonyms	Proportion Probability		Incidence density†	

*Equivalent to person-time when events and losses (or additions) are homogeneously distributed over the time interval of interest.
†In the text, the term *density* is used to refer to the situation in which the exact follow-up time for each individual is available; in real life, however, the terms *rate* and *density* are often used interchangeably.

EXHIBIT 2-2 Comparing absolute numerical values of cumulative incidence based on the actuarial life table and rate (assuming that follow-up interval equals person-time unit).

Notice that (as long as $x > 0$) the denominator of the rate will always be smaller than that of the cumulative incidence (960 vs 985 in the example), thus explaining the larger absolute value of the rate.

Cumulative incidence	In the absence of censoring	In the presence of censoring (C)	Example: $N = 1000$ individuals followed for 1 year, $x = 50$ events, $C = 30$ censored observations
(***q***) is calculated based on number of individuals at risk *at the beginning* of the interval (***N***)	$q = \frac{x}{N}$	$q = \frac{x}{N - \frac{1}{2}C}$	$= \frac{50}{1000 - \frac{1}{2}30}$ $= \frac{50}{985} = 0.0508$
Rate is calculated based on person-time of observation over the follow-up, subtracting person-time lost by the cases (***x***)	$\text{Rate} = \frac{x}{N - \frac{1}{2}x}$	$\text{Rate} = \frac{x}{N - \frac{1}{2}x - \frac{1}{2}C}$	$= \frac{50}{1000 - \frac{1}{2}50 - \frac{1}{2}30}$ $= \frac{50}{960} = 0.0521$

person-year). Under this circumstance, the general rule is that, in absolute value, the rate will always be larger than the cumulative incidence. The rationale for this rule is best explained when comparing a rate with a cumulative incidence based on the classic life table, as illustrated in **EXHIBIT 2-2**. Although losses because of censoring are similarly taken into account in the denominator of both cumulative incidence and rate, the observation time "lost" by the cases is subtracted from the denominator of the rate but not from the probability-based cumulative incidence (which uses number of observations at the start of the interval corrected for losses regardless of how many cases occur subsequently). As a result, the denominator for the rate will always tend to be smaller than that of the cumulative incidence and thus the larger absolute value of the former when compared with the latter. When

EXHIBIT 2-3 Assumptions necessary for survival and person-time analyses.

	Survival analysis	Person-time
If there are losses to follow-up:	Censored observations have an outcome probability that is similar to that of individuals remaining under observation.	
If intervals are used and there are losses during a given interval:	Losses are uniform over the interval.	
If risk is calculated over intervals:	Risk is uniform during the interval.	*N* individuals followed for *T* units of time have the same risks as *T* individuals followed for *N* units of time.
If accrual of study subjects is done over a relatively long time period:	There are no secular trends over the calendar period covered by the accrual.	

events are relatively rare, the discrepancy is very small (e.g., example in Exhibit 2-2); as the frequency of the event increases, so will the numerical discrepancy between cumulative incidence and rate pertaining to the same time unit.

Finally, regarding assumptions, all methods for the calculation of incidence share the fundamental assumptions in the analysis of cohort data that were discussed in previous sections and are summarized in **EXHIBIT 2-3**: independence between censoring and survival and lack of secular trends. Additional assumptions are needed depending on the specific requirements of the method (e.g., uniformity of risk across defined interval in classic life-table and/or person-time–based analyses).

Experienced epidemiologists have learned that, whereas each approach has advantages and disadvantages, the ultimate choice of how to present incidence data is dictated by pragmatism (and/or personal preference). Thus, in a cohort study without an "internal" comparison (e.g., unexposed) group—as may be the case in occupational epidemiology research—estimation of densities, rather than probabilities, allows using available population rates as control rates. On the other hand, probabilities are typically estimated in studies with a focus on the temporal behavior (or "natural history") of a disease, as in studies of survival after diagnosis of disease.

2.2.4 The Hazard Rate

The instantaneous incidence rate (density) is the so-called *hazard rate*, also named *instantaneous conditional incidence* or *force of morbidity* (or *mortality*). In the context of a cohort study, the hazard rate is defined as

each individual's instantaneous probability of the event at precisely time t (or at a small interval $[t, t + \Delta t]$), given (or "conditioned" on) the fact that the individual was at risk at time t. The hazard rate is defined for each particular point in time during the follow-up. In mathematical terms, this is defined for a small time interval (Δt close to zero) as follows:

$$h(t) = \frac{P(\text{event in interval between } t \text{ and } [t + \Delta t] \mid \text{alive at } t)}{\Delta t}$$

The hazard is analogous to the conditional probability of the event that is calculated at each event time using Kaplan–Meier's approach (Table 2-3, column 4); however, because its denominator is "time-at-risk," the rate is measured in units of time^{-1}. Another important difference is that, in contrast to Kaplan–Meier's conditional probability, the hazard rate *cannot* be directly calculated, as it is defined for an infinitely small time interval; however, the hazard *function* over time can be estimated using available parametric survival analysis techniques.[15]

The hazard rate is a useful concept when trying to understand some of the statistical techniques used in survival analysis, particularly those pertaining to proportional hazards regression (see Chapter 7, Section 7.4.4). It is outside the scope of this textbook, however, to discuss the complex mathematical properties of the hazard rate; the interested reader should consult more advanced statistical textbooks, such as Collett's.[15]

2.3 Measures of Prevalence

Prevalence is defined as the frequency of *existing* cases of a disease or other condition in a given population at a certain time or period. Depending on how "time" is defined, there are two kinds of prevalence, point prevalence and period prevalence (Table 2-1). *Point prevalence* is the frequency of a disease or condition at a point in time; it is the measure estimated in the so-called prevalence or cross-sectional surveys, such as the National Health and Nutrition Examination Surveys conducted by the U.S. National Center for Health Statistics.[16] For the calculation of point prevalence, it is important to emphasize that all existing cases at a given point in time are considered prevalent regardless of whether they are old or more recent. *Period prevalence* is less commonly used and is defined as the frequency of an existing disease or condition during a defined time period. For example, the period prevalence of condition Y in year 2010 includes all existing cases on January 1, 2010, plus the new (incident) cases occurring during the year. A special type of period prevalence is the *cumulative lifetime prevalence*, which provides an estimate of the occurrence of a condition at any time during an individual's past (up to the present time). For example, the United States-based 2003 National Youth Risk Behavior Survey estimated the lifetime prevalence of asthma among high school students to be 18.9%, whereas the estimated point prevalence at the time of the survey was estimated to be 16.1%.[17]

In the case of period prevalence, the denominator is defined as the average reference population over the period. In general, when the term *prevalence* is not specified, it can be taken to mean *point prevalence*. As a descriptive measure, point prevalence is a useful index of the magnitude of current health problems and is particularly relevant to public health and health policy (Chapter 10). In addition, prevalence is often used as the basis for the calculation of the *point prevalence rate ratio*, a measure of association in cross-sectional studies or in cohort studies using baseline data.

Because the point prevalence rate ratio is often used as a "surrogate" of the incidence ratio in the absence of prospective cohort data (see Chapter 3, Section 3.3), it is important to understand prevalence's dependence on *both* incidence and duration of the disease after onset; duration is, in turn, determined by either survival for fatal diseases or recovery for nonfatal diseases. In a population in a steady-state situation (i.e., no major migrations or changes over time in incidence/prevalence of the condition of interest), the relationship between prevalence and disease incidence and duration can be expressed by the following formula:[*]

$$\frac{\text{Point prevalence}}{(1 - \text{Point prevalence})} = \text{Incidence} \times \text{Duration} \qquad \text{(Eq. 2.3)}$$

The term [Point prevalence ÷ (1 – Point prevalence)] is the odds of point prevalence (see Section 2.4). Also, in this equation and those derived from it, the time unit for incidence and duration should be the same; that is, if incidence is given as a yearly average, duration should be given using year(s) or a fraction thereof. Equation 2.3 can be rewritten as follows:

$$\text{Point prevalence} = \text{Incidence} \times \text{Duration} \times \left(1 - \text{Point prevalence}\right) \qquad \text{(Eq. 2.4)}$$

As discussed in Chapters 3 and 4, Equation 2.4 underscores the two elements of a disease that are responsible for the difference between incidence and point prevalence: its duration and the magnitude of its point prevalence. When the point prevalence is relatively low (e.g., 0.05 or less), the term (1 – Point prevalence) is almost equal to 1.0, and the following well-known simplified formula defining the relationship between prevalence and incidence is obtained:

$$\text{Point prevalence} \approx \text{Incidence} \times \text{Duration}$$

For example, if the incidence of the disease is 0.5% per year and its average duration (survival after diagnosis) is 10 years, the point prevalence will be approximately 5%.

---

[*]The derivation of this formula is fairly straightforward. Under the assumption that the disease is in steady state, the incidence and number of existing cases at any given point (e.g., X) are approximately constant. For an incurable disease, this implies that the number of new cases during any given time period is approximately equal to the number of deaths among the cases. If N is the population size, I is the incidence, and F is the case fatality rate, the number of new cases can be estimated by multiplying the incidence times the number of potentially "susceptible" ($N - X$); in turn, the number of deaths can be estimated by multiplying the case fatality rate (F) times the number of prevalent cases. Thus, the assumption that the number of new cases approximates the number of deaths among the cases in the period can be formulated as follows: $I \times \left(N - X\right) \approx F \times X$. If there is no immigration, the case fatality rate is the inverse of the duration (D).[3] Thus, after a little arithmetical manipulation and dividing numerator and denominator of the right-hand side term by N, the following equation is obtained:

$$I \times D \approx \frac{X}{(N - X)} = \frac{\text{Prevalence}}{(1 - \text{Prevalence})}$$

An analogous reasoning can be applied to nonfatal diseases, for which F is the proportion cured.

2.4 Odds

Odds are the ratio of the probability of the event of interest to that of the nonevent. This can be defined for both incidence and prevalence. For example, when dealing with incidence probabilities, the odds are as follows:

$$\text{Incidence odds} = \frac{q}{1-q}$$

(Alternatively, knowing the odds allows the calculation of probability: q = Odds ÷ [1 + Odds].)

The point prevalence odds are as follows (see also Equation 2.3):

$$\text{Point prevalence odds} = \frac{\text{Point prevalence}}{1 - \text{Point prevalence}}$$

Both odds and proportions can be used to express "frequency" of the disease. An odds approximates a proportion when the latter is small (e.g., less than 0.1). An example follows:

$$\text{Proportion} = 0.05$$

$$\text{Odds} = 0.05 / (1 - 0.05) = 0.05/0.95 = 0.0526$$

It is easier to grasp the intuitive meaning of the proportion than that of the odds perhaps because, in a description of odds, the nature of the latter as a ratio is often not clearly conveyed. For example, if the proportion of smokers in a population is 0.20, the odds are as follows:

$$\text{Odds} = \frac{\text{Proportion of smokers}}{1 - \text{Proportion of smokers}} = \frac{\text{Proportion of smokers}}{\text{Proportion of nonsmokers}}$$

$$\text{or } 0.20 \div (1 - 0.20) = 0.20 \div 0.80 = 1{:}4 = 0.25.$$

Thus, there are two alternative ways to describe an odds estimate: either as an isolated number, 0.25, implying that the reader understands that it intrinsically expresses a ratio, 0.25:1.0, or clearly as a ratio—in the example, 1:4—conveying more explicitly the message that, in the study population, for every smoker there are four nonsmokers.

Traditionally and commonly used in certain venues (e.g., in the horse-race betting world), the odds are rarely if ever used by epidemiologists as measures of disease occurrence. However, the ratio of two odds (the *odds ratio*) is a very popular measure of association in epidemiology because the odds ratio allows the estimation of the easier-to-grasp relative risk in case-based case-control studies and it is the measure of association derived from *logistic regression*, one of the most widely used methods for multivariate analysis of epidemiologic data (see Chapter 1, Section 1.4.2; Chapter 3, Section 3.4.1; and Chapter 7, Section 7.4.3).

References

1. *Merriam-Webster's Collegiate Dictionary*. 11th ed. Springfield, MA: Merriam-Webster Inc.; 2003.
2. Centers for Disease Control and Prevention. Revision of the CDC surveillance case definition for acquired immunodeficiency syndrome. *J Am Med Assoc*. 1987;258:1143-1145.
3. Rothman K, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
4. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.

5. Reed LJ, Merrell M. A short method for constructing an abridged life table: 1939. *Am J Epidemiol*. 1995;141:993-1022; discussion 1991-1022.
6. Pooling Project Research Group. Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report of the Pooling Project. *J Chron Dis*. 1978;31:201-306.
7. Gordis L. *Epidemiology*. 5th ed. Philadelphia: Elsevier Saunders; 2014.
8. Kahn HA, Sempos CT. *Statistical Methods in Epidemiology*. New York: Oxford University Press; 1989.
9. Young T, Finn L, Peppard PE, et al. Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin Sleep Cohort. *Sleep.* 2008;31:1071-1078.
10. Halanych JH, Safford MM, Kertesz SG, et al. Alcohol consumption in young adults and incident hypertension: 20-year follow-up from the Coronary Artery Risk Development in Young Adults Study. *Am J Epidemiol.* 2010;171:532-539.
11. Checkoway H, Pearce NE, Crawford-Brown DJ. *Research Methods in Occupational Epidemiology*. New York: Oxford University Press; 1989.
12. Monson RR. Analysis of relative survival and proportional mortality. *Comput Biomed Res*. 1974;7:325-332.
13. Macaluso M. Exact stratification of person-years. *Epidemiology*. 1992;3:441-448.
14. Pearce N, Checkoway H. A simple computer program for generating person-time data in cohort studies involving time-related factors. *Am J Epidemiol*. 1987;125:1085-1091.
15. Collett D. *Modelling Survival Data in Medical Research*. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC; 2014.
16. Curtin LR, Mohadjer LK, Dohrmann SM, et al. The National Health and Nutrition Examination Survey: sample design, 1999-2006. *Vital Health Stat.* 2012;155:1-39.
17. Centers for Disease Control and Prevention. Self-reported asthma among high school students: United States, 2003. *MMWR*. 2005;54:765-767.

Exercises

1. A prospective study with a 2-year (24-month) follow-up was conducted. Results are shown in the table for individuals who either died or were censored before the end of the follow-up period.

Survival data for 20 participants of a hypothetical prospective study.

Follow-up time (months)	Event
2	Death
4	Censored
7	Censored
8	Death
12	Censored
15	Death
17	Death
19	Death
20	Censored
23	Death

   a. Using the data from the table, for all deaths calculate (1) the probability of death at the exact time when each death occurred, (2) the probability of survival beyond the time when each death occurred, and (3) the cumulative probabilities of survival.
   b. What is the cumulative survival probability at the end of the follow-up period?
   c. Using arithmetic graph paper, plot the cumulative probabilities of survival.
   d. What is the simple proportion of individuals apparently surviving (i.e., not observed to die) through the end of the study's observation period?
   e. Why are the simple proportion surviving and the cumulative probability of survival different?
   f. Using the same data, calculate the overall death rate per 100 person-years. (To facilitate your calculations, you may wish to calculate the number of person-months and then convert that to the number of person-years.)

g. Calculate the rates separately for the first and second years of follow-up. (For this calculation, assume that the individual who withdrew at month 12 withdrew just after midnight on the last day of the month.)
h. Assuming that there was no random variability, was it appropriate to calculate the rate per person-year for the total 2-year duration of follow-up?
i. What is the most important assumption underlying the use of both survival analysis and the person-time approach?
j. Now, assume that the length of follow-up was the same for all individuals (except those who died). Calculate the proportion dying and the odds of death in this cohort.
k. Why are these figures so different in this study?

2. In a cohort study of individuals aged 65 years and older and free of dementia at baseline,[*] the associations of age and APOE $\varepsilon 4$ with the risk of incident Alzheimer's disease (AD) were investigated. The table shows the number of individuals, person-years, and probable cases of AD overall and according to age ($<$ 80 and $\geq$ 80 years old) and, separately, to presence of APOE $\varepsilon 4$.

Probable Alzheimer's disease (AD) by age and APOE $\varepsilon 4$.

	Number of individuals	Number with probable AD/person-years	Density of AD per 100 person-years	Average duration of follow-up
All subjects	3099	263/18,933		
$<$ 80 years	2343	157/15,529		
$\geq$ 80 years	756	106/3404		
APOE$\varepsilon 4(+)$	702	94/4200		
APOE$\varepsilon 4(-)$	2053	137/12,894		

a. Calculate the densities of AD per 100 person-years and the average durations of follow-up for all subjects and for each subgroup.
b. Why is it important that the follow-up durations be similar for the "exposed" and "unexposed" categories particularly in this study?
c. When calculating a density for a given long follow-up period, the assumption is that the risk remains the same throughout the duration of follow-up. Is this a good assumption in the case of Alzheimer's disease? Why or why not?

[*] Li G, Shofer JB, Rhew IC, et al. Age-varying association between statin use and incident Alzheimer's disease. *J Am Geriatr Soc*. 2010;58:1311-1317.

3. In a case-based case-control study of risk factors for uterine leiomyoma, the authors assessed a history of hypertension in cases and controls, as shown in the table here.

History of hypertension	Cases		Controls	
	Number	Percentage	Number	Percentage
Absent	248	78.0	363	92.4
Present	70	22.0	30	7.6
Total	318	100.0	393	100.0

   a. Using the absolute numbers of cases and controls with and without a history of hypertension, calculate the absolute odds of history of hypertension separately in cases and in controls.
   b. Now, calculate the odds of hypertension using the percentages of cases and controls with a history of hypertension.
   c. What can you conclude from comparing the response to Exercise 2a to the response to Exercise 2b?
   d. Why are the odds of a history of hypertension more similar to the proportion of individuals with a history of hypertension in controls than in cases?

4. The baseline point prevalence of hypertension in African American women aged 45 to 64 years included in the Atherosclerosis Risk in Communities (ARIC) cohort study was found to be 56%.[†] In this study, over a follow-up period of 6 years, the average yearly incidence of hypertension in African American women was estimated to be about 5% and stable over the years.[‡] Using these data, estimate the average duration of hypertension in African American women in the ARIC Study.

5. Wakelee et al. reviewed lung cancer incidence rates in never smokers in several cohort studies with varying lengths of follow-up and found them to vary from 4.8 to 20.8 per 100,000 person-years.[§]
   a. How can rates per person-years in this review by Wakelee et al. be interpreted given the variation in the length of follow-up among the cohorts?
   b. What are the conditions that render this interpretation incorrect?
   c. Which are the assumptions common to survival analysis and the person-time strategy?

[†]Harris MM, Stevens J, Thomas N, Schreiner P, Folsom AR. Associations of fat distribution and obesity with hypertension in a bi-ethnic population: the ARIC Study. *Obesity Res*. 2000;8:516-524.
[‡]Fuchs FD, Chambless LE, Whelton PK, Nieto FJ, Heiss G. Alcohol consumption and the incidence of hypertension: the Atherosclerosis Risk in Communities Study. *Hypertension*. 2001;37:1242-1250.
[§]Wakelee HA, Chang ET, Gomez SL, et al. Lung cancer incidence in never smokers. *J Clin Oncol*. 2007;25:472-478.