Chapter 4
Measures of occurrence of disease and other health-related events

4.1 Introduction

Epidemiological research is based on the ability to quantify the occurrence of disease (or any other health-related event) in populations. To do this, the following must be clearly defined:

(1) What is meant by a case, i.e., an individual in a population who has the disease, or undergoes the event of interest (e.g., death).
(2) The population from which the cases originate.
(3) The period over which the data were collected.

4.1.1 Defining a case—the numerator

In epidemiology, it is seldom easy to define what is meant by a 'case', even for a well known condition. The epidemiological definition of a case is not necessarily the same as the clinical definition, and epidemiologists are often forced to rely on diagnostic tests that are less invasive and cheaper than those normally used by clinicians. Nevertheless, for study purposes, it is important to standardize the case definition. For instance, should 'cancer cases' comprise only those that were confirmed histologically? Should *in situ* lesions be included? For cancers of paired organs (e.g., breast, testis, kidney), should the number of cases counted reflect the number of individuals who develop the cancer or the number of organs affected? Cancer epidemiologists are also interested in measuring the frequency of other health-related event, so, for example, someone who smokes, uses oral contraceptives or uses a certain health service might be counted as a case.

Another important consideration when dealing with recurrent non-fatal conditions (e.g., the common cold) is to decide whether, for a given individual, each episode or occurrence should be counted as a case, or only the first attack. In this chapter, we assume that individuals can only suffer from one episode of the condition of interest; however, the measures of occurrence presented can be modified to cover recurrent episodes.

Cases may be identified through disease registries, notification systems, death certificates, abstracts of clinical records, surveys of the general population, etc. It is important, however, to ensure that the numerator both includes all cases occurring in the study population, and excludes cases from elsewhere. For instance, when measuring the occurrence of a disease in a particular town, all cases that occurred among its residents should be

included in the numerator, even those diagnosed elsewhere. In contrast, cases diagnosed in people who are normally resident elsewhere should be excluded.

4.1.2 Defining the population at risk—the denominator

Knowing the number of cases in a particular population is on its own of little use to the epidemiologist. For example, knowing that 100 cases of lung cancer occurred in city A and 50 cases in city B does not allow the conclusion that lung cancer is more frequent in city A than in city B: to compare the frequency of lung cancer in these two populations, we must know the size of the populations from which the cases originated (i.e., the denominator).

The population at risk must be defined clearly, whether it be the residents of one particular town, the population of a whole country or the catchment population of a hospital. The definition must exclude all those who are not usually resident in that area. If possible, it should also exclude all those who are not at risk of the event under investigation. For instance, in quantifying the occurrence of cervical cancer in a population, women who have undergone hysterectomy should ideally be excluded, as they cannot develop this cancer. However, as the data necessary to exclude such women are seldom available, all women are usually included in the denominator.

4.1.3 Time period

As most health-related events do not occur constantly through time, any measure of occurrence is impossible to interpret without a clear statement of the period during which the population was at risk and the cases were counted. The occurrence of lung cancer in most western countries illustrates this point: incidence of this disease was much lower in the early years of this century than today.

4.2 Measures of occurrence

There are two principal measures of occurrence: prevalence and incidence.

4.2.1 Prevalence

Point prevalence is the proportion of existing cases (old and new) in a population at a single point in time.

$$\text{Point prevalence} = \frac{\text{No. of existing cases in a defined population at one point in time}}{\text{No. of people in the defined population at the same point in time}}$$

This measure is called point prevalence[a] because it refers to a single point in time. It is often referred to simply as prevalence.

[a] Period prevalence is a variation that represents the number of people who were counted as cases at any time during a specified (short) period, divided by the total number of people in that population during that time. This measure is used when the condition is recurrent and non-fatal, and so is seldom used in cancer epidemiology. An example of period prevalence would be the proportion of women who have used oral contraceptives at any time during the 12-month period preceding the day of the survey.

Example 4.1. *Each line in Figure 4.1 represents an individual (subject) in a particular population. Some subjects developed the condition of interest and either recovered or died from it. Others left the population and went to live elsewhere. Because of these dynamic changes, the magnitude of the prevalence varies from one point in time to another.*

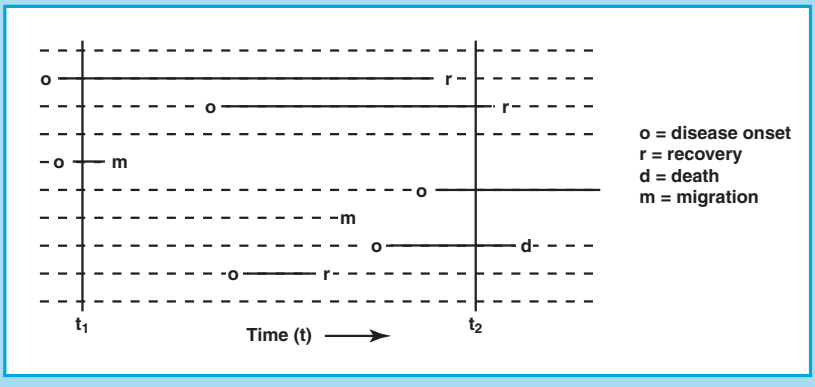


Figure 4.1.
Changes in the disease status and migration of members of a population over time, and how these changes affect the prevalence of the disease in the population.

Prevalence at time $t_1 = {}^{2}/_{10} = 0.20 = 20\%$
Prevalence at time $t_2 = {}^{3}/_{8} = 0.38 = 38\%$

Although as with any proportion, prevalence has no time units, the point in time to which it refers must always be specified (Examples 4.1 and 4.2). The term 'prevalence rate' is often wrongly used instead of 'prevalence': this is incorrect, as prevalence is, by definition, a proportion not a rate (see Section 4.2.2).

Example 4.2. *In 1985, a study was carried out in a small town to determine the prevalence of oral contraceptive use among women aged 15–44 years. All women between these ages resident in the town were interviewed and asked about their current use of oral contraceptives. The prevalence of oral contraceptive use in that town in 1985 among women aged 15–44 years was 0.5 (50%).*

It may be difficult to define a prevalent cancer case. Cancer registries generally assume that once diagnosed with cancer, an individual represents a prevalent case until death (see Section 17.6.1). However, this assumption is not always correct, as people diagnosed with cancer may survive for a long period without any recurrence of the disease and may die from another cause.

Prevalence is the only measure of disease occurrence that can be obtained from cross-sectional surveys (see Chapter 10). It measures the

burden of disease in a population. Such information is useful to public-health professionals and administrators who wish to plan the allocation of health-care resources in accordance with the population's needs.

4.2.2 Incidence

The number of cases of a condition present in a population at a point in time depends not only on the frequency with which new cases occur and are identified, but also on the average duration of the condition (i.e., time to either recovery or death). As a consequence, prevalence may vary from one population to another solely because of variations in duration of the condition.

Prevalence is therefore not the most useful measure when attempting to establish and quantify the determinants of disease; for this purpose, a measurement of the flow of new cases arising from the population is more informative. Measurements of incidence quantify the number of new cases of disease that develop in a population of individuals at risk during a specified time interval. Three distinct measures of incidence may be calculated: risk, odds of disease, and incidence rate.

Risk

Risk is the proportion of people in a population that is initially free of disease who develop the disease within a specified time interval.

$$\text{Risk} = \frac{\text{No. of new cases of disease arising in a defined population over a given period of time}}{\text{No. of disease free people in that population at the beginning of that time period}}$$

Both numerator and denominator include only those individuals who are free from the disease at the beginning of the given period and are therefore at risk of developing it. This measure of incidence can be interpreted as the average probability, or risk, that an individual will develop a disease during a specified period of time.

Often, other terms are used in the epidemiological literature to designate risk, for example, incidence risk and incidence proportion.

Like any proportion, risk has no time units. However, as its value increases with the duration of follow-up, the time period to which it relates must always be clearly specified, as in Example 4.3.

Example 4.3. *A group of 5000 healthy women aged 45–75 years was identified at the beginning of 1981 and followed up for five years. During this period, 20 new cases of breast cancer were detected. Hence, the risk of developing breast cancer in this population during this five-year period was 20/5000 = 0.4%.*

Example 4.4. *A total of 13 264 lung cancer cases in males were diagnosed in a certain population in 1971. These cases were followed up for five years. At the end of this follow-up period, only 472 cases were still alive. The probability of surviving during this five-year period was 472/13 264 = 3.6%. Thus, the probability of dying during the period was 100% – 3.6% = 96.4%. These measures are risks, as they represent the proportion of lung cancer cases who were still alive (or who died) at the end of the follow-up period out of all cases diagnosed at the beginning of the study. These calculations assume that all individuals were followed up for the entire five-year period (or until death if it occurred earlier).*

Risk is a measure commonly used to quantify the survival experience of a group of subjects, as in Example 4.4.

The measures in Example 4.4 are often called survival and fatality 'rates'; this is incorrect as, by definition, they are proportions (see later in this section). These two measures are discussed further in Chapter 12.

Odds of disease

Another measure of incidence is odds of disease, which is the total number of cases divided by the total number of persons who remained disease-free over the study period.

$$\text{Odds of disease} = \frac{\text{No. of new cases of disease arising in a defined population over a given period of time}}{\text{No. of people in that population who remain disease-free during that period}}$$

Example 4.5. *One hundred disease-free individuals were followed up for a certain period of time. By the end of this period, ten had developed the disease of interest (Figure 4.2).*

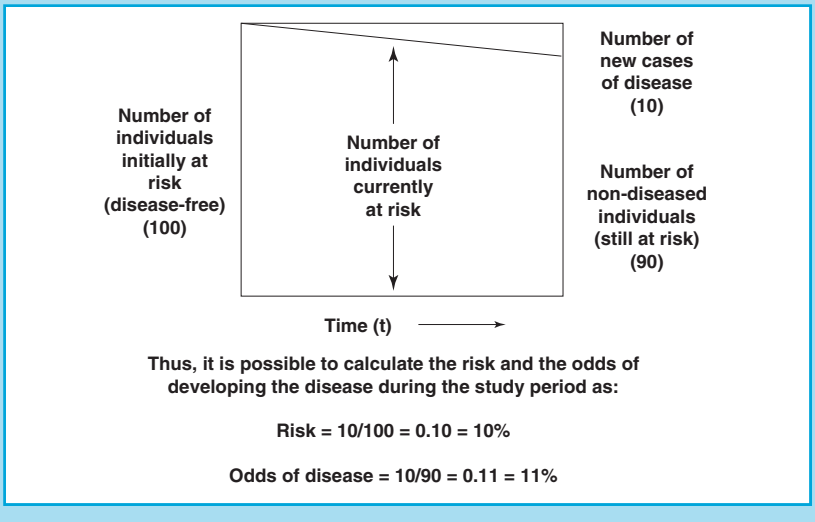


Figure 4.2.
Follow-up of the 100 disease-free individuals described in Example 4.5.

This measure is a *ratio* of the probability of getting the disease to the probability of not getting the disease during a given time period. Thus, it can also be expressed as:

Odds of disease = risk/(1 – risk)

Risk and odds of disease use the same numerator (number of new cases) but different denominators. In the calculation of risk, the denominator is the total number of disease-free individuals at the beginning of the study period, whereas when calculating the odds of disease, it is the number of individuals who remained disease-free at the end of the period (Example 4.5).

Incidence rate

Calculations of risk and odds of disease assume that the entire population at risk at the beginning of the study period has been followed up during the specified time period. Often, however, some participants enter the study some time after it begins, and some are lost during the follow-up; i.e., the population is dynamic. In these instances, not all participants will have been followed up for the same length of time. Moreover, neither of these two measures of incidence takes account of the time of disease onset in affected individuals.

To account for varying lengths of follow-up, the denominator can be calculated so as to represent the sum of the times for which each individual is at risk, i.e., the sum of the time that each person remained under observation and was at risk of becoming a case. This is known as person-time at risk, with time being expressed in appropriate units of measurement, such as person-years (often abbreviated as pyrs).

Example 4.6. *Consider a hypothetical group of nine persons who were followed up from the beginning of 1980 to the end of 1984. Subjects joined the study at different points, as shown in Figure 4.3. Three subjects, (2), (6) and (7), developed the disease of interest during the study period and one, (4), was last contacted at the end of 1983.*

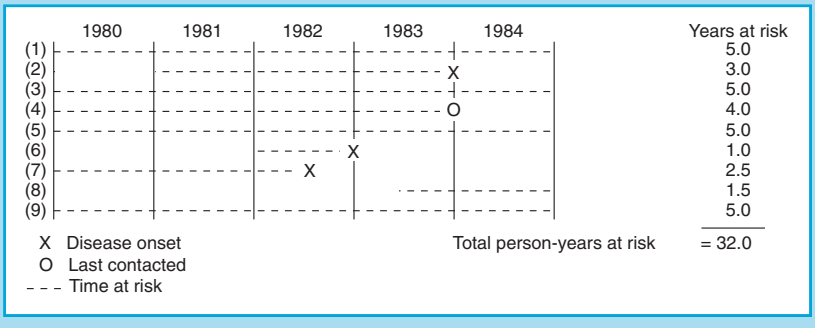


Figure 4.3.
Calculation of an individual's time at risk and total person-time at risk for the nine study subjects described in Example 4.6.

Example 4.6 illustrates the calculation of person-time at risk using a hypothetical group of nine persons. Subject (1) joined the study at the beginning of 1980 and was followed up throughout the study period. Therefore, (1) was at risk of becoming a case for the entire five years of the study. Subject (4) also joined at the beginning of the study, but was last contacted at the end of 1983; thus, (4) was at risk for only four years. Subject (6) joined the study at the beginning of 1982, and developed the disease by the end of that year; after that, (6) was no longer at risk (assuming there can be no recovery from the disease of interest). The total person-years at risk is the sum of all the individuals' time at risk.

The incidence rate accounts for differences in person-time at risk and is given by:

$$\text{Incidence rate} = \frac{\text{No. of new cases of disease arising in a defined population over a given time period}}{\text{Total person–time at risk during that period}}$$

This measure of disease frequency is also called incidence density or force of morbidity (or mortality). Like risk and odds, the numerator of the incidence rate is the number of new cases in the population. The denominator, however, is the sum of each individual's time at risk. In the above example, the incidence rate will be equal to:

3/32 = 0.094 per person-year or 9.4 per 100 person-years

When presenting an incidence rate, the time units must be specified; that is, whether the rate measures the number of cases per person-day, person-month, person-year, etc. Although the above definitions of risk, odds and rate are now widely accepted, the terms risk and rate are used interchangeably in much of the literature, and especially in older publications.

4.2.3 The relationship between prevalence, rate and risk

As stated in Section 4.2.2, prevalence depends on both the incidence and the duration of the disease. When both incidence and duration are stable and the prevalence of the disease is low (as with cancer), this association may be expressed as follows:

Prevalence = incidence rate × average duration of disease

Example 4.7 provides an illustration of the relationship between prevalence, incidence and duration of the disease.

Example 4.7. *A total of 50 new cases of a particular cancer are diagnosed every year in a population of 100 000 people. The average duration of (i.e., survival from) this cancer is four years. Thus, the prevalence of the cancer in that population is:*

Prevalence = 0.0005 per person-year × 4 years = 0.2%

Risk depends both on the incidence rate and on the duration of the at-risk period. It is also affected by mortality from diseases other than the disease of interest; some of those who died from other diseases would have been expected to develop the disease of interest had they survived. If mortality from other diseases is disregarded, and if the incidence rate is constant throughout the period at risk, the following relationship applies:

Risk = 1 – exp (– incidence rate × duration of the period at risk)

The symbol exp indicates that the mathematical constant e = 2.72 should be raised to the power of the expression in parentheses. For diseases that have a low incidence rate or when the period at risk is short, the following approximation may be used:

Risk = incidence rate × duration of the period at risk.

This is clearly illustrated in Example 4.8.

Example 4.8. *The incidence rate of a particular condition in a population is 50 per 100 000 person-years. The risk for an individual in this population of developing this condition during a five-year period (assuming no other causes of death) is given by:*

Five-year risk = 1 – exp (– 0.0005 per person-year × 5 years) = 0.0025 = 0.25%.

The same value can be obtained using the simplified formula:

Five-year risk = 0.0005 per person-year × 5 years = 0.0025 = 0.25%

Consider now a common condition with an incidence rate of 300 per 1000 person-years:

Five-year risk = 1 – exp (– 0.3 per person-year × 5 years) = 0.78 = 78%

In this instance, the simplified formula yields a meaningless result:

Five-year risk = 0.3 per person-year × 5 years = 1.5 = 150%

(As risk is a proportion, it can never have a value above 1, or 100%.)

4.3 Using routine data to measure disease occurrence

Rates can be estimated from routinely collected data (e.g., vital statistics data, cancer registration data), even though direct measures of the person-time at risk are not available (Example 4.9). An estimate of the person-time at risk during a given period can be made as follows:

> Population at the mid-point of the calendar period of interest × length of the period (in suitable units of time, usually years).

Provided that the population remains stable throughout this period, this method yields adequate estimates of person-time at risk.

Example 4.9. *Suppose that we wish to estimate the incidence of stomach cancer in men living in Cali, Colombia. Volume VI of* Cancer Incidence in Five Continents *(Parkin* et al.*, 1992) provides data on the total number of stomach cancer cases that occurred in Cali during the years 1982–86 and on the total male population in 1984. The incidence rate of stomach cancer can be calculated from these data as shown below:*

No. of male stomach cancer cases, Cali, 1982–86 = 655
Total male population, Cali, 1984 = 622 922
Total person-years at risk, 1982–86 = 5 (years) × 622 922 = 3 114 610 pyrs
Mean annual incidence rate, Cali, 1982–86 = 655/3 114 610 = 21.03 per 100 000 pyrs

Thus the mean annual incidence rate of stomach cancer in men living in Cali during the years 1982–86 was 21 per 100 000 pyrs.

This method of estimating person-time at risk is appropriate for rare conditions such as cancer. However, common conditions demand more sophisticated approaches that exclude from the denominator those who have the disease and are therefore no longer at risk.

In most developed countries and many developing countries, a population census is taken, usually once every ten years. This provides the baseline count of the total population. As a source of denominator data, censuses are somewhat limited: they are relatively incomplete for some population subgroups (e.g., homeless and nomadic people) and can rapidly become out of date. Most census offices provide estimates of the population size between censuses (for intercensal years), which are based on population birth, death and migration rates. When available, these annual population estimates can be taken as the best estimates of the person-time at risk in each calendar year. Thus in the above example, the sum of the annual population estimates for the years 1982–86 could have been used to provide an estimate of the total person-years at risk for the entire study period.

4.3.1 Crude and stratum-specific measures

The measures of disease occurrence discussed in Section 4.2 may be calculated for a whole population—so-called crude measures—or separately for specific sub-groups (strata) of the population—called stratum-specific measures. For example:

$$\text{Crude incidence rate per 100 000 pyrs} = \frac{\text{No. of new cases arising in a defined population in a specific period of time}}{\text{Total person - years at risk in that population during that period of time}} \times 100\,000$$

Crude rates are widely used, in part because they are summary measures and so are easily interpreted, and in part because their calculation requires relatively little information. Crude rates may obscure the fact that subgroups of the population have marked differences in incidence; for instance, people in different age groups have a different risk of death. This should be borne in mind when comparing crude rates from various populations, as disparities might reflect differences in their population structure rather than in disease incidence (see Section 4.3.3).

To gain an understanding of certain epidemiological aspects of a disease, more detailed rates, specific for sex and other demographic characteristics such as age, are needed. For example, age-specific rates can be calculated as follows:

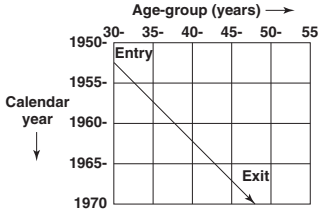
$$\text{Age-specific incidence rate per 100 000 pyrs} = \frac{\text{No. of new cases arising in a certain age-group in a defined population and in a specific period of time}}{\text{Person - years at risk in that age group in the same population and during that period of time}} \times 100\,000$$

Person-time at risk is calculated separately for each age group. Plotting these age-specific rates against age yields an age–incidence curve, which can reveal important clues to the epidemiology of a disease (see Figure 4.5a). Note that cancer rates are usually sex-specific, i.e., calculated separately for males and females, because cancer incidence for most sites differs markedly between the sexes.

4.3.2 Changes in disease incidence with time

The risk of getting a disease also changes with calendar time, and this should be taken into account during follow-up. This is illustrated in Example 4.10.

Example 4.10. *Consider a group of people (cohort) aged from 30 to 54 years who were followed up from 1950 to the end of 1969. Study subjects contributed person–time at risk from the time they joined the cohort to the end of the study in 1969 (or until their 55th birthday if it occurred earlier). The experience of one study subject is shown in Figure 4.4; this subject joined the cohort on 1 October 1952, on his 30th birthday, and was 47 years and 3 months old when the study ended on 31 December 1969.*



Calendar period	Age group	Person-months	Total person-months in calendar period and age stratum
1950-54	30-34	3 months in 1952 12 months in 1953 12 months in 1954	27
1955-59	30-34	12 months in 1955 12 months in 1956 9 months in 1957	33
	35-39	3 months in 1957 12 months in 1958 12 months in 1959	27
1960-64	35-39	12 months in 1960 12 months in 1961 9 months in 1962	33
	40-44	3 months in 1962 12 months in 1963 12 months in 1964	27
1965-69	40-44	12 months in 1965 12 months in 1966 9 months in 1967	33
	45-49	3 months in 1967 12 months in 1968 12 months in 1969	27

Figure 4.4.
Lexis diagram showing the follow-up of the study subject described in Example 4.10 and the calculation of his person-months contribution to each calendar period and age stratum.

The experience of a whole cohort can be represented in a *Lexis diagram*, which consists of age and calendar time cells or strata (see Figure 4.4). This diagram can be used to assess individual follow-up simultaneously in relation to two different time-scales: age and calendar period. Once a subject enters the cohort, he moves diagonally through the Lexis diagram as he ages, contributing person-time at risk to various strata as he moves through them. Stratum-specific rates can be calculated by dividing the total number of cases arising in each age and calendar period stratum by the corresponding total person-time at risk.

Even when data on date of birth are not available, the Lexis diagram can be used with routine data to describe the incidence of a disease in successive generations. Mortality rates from cancer of the lung in men in England and Wales during 1941–78 are shown in Table 4.1: columns show changes in the incidence rates with age, and rows show changes in the age-specific rates over calendar time. In any age × calendar-period two-way table, diagonal lines represent successive birth cohorts, although the earliest and the most recent birth cohorts (in the extremes of the table) will have very few data points.

Table 4.1.
Mortality (per million person-years) from cancer of the lung, in males in England and Wales, 1941–78, by age group and calendar period; the central year of birth is indicated on the diagonals.[a]

	Age-group (years)																	
Year of death	0-	5-	10-	15-	20-	25-	30-	35-	40-	45-	50-	55-	60-	65-	70-	75-	80-	85 and over
1941-45	1	1	2	4	9	17	39	87	203	404	626	924	1073	1031	799	685	396	
1946-50	1	2	1	4	12	18	41	97	242	555	972	1375	1749	1798	1442	1130	791	497
1951-55	1	0	1	2	8	14	37	100	250	584	1232	2018	2575	2945	2651	2087	1444	927
1956-60	1	0	0	2	5	13	36	94	253	594	1254	2326	3330	3941	3896	3345	2271	1438
1961-65	1	0	1	2	5	11	33	91	225	566	1226	2290	3673	4861	4994	4530	3423	2062
1966-70	0	0	0	3	4	11	25	76	218	532	1165	2208	3703	5281	6223	5931	4578	3490
1971-75	0	0	0	1	4	10	25	58	178	505	1074	2082	3552	5185	6834	7284	6097	4384
1976-78	0	1	0	1	3	8	18	60	146	422	1056	1907	3384	5061	6782	7982	7395	5285
Birth cohort (diagonal)							1951	1946	1941	1936	1931	1926	1921	1916	1911	1906	1901	1896	1891

1866
1871
1876
1881
1886

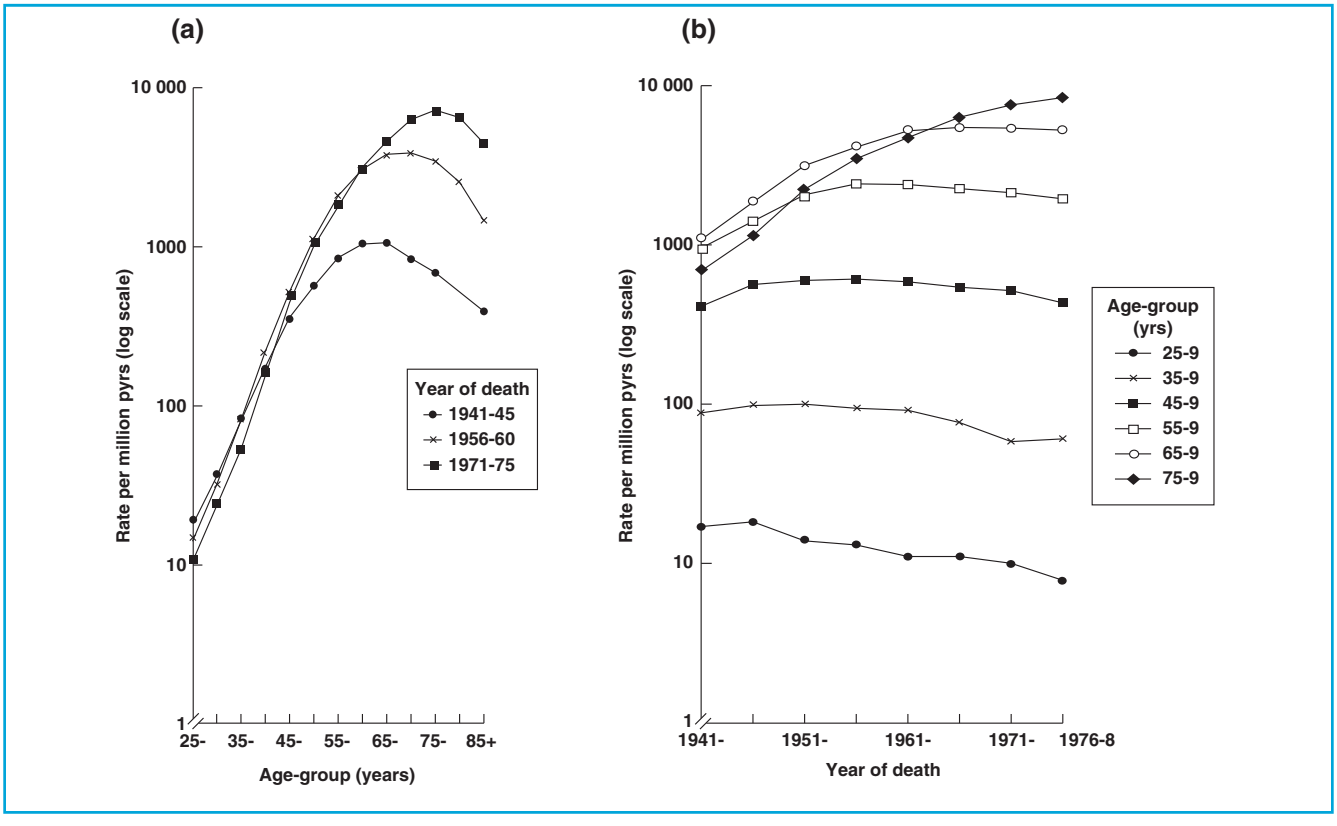
[a] Data from OPCS, (1981).

The diagonals of Table 4.1 (from upper left to lower right), for instance, define the lung cancer mortality experience for successive generations of men who were born together and hence aged together. For example, a man aged 40–44 years in 1941–45, was aged 45–49 in 1946–50, 50–54 in 1951–55, etc. To be 40–44 in 1941–45, he could have been born at any time between January 1896 (44 in 1941) and December 1905 (40 in 1945). These so-called birth cohorts are typically identified by their central year of birth; for example, the 1901 birth cohort, or more precisely the 1900/1 birth cohort, contains those men born during the 10-year period from 1896 to 1905. The diagonal just above this one shows the rates pertaining to the 1896 cohort, i.e., those men born between 1891 and 1900. As the years of birth for each cohort are estimated from the age and calendar period data, adjacent cohorts inevitably overlap, i.e., they have years of birth in common. When data on exact year of birth are available, these estimates need not be made, and so successive birth cohorts do not overlap.

Analyses by birth cohort thus use the same age-specific rates as in calendar time period analyses, but these rates are arranged in a different way. Comparison of rates in successive birth cohorts allows us to assess how incidence may have changed from one generation to another.

The data in Table 4.1 can be plotted in different ways to illustrate changes in age-specific rates over calendar time—*secular trends*—or changes from generation to generation—*cohort trends*.

Figure 4.5 clearly illustrates that secular trends in lung cancer mortality differ by age. In older age-groups, rates increased over the study period, while in younger groups, they declined. When rates are presented by year of birth (Figure 4.6), it becomes apparent that while rates for successive generations of men born until the turn of the century increased, they declined for generations born since then. These trends closely parallel trends in cigarette smoking (not shown).



In certain situations, cohort analysis gives the most accurate picture of changes in the patterns of disease over time, for example, if exposure to a potential risk factor occurs very early in life and influences the lifetime risk of a particular disease, or if the habits of adults are adopted by successive generations (as with cigarette smoking and lung cancer, and exposure to sunlight and malignant melanoma of skin). In other situations, secular analysis might be more appropriate: for example, if exposure to the risk factor affects all age groups simultaneously (as with the introduction of a new medical treatment). However, in most situations, it is not clear which analysis is most appropriate to describe temporal trends, and the results of both should be examined.

Figure 4.5.
Mortality from lung cancer in men in England and Wales, 1941-78. (*a*) Rates presented to show differences in age-specific curves between three selected calendar periods; (*b*) rates presented to show secular (calendar) trends in age-specific rates. For clarity, only rates for alternate age-groups are shown; the first five age-groups are omitted because of the small number of deaths (data from Table 4.1).

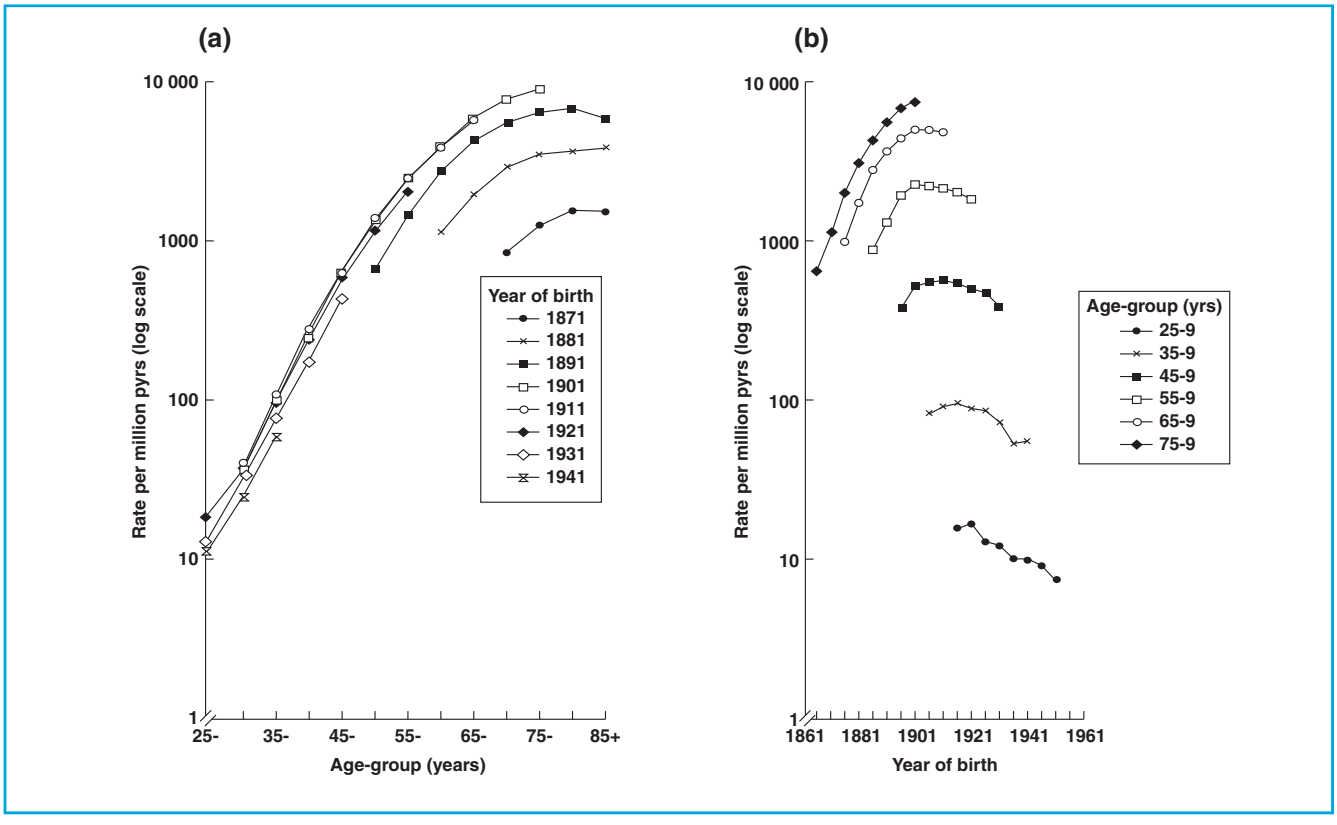


Figure 4.6.
Mortality from lung cancer in men in England and Wales, 1941–78. (*a*) Rates presented to show differences in age-specific curves for successive birth cohorts; (*b*) rates presented to show cohort trends in age-specific rates. For clarity, only rates for alternate age-groups or cohorts are shown; the first five age-groups are omitted because of the small number of deaths (data from Table 4.1).

Descriptive analyses by age, calendar time and cohort are a popular epidemiological tool for examining temporal changes in the incidence of a disease. These analyses are based on the inspection of tables and graphs, in much the same way as described here, although statistical models can also be used to assess whether there is a statistically significant trend in rates over calendar time, or between birth cohorts (Clayton & Schifflers, 1987a,b).

4.3.3 Controlling for age

For comparison of incidence between populations, crude rates may be misleading. As an example, let us compare stomach cancer incidence among men living in Cali, Colombia and Birmingham, England. The data are extracted from *Cancer Incidence in Five Continents* (Parkin *et al.*, 1992).

Table 4.2 shows that the crude incidence rate (the rate for all ages combined) for Birmingham was much higher than that for Cali. However, before concluding that the incidence of male stomach cancer in Birmingham (1983–86) was higher than in Cali (1982–86), the age-specific rates for the two must be compared. Surprisingly, age-specific rates were higher for Cali in all age-groups. The discrepancy between crude and age-specific rates is because these two populations had markedly different age-structures (Table 4.3), with Birmingham having a much older population than Cali.

Age (years)[b]	Cali			Birmingham		
	No. of cancers (1982–86)	Male population (1984)	Mean annual rate[c] (1982–86)	No. of cancers (1983–86)	Male population (1985)	Mean annual rate[c] (1983–86)
0–44	39	524 220	1.5	79	1 683 600	1.2
45–64	266	76 304	69.7	1037	581 500	44.6
65+	315	22 398	281.3	2352	291 100	202.0
All ages	620	622 922	19.9[d]	3468	2 556 200	33.9

[a] Data from Parkin *et al.* (1992)
[b] For simplicity, only three broad age-groups are used throughout this example.
[c] Rate per 100 000 person-years.
[d] This crude rate is slightly lower than in Example 4.9 (21.03 per 100 000 person-years) because cases of unknown age (35 in total) were excluded here. The exclusion of two cases of unknown age in Birmingham did not affect the value of the crude rate calculated here.

Table 4.2.
Incidence of stomach cancer in males by age group in Cali, 1982–86, and Birmingham, 1983–86.[a]

The lower crude rate for Cali is thus explained by its male population being younger than that of Birmingham, and the fact that younger people have a much lower incidence of stomach cancer than older people (Figure 4.7). In this situation, age is a confounding variable, i.e., age is related to exposure (locality) and it is itself a risk factor for the outcome of interest, stomach cancer (see also Chapter 13).

Age (years)	Percentage of total male population	
	Cali (1984)	Birmingham (1985)
0–44	84	66
45–64	12	23
65+	4	11
All ages	100	100

[a] Data from Parkin *et al.* (1992).

Table 4.3.
Age distribution of the male population in Cali, 1984, and Birmingham, 1985.[a]

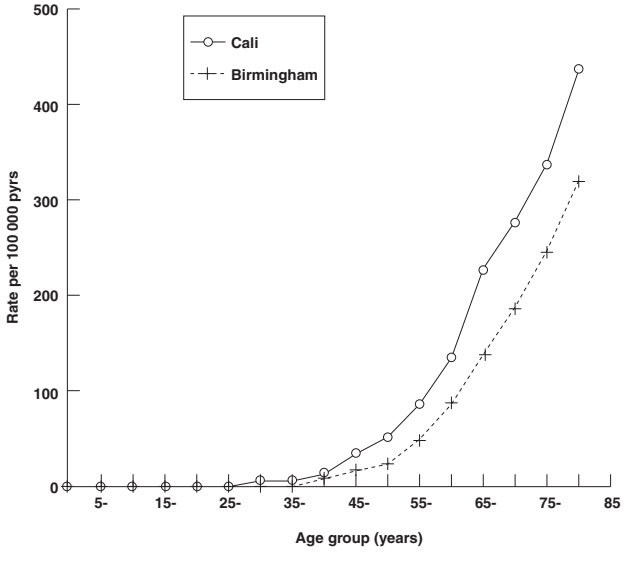


Figure 4.7.
Age-incidence curve of stomach cancer in males in Cali, 1982-86, and Birmingham, 1983-86 (data from Parkin *et al.*, 1992).

As incidence rates for stomach cancer change considerably with age, differences in the age distribution of populations need to be considered before attempting to compare incidence. One approach is to compare age-specific rates, as in the example above; however, this can become cumbersome when comparing several populations each with many age-groups. The ideal would be to have a summary measure for each population, which has been controlled, or adjusted, for differences in the age structure. Several statistical methods can be used to control for the effects of confounding variables, such as age (see also Chapters 13 and 14). Here, only one such method, *standardization*, is discussed.

Standardization is by far the most common method used when working with routine data. Although this method is usually employed to adjust

for the effect of age, it can equally be used to control for any other confounding variable such as social class, area of residence, etc. There are two methods of standardization: direct and indirect.

Age (years)	Population
0–44	74 000
45–64	19 000
65+	7 000
All ages	100 000

Table 4.4.
A standard population.

Direct method of standardization

Let us take a hypothetical population, and call it our standard population, the age-structure of which is shown in Table 4.4. How many cases of stomach cancer would we expect in males in Cali if its male population had the same age distribution as this standard population?

As shown in Figure 4.8, this is relatively easy to calculate. Each age-specific rate for Cali is simply multiplied by the standard population figures in the corresponding age-group; the sum over all age categories will give the total number of male stomach cancer cases expected in Cali if its male population had the same age distribution as the standard.

It is also possible to determine how many male stomach cancer cases would be expected in Birmingham if its male population had the same age distribution as the standard population; the calculations are similar to those described for Cali, but are based on the age-specific rates for Birmingham (Figure 4.8).

Summary incidence rates for Cali and Birmingham, assuming the age-structure of the standard population, can be obtained by dividing the total expected cases by the total person-years at risk in the standard population. These rates are called mean annual age-adjusted or age-standardized incidence rates; they can be seen as the crude incidence rates that these populations would have if their age distributions were shifted from their actual values in the mid-1980s to the age distribution of the standard population. These standardized rates are a fiction: they are not the stomach cancer incidence rates that actually existed, but rather those that these two populations would have had if, while retaining their own age-specific rates, they had a hypothetical (standard) population. The fiction is useful, however, because it enables the epidemiologist to make summary comparisons between populations from different areas, or during different time periods, which are free from the distortion that arises from age differences in the actual populations.

Figure 4.8.
The direct method of standardization.

Mean annual age-specific rates in Cali, 1982-86 (per 100 000 pyrs)

Age	Rate
0–44	1.5
45–64	69.7
65+	281.3

Mean annual age-specific rates in Birmingham, 1983-86 (per 100 000 pyrs)

Age	Rate
0–44	1.2
45–64	44.6
65+	202.0

Standard Population

Age	Population
0–44	74 000
45–64	19 000
65+	7 000

No. of male stomach cancer cases expected if the male population of Cali and Birmingham had the same age distribution as the standard population

a) Cali

Age	Expected cases
0–44	0.000015 × 74 000 = 1.11
45–64	0.000697 × 19 000 = 13.24
65+	0.002813 × 7 000 = 19.69
Total expected	**= 34.04**

b) Birmingham

Age	Expected cases
0–44	0.000012 × 74 000 = 0.89
45–64	0.000446 × 19 000 = 8.47
65+	0.002020 × 7 000 = 14.14
Total expected	**= 23.50**

Mean annual age-adjusted rate for Cali, 1982-86 =
= 34.04/100 000 =
= 34.0 per 100 000 pyrs

Mean annual age-adjusted rate for Birmingham, 1983-86 =
= 23.5/100 000 =
= 23.5 per 100 000 pyrs

The age-standardized rate can be seen as a weighted average of the age-specific rates, the weights being taken from the standard population. Age-adjusted rates can be compared directly, provided that they refer to the same standard population, i.e., that the weights given to the age-specific rates are the same. In the example above, the mean annual age-standardized incidence rate for Cali is higher than that for Birmingham; this is in agreement with the age-specific rates. An age-standardized rate ratio can be calculated by dividing the rate for Cali by that for Birmingham, to yield a rate ratio:

34.0 per 100 000 pyrs/23.5 per 100 000 pyrs = 1.45

This measure is called the *standardized rate ratio (SRR)* or *comparative morbidity (or mortality) figure (CMF)*. In this example, it reveals that the estimated incidence of stomach cancer was 45% higher in Cali than in Birmingham in the mid-1980s, and that this excess is independent of age differences between these two populations.

This method of adjusting for age is called the direct method of standardization. It requires knowledge of the age-specific rates (or the data to calculate them) for all the populations being studied, and also the definition of a standard population. The standard population can be any population: one of those being compared or any other. However, the standard population used must always be specified, as its choice may affect the comparison. Conventional standard populations, such as the world standard and the European standard populations, have been defined and are widely used so as to allow rates to be compared directly (see Appendix 4.1). The standard population given in Table 4.4 is in fact a summary of the world standard population.

For simplicity, we use only three broad age-groups in the example given in this section. However, this does not provide an adequate age-adjustment, and narrower age groups should be used. Five-year age-groups are usually employed, as they are the most common grouping in publications on site-specific cancer data. When five-year age-groups are used for age-adjustment of the data on stomach cancer presented in the example above (Figure 4.8), the age-adjusted rates per 100 000 person-years are 36.3 for Cali and 21.2 for Birmingham; the rate ratio is now 1.71. When rates change dramatically with age, narrower age groups (e.g., one-year groups) may be required to obtain an adequate age-adjustment.

It is important to remember that an age-standardized rate is not an actual rate but rather an artificial one, which permits the incidence of a disease in one population to be compared with that in another, controlling for differences in their age composition. Therefore, age-standardized rates should not be used when what is needed is an accurate measurement of disease occurrence in a population, rather than a comparison.

Indirect method of standardization

Suppose that the total number of stomach cancers in Cali in 1982–86 is known, but their distribution by age is not available (Table 4.5). In this case, the direct method of standardization cannot be used.

Table 4.5.
Incidence of stomach cancer in Cali, 1982–86, and Birmingham, 1983–86.[a]

	Cali			Birmingham		
Age (years)[b]	**No. of cancers (1982–86)**	**Male population (1984)**	**Mean annual rate[b] (1982–86)**	**No. of cancers (1983–86)**	**Male population (1985)**	**Mean annual rate[b] (1983–86)**
0–44	NA	524 220	–	79	1 683 600	1.2
45–64	NA	76 304	–	1037	581 500	44.6
65+	NA	22 398	–	2352	291 100	202.0
All ages	620	622 922	19.9	3468	2 556 200	33.9

NA, data assumed to be not available (see Table 4.2).
[a] Data from Parkin *et al.*, 1992.
[b] Rate per 100 000 person-years.

It is, however, possible to calculate how many male cases of stomach cancer would be expected in Cali if males in both Cali and Birmingham had the same age-specific incidence rates. In other words, the Birmingham age-specific rates can be treated as a set of standard rates. The calculations are shown in Figure 4.9. The expected number of cancer cases in Cali is calculated by multiplying the mean annual age-specific rates for Birmingham by the person-years at risk in the corresponding age-group in Cali; the sum over all age categories will give the total number of male cancer cases that would be expected in Cali if its male population had the same age-specific incidence rates for stomach cancer as that of Birmingham. Evidently, the number of expected cases in Birmingham is equal to the number observed.

Note that these expected stomach cancer cases relate to what would happen if Cali and Birmingham had the same age-specific incidence rates for stomach cancer rather than the same population structure. So it would be meaningless to calculate summary rates for each locality by dividing the total number of expected cases by the corresponding total person-years at risk. However, for

Figure 4.9.
The indirect method of standardization.

Mean annual age-specific rates in Birmingham, 1983-86 (per 100 000 pyrs)

Age	Rate
0–44	1.2
45–64	44.6
65+	202.0

Total person-years at risk in Cali, 1982-86

Age	Person-years
0–44	524 220 × 5 = 2 621 100
45–64	76 304 × 5 = 381 520
65 +	22 398 × 5 = 111 990
All ages	**= 3 114 610**

Total person-years at risk in Birmingham, 1983-86

Age	Person-years
0–44	1 683 600 × 4 = 6 734 400
45–64	581 500 × 4 = 2 326 000
65 +	291 100 × 4 = 1 164 400
All ages	**= 10 224 800**

No. of expected male stomach cancer cases if the populations have the same stomach cancer age-specific incidence rates as Birmingham

a) Cali

Age	Expected cases
0–44	0.000012 × 2 621 100 = 31.45
45–64	0.000446 × 381 520 = 170.15
65+	0.002020 × 111 990 = 226.22

Total expected (E), 1982-86 427.82
Total observed (O), 1982-86 620
O/E (%) = 145

b) Birmingham

Age	Expected cases
0–44	0.000012 × 6734400 = 79
45–64	0.000446 × 2326000 = 1037
65+	0.002020 × 11644400 = 2352

Total expected (E), 1983-86 = 3 468
Total observed (O), 1983-86 = 3 468
O/E (%) = 100

each locality, the numbers of cases observed and expected can be compared, because both refer to the same population. The ratio of the observed number of cases to that expected is called the standardized incidence ratio (SIR) or the standardized mortality ratio (SMR) if a case is defined as death. These ratios are usually expressed as a percentage.

In the example above, the SIR (%) for Birmingham is 100; by definition, the number of observed cases of stomach cancer is equal to the number of expected cases when using the age-specific stomach cancer incidence rates for Birmingham as the standard rates. The SIR (%) for Cali is 145, meaning that the number of cases observed was 45% higher than that expected if Cali had the same incidence of stomach cancer as Birmingham. This result is similar to that obtained using the direct method of standardization.

This method is called the indirect method of standardization. As with the direct method, the results depend in part upon the standard chosen. However, the indirect method of standardization is less sensitive to the choice of standard than the direct one.

Which method is the best?

In comparisons of incidence of disease between two or more populations, direct and indirect standardization tend to give broadly similar results in practice. However, the choice of method might be affected by several considerations:

(1) The direct method requires that stratum-specific rates (e.g., age-specific rates) are available for all populations studied. The indirect method requires only the total number of cases that occurred in each study population. If stratum-specific rates are not available for all study populations, the indirect method may be the only possible approach.

(2) Indirect standardization is preferable to the direct method when age-specific rates are based on small numbers of subjects. Rates used in direct adjustment would thus be open to substantial sampling variation (see Section 6.1.4). With the indirect method, the most stable rates can be chosen as the standard, so as to ensure that the summary rates are as precise as possible.

(3) In general, when comparing incidence in two or more populations, direct standardization is open to less bias than indirect standardization. The reasons for this are subtle and are beyond the scope of this text.

For a more detailed discussion of the advantages and disadvantages of each method of standardization, see pp. 72–76 in Breslow & Day (1987).

Is the use of adjusted summary measures always appropriate?

Although age-adjusted measures provide a convenient summary of age-specific rates, the age-specific rates themselves give the most infor-

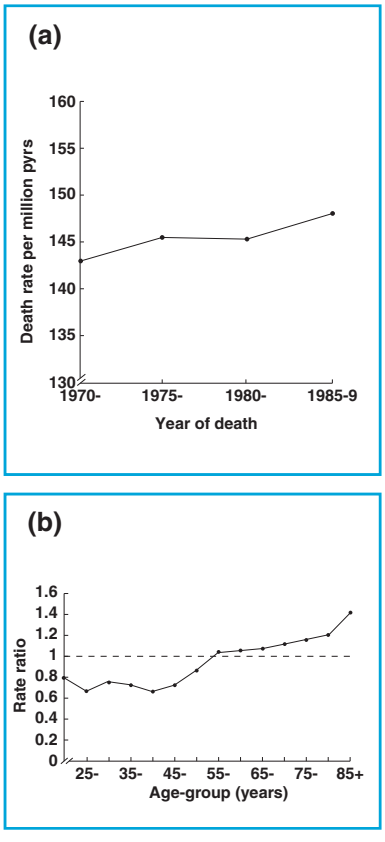


Figure 4.10.
Ovarian cancer mortality in England and Wales, 1970–89. (a) Rates are age-standardized to the 1981 female population of England and Wales; (b) age-specific mortality rate ratios, 1970–74 and 1985–89 (rates in 1970–74 taken as the baseline).

mation. It must be emphasized that, under certain circumstances, it may not be appropriate to summarize disease rates in a single summary measure. Consider this example. We wish to monitor trends in mortality from ovarian cancer in England and Wales. The age-adjusted death rates for this cancer in England and Wales increased slightly from 1970–74 to 1985–89, as shown in Figure 4.10a. However, trends in the age-specific rates for this period reveal that they did not increase across all age-groups. Figure 4.10b shows the rate ratio of the 1985–89 to the 1970–74 age-specific rates. It becomes apparent that while death rates were increasing at older ages (rate ratios above 1), there was no increase in women below age 55 years (rate ratios below 1). If age-standardized death rates for all ages are calculated, this information is lost, because mortality rates from this cancer at younger ages were so low that they were dominated by the much higher mortality rates in the older age-groups. So in this case, the age-adjusted summary measure is misleading.

Before age-adjusted summary measures are calculated, the age-specific rate ratios should always be examined to determine whether this approach is appropriate. If these ratios vary systematically with age, this information would inevitably be lost in the summary age-adjusted measure.

4.3.4 Cumulative rate

The cumulative rate is another measure of disease occurrence that is increasingly used in cancer epidemiology. This measure has been included in recent editions of *Cancer Incidence in Five Continents* (see, for example, Parkin *et al.*, 1992).

A cumulative rate is the sum of the age-specific incidence rates over a certain age range. The age range over which the rate is accumulated must be specified, and depends on the comparison being made. Thus for childhood tumours, this might be age 0–14 years. In general, however, the most appropriate measure is calculated over the whole life span, usually taken as 0–74 years.

The cumulative rate can be calculated by the sum of the age-specific incidence rates (provided they are expressed in the same person-time units, e.g., 100 000 pyrs), multiplied by the width of the age-group. Thus for five-year age-groups, the cumulative rate would be five times the sum of the age-specific incidence rates over the relevant age range (Figure 4.11). If the age-groups are of different width, each age-specific rate should be first multiplied by the width of the corresponding age-group; the sum over all age-categories yields the cumulative rate. This measure is usually expressed as a percentage.

The cumulative rate can be interpreted as a form of direct age-standardization with the same population size (i.e., denominator) in each age-group. Thus, it avoids the arbitrary choice of a standard population.

Another advantage of the cumulative rate is that it provides an estimate of cumulative risk, i.e., the risk an individual would have of devel-

Age-group (years)	Mean annual age-specific incidence rate (per 100 000 pyrs)
0–4	0
5–9	0
10–14	0
15–19	0
20–24	0.1
25–29	0.1
30–34	0.9
35–39	3.5
40–44	6.7
45–49	14.5
50–54	26.8
55–59	52.6
60–64	87.2
65–69	141.7
70–74	190.8
Total	**524.9**
Total x 5	**2624.5**

Cumulative rate = 100 x (2624.5/100 000) = 2.6%
Cumulative risk = 100 x {1–exp(–cumulative rate/100)} = 2.6%

Figure 4.11.
Calculation of cumulative rate and cumulative risk over the age range 0–74 years for male stomach cancer in Birmingham, 1983-86 (data from Parkin *et al.*, 1992).

oping a particular cancer over a defined life span in the absence of any other cause of death. The cumulative risk can be calculated as follows:

$$\text{Cumulative risk} = 100 \times \{1–\exp(–\text{cumulative rate}/100)\}$$

However, if the cumulative rate is lower than 10%, its value is practically equal to that of the cumulative risk. Thus, in Figure 4.11, the estimated risk for a Birmingham male of developing stomach cancer between the ages of 0–74 years is 2.6% (assuming no other cause of death). This is equal to the cumulative rate.

Table 4.6 shows the crude rates, five-year age-standardized rates and cumulative rates for male stomach cancer in Cali and Birmingham. In contrast to the crude rates, both age-standardized rates and cumulative rates give an accurate picture of the relative incidence of stomach cancer in the two populations.

4.3.5 Lack of proper denominators

Sometimes, no suitable denominator is available to permit calculation of one of the measures of incidence discussed so far. This may be because there are no data on denominators (e.g., no census has been carried out), because a catchment population cannot be defined (e.g., for a hospital-based registry), or because case-finding has been so incomplete that denominators derived from other sources (e.g., the census) are not comparable with the numerator data. In these circumstances, it is traditional

Table 4.6.
Mean annual crude incidence rates, mean annual age-standardized rates and cumulative rates for male stomach cancer, Cali, 1982–86, and Birmingham, 1983–86.

	Cali, 1982–86	Birmingham, 1983–86	Rate ratio
Crude rates (per 100 000 pyrs)	19.9	33.9	0.59
Age-standardized rate[a] (per 100 000 pyrs)	36.3	21.2	1.71
Cumulative rate, 0–74 years (%)	4.6	2.6	1.77

[a] Standardized to the world standard population. These age-standardized rates differ slightly from those in Figure 4.8, being age-adjusted within five-year age-groups.

to calculate proportional measures; that is, to express the number of cases of a particular condition as a proportion of the total number of cases of all conditions:

$$\text{Proportional incidence (\%)} = \frac{\text{No. of cases of the disease of interest in a specified time period}}{\text{Total number of cases of all conditions in the same time period}} \times 100$$

Comparisons of incidence between populations can then be made by calculating proportional incidence ratios (PIRs); likewise, mortality can be compared by using mortality data to calculate proportional mortality ratios (PMRs). These ratios are calculated as follows:

$$\text{PIR (\%)} = \frac{\text{Proportion of cases from a specific cause in population A}}{\text{Proportion of cases from the same cause in population B}} \times 100$$

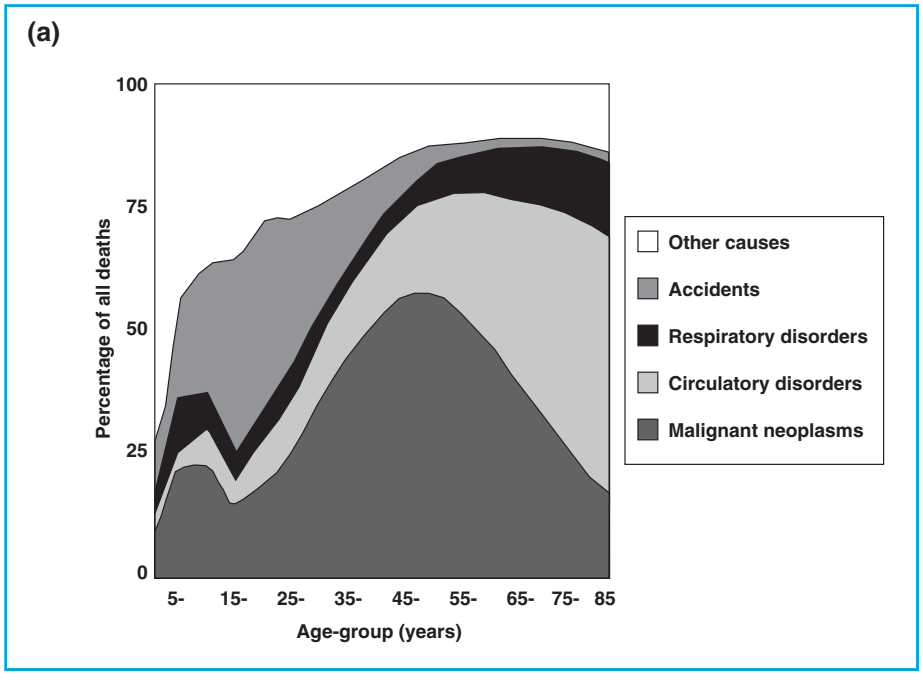
As with rates, these proportions can be standardize for age (or any other potential confounding factor).

Note that a proportional measure is not equivalent to a rate, as the denominator is derived from the total number of cases, and not from the population at risk. While proportional measures reveal the proportion of cases (or deaths) that can be attributed to a particular disease, a cause-specific rate reflects the risk of developing (or dying from) a particular disease for members of a specific population.

Proportional measures can be misleading because their denominator is the total number of cases (or deaths), a measure that depends on the number of cases (or deaths) from all causes, not just that being studied. For example, although the proportion of deaths due to cancer is greater in middle-aged women than in elderly women, death rates from cancer are actually higher among the elderly (Figure 4.12). This is because the total

number of deaths from other causes, particularly from cardiovascular disease, is also considerably higher in the elderly. Thus, although the total number of deaths from cancer is greater in the elderly, they constitute a smaller proportion of all deaths than at younger ages.

These measures (and the use of odds ratios as an alternative) are discussed further in Chapter 11.



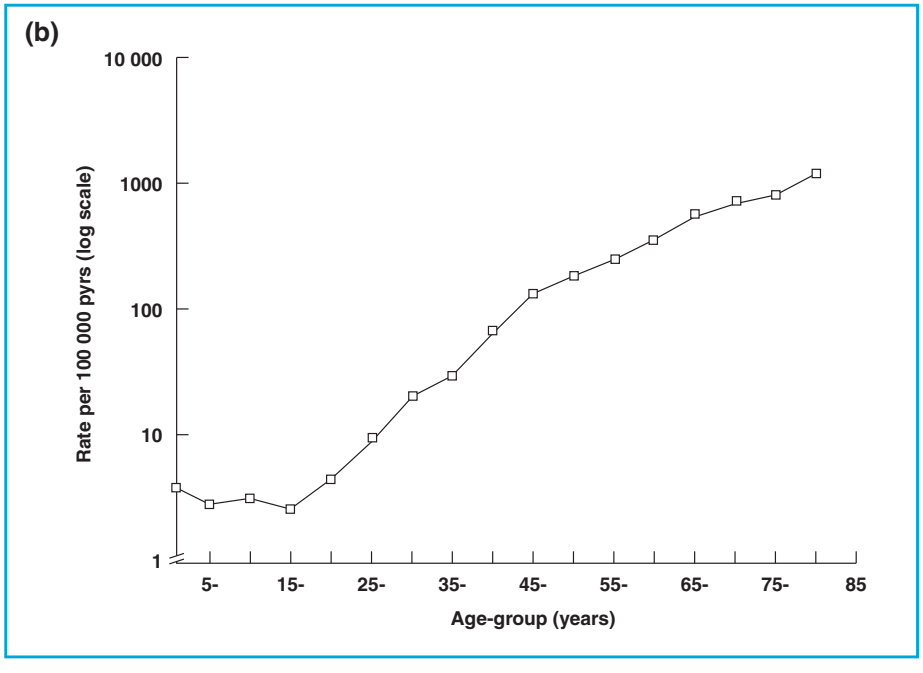


Figure 4.12.
Female deaths in England and Wales, 1993. (a) Proportion of deaths due to cancer and other causes by age; (b) cancer mortality rates by age (data from OPCS, 1995).

Further reading

* Most of the measures of disease occurrence discussed here are dealt with in more detail in Breslow & Day (1987) and Estève *et al.* (1994).

* A more elaborate discussion of age, calendar time and cohort effects can be found in Clayton & Schifflers (1987a,b).

Box 4.1. Key issues

- Quantification of the occurrence of disease and other health-related events in populations requires a clear definition of the cases (numerator), the population at risk (denominator), and the time frame to which these refer.

- There are two major measures of disease occurrence: *prevalence* and *incidence*. Prevalence refers to the total number of existing (new and old) cases of a condition in a population at a specific point in time. Incidence refers to the occurrence of new cases in a population over a specific time period.

- Incidence can be measured as either *risk*, *odds of disease*, or *rate*. The calculation of risk and odds requires complete follow-up of all study subjects for the entire study period. In contrast, the calculation of rates takes into account individual differences in length of follow-up.

- Measures of disease occurrence can be calculated for the whole population, as *crude* measures; or separately for certain subgroups of the population, as *stratum-specific* measures.

- Incidence of a disease varies with time. These changes occur simultaneously according to three different time scales: age, calendar period (*secular trends*) and date of birth (*cohort trends*), but can be examined separately in an age-by-calendar period two-way Lexis diagram. The diagonals in this diagram represent successive birth cohorts.

- Crude rates can be misleading when comparing incidence from different populations because they do not take into account differences in population age-structure. *Age-standardization*, either direct or indirect, can be used to obtain summary measures that are adjusted for differences in age structure. Alternatively, *cumulative rates* may be calculated.

- Proportional measures, such as *proportional incidence ratios*, can be calculated when no suitable denominators are available. These ratios should be interpreted cautiously, as their denominator is the total number of cases, not the population at risk.

Appendix 4.1

Conventional standard populations

The choice of the standard population to be used in the direct method of standardization is, to a certain extent, arbitrary. For example, if the aim is to compare disease occurrence in several groups in England and Wales, an appropriate standard might be the adult population of England and Wales. On the other hand, this may not be an appropriate standard when making comparisons between countries.

For international comparisons, various conventional standard populations have been used (Table A4.1). These standard populations range from an African population with a low proportion of old people, through an intermediate world population, to a European population with a high proportion of old people. In the earliest volumes of *Cancer Incidence in Five Continents*, rates were standardized to these three populations; however, the European and African standards were dropped in Volume IV (Waterhouse *et al.*, 1982) and replaced by cumulative rates over the age ranges 0–64 and 0–74 years.

The truncated population is derived from the world population but comprises only the middle age-groups. This truncated population was often used in the past, because data for older age-groups were likely to be less reliable than those for the middle age-groups, and because for most forms of cancer, virtually no cases arise in groups under 35 years.

Table A4.1.
Conventional standard populations used for international comparisons.

Age group (years)	African	World	European	Truncated
0	2 000	2 400	1 600	–
1–4	8 000	9 600	6 400	–
5–9	10 000	10 000	7 000	–
10–14	10 000	9 000	7 000	–
15–19	10 000	9 000	7 000	–
20–24	10 000	8 000	7 000	–
25–29	10 000	8 000	7 000	–
30–34	10 000	6 000	7 000	–
35–39	10 000	6 000	7 000	6 000
40–44	5 000	6 000	7 000	6 000
45–49	5 000	6 000	7 000	6 000
50–54	3 000	5 000	7 000	5 000
55–59	2 000	4 000	6 000	4 000
60–64	2 000	4 000	5 000	4 000
65–69	1 000	3 000	4 000	–
70–74	1 000	2 000	3 000	–
75–79	500	1 000	2 000	–
80–84	300	500	1 000	–
85+	200	500	1 000	–
Total	**100 000**	**100 000**	**100 000**	**31 000**