Chapter 11. Statistical methods for registries

P. Boyle and D.M. Parkin

International Agency for Research on Cancer,
150 cours Albert Thomas, 69372 Lyon Cédex 08, France

This chapter is not intended to replace statistical reference books. Its objective is solely to assist those involved in cancer registration to understand the calculations necessary for the presentation of their data. For population-based registries this will be as incidence rates. The methods required for using these rates in comparative studies—for example, comparing incidence rates from different time periods or from different geographical areas—are also described. Where incidence rates cannot be calculated, registry results must be presented as proportions, and analogous methods for such registries are also included.

PART I. METHODS FOR THE STUDY OF INCIDENCE

Definitions

The incidence rate

The major concern of population-based cancer registries will be the calculation of cancer incidence rates and their use to study the risk of individual cancers in the registry area compared to elsewhere, or to compare different subgroups of the population within the registry area itself (see Chapter 3).

Incidence expresses the number of new cases of cancer which occur in a defined population of disease-free individuals, and the incidence rate is the number of such events in a specified period of time. Thus:

$$\text{Incidence rate} = \frac{\text{Number of new cases of disease}}{\text{Population at risk}} \text{ in a period of time}$$

This measure provides a direct estimate of the probability or risk of illness, and is of fundamental importance in epidemiological studies.

Since incidence rates relate to a period of time, it is necessary to define the exact date of onset of a new case of disease. For the cancer registry this is the incidence date (Chapter 6, item 16). Although this does not correspond to the actual time of onset of a cancer, other possibilities are less easy to define in a consistent manner—for example, the date of onset of symptoms, date of entry to hospital, or the date of treatment.

Period of observation

The true instantaneous risk of disease is given by the incidence rate for an

infinitely short time period, the 'instantaneous' rate or 'force of morbidity'. With longer time periods the population-at-risk becomes less clearly defined (owing to births, deaths, migrations), and the rate itself may be varying with time. In practice, cancer in human populations is a relatively rare event and to study it quite large populations must be observed over a period of several years. Incidence rates are conventionally expressed in terms of annual rates (i.e., per year), and when data are collected over several years the denominator is converted to an estimate of person-years of observation.

Population at risk

In epidemiological cohort studies, relatively small populations of individuals on whom information has been collected about the presence or absence of risk factors are followed up. There will inevitably be withdrawal of individuals from the group under study (owing to death, migration, inability to trace), and often new individuals will be added to the cohort.

The result is that individuals are under observation and at risk of disease for varying periods of time; the denominator for the incidence rate is thus calculated by summing for each individual the person-years which are contributed.

Cancer registries are usually involved in calculating incidence rates for entire populations, and the denominator for such rates cannot be derived from a knowledge of each individual's contribution to the population at risk. This is therefore generally approximated by the mid-year population (or the average of the population at the beginning and end of the year or period), which is obtained from a census department. The variance of the estimate of the incidence rate is determined by the number of cases used in the numerator of the rate; for this reason it is usual to accumulate several years of observation, and to calculate the average annual rate. The denominator in such cases is again estimated as person-years, ideally by summing up the mid-year population estimates for each of the years under consideration. When these are unavailable, the less accurate solution of using the population size from one or two points during the time period to estimate person-years has to be used, an approximation that is likely to be reasonable providing no rapid or irregular changes in population structure are taking place. Examples, illustrating estimates of person-years of observation with differing availabilities of population data, are shown in Table 1. Conventionally, incidence rates of cancer are expressed as cases per 100 000 person-years, since this avoids the use of small decimals. For childhood cancers, the rate is often expressed per million.

When population estimates are used to approximate person-years at risk, the denominator of the rate will include a few persons who are not truly at risk. Fortunately for the study of incidence rates of particular cancers, this makes little difference, since the number of persons in the population who are alive and already have a cancer of a specific site is relatively small. However, if a substantial part of the population is genuinely not at risk of the disease, it should be excluded from the denominator. An obvious example is to exclude the opposite sex from the denominator of rates for sex-specific cancers, and incidence rates for uterine cancer

Table 1. Calculation of person-years at risk, with different availabilities of population data using data for the age group 45–49 for males in Scotland from 1980 to 1984

Year	1. Each year[a]	2. Mid-point[b]	3. Irregular points[c]
1980	140 800	—	—
1981	142 700	—	142 700
1982	140 600	140 600	—
1983	141 200	—	141 200
1984	141 500	—	—

[a] Method 1. Person-years = 140 800 + 142 700 + 140 600 + 141 200 + 141 500 = 706 800
[b] Method 2. Person-years = 140 600 × 5 = 703 000
[c] Method 3. Decrease in population, year 2 to year 4 = 1500; annual decrease = 1500/2 = 750; person-years = (142 700 + 750) + 142 700 + (142 700−750) + 141 200 + (141 200−750) = 709 750

are better calculated only for women with a uterus (quite a large proportion of middle-aged women may have had a hysterectomy)—especially when comparisons are being made for different time periods or different locations where the frequency of hysterectomy may vary (Lyon & Gardner, 1977; Parkin *et al.*, 1985a).

Calculation of rates

Many indices have been developed to express disease occurrence in a community. These have been clearly outlined by Inskip and her colleagues (Inskip *et al.*, 1983) and other sources of information also provide good discussions of this subject (Armitage, 1971; Armitage & Berry, 1987; Breslow & Day, 1980, 1987; Doll & Cook, 1967; Fleiss, 1981; MacMahon & Pugh, 1970). This chapter will concentrate on those methods which are generally most appropriate for cancer registration workers and will provide illustrative, worked examples. Whenever possible the example has been based on incidence data on lung cancer in males in Scotland. While an attempt has been made to enter results of as many of the intermediate steps on the calculation as possible, it has not been feasible to enter them all. Also, repetition of some of the intermediate steps may produce slightly different results owing to different degrees of precision used in the calculations and rounding. Thus the reader who attempts all the recalculations should get the same final result but should expect some minor imprecision in the intermediate results presented in the text.

Crude (all-ages) and age-specific rates

Suppose that there are A age groups for which the number of cases and the corresponding person-years of risk can be assessed. Frequently, the number of groups is 18 ($A = 18$) and the categories used are 0–4, 5–9, 10–14, 15–19 . . . 80–84 and 85 and over (85+). However, variations of classification are often used, for example by separating children aged less than one year (0) from those aged between 1 and 4 (1–4) or by curtailing age classification at 75, i.e., having age classes up to 70–74 and 75+.

Let us denote by r_i to be the number of cases which have occurred in the ith age class. If all cases are of known age, then the total number of cases R can be written as

$$R = \sum_{i=1}^{A} r_i = r_1 + r_2 + r_3 + \cdots + r_A \tag{11.1}$$

Similarly, denoting by n_i the person-years of observation in the ith age class during the same period of time as cases were counted, the total person-years of observation N can be written as

$$N = \sum_{i=1}^{A} n_i = n_1 + n_2 + n_3 + \cdots + n_A \tag{11.2}$$

The crude, all-ages rate per 100 000 can be easily calculated by dividing the total number of cases (R) by the total number of person-years of observation (N) and multiplying the result by 100 000.

$$\text{Crude rate} = C = \frac{R}{N} \times 100\,000 \tag{11.3}$$

i.e., when all cases are of known age,

$$C = \frac{\sum_{i=1}^{A} r_i}{\sum_{i=1}^{A} n_i} \times 100\,000 \tag{11.4}$$

The age-specific rate for age class i, which we denote as a_i, can also be simply calculated as a rate per 100 000 by dividing the number of cases in the age-class (r_i) by the corresponding person-years of observation (n_i) and multiplying the result by 100 000. Thus,

$$a_i = \frac{r_i}{n_i} \times 100\,000 \tag{11.5}$$

Age-standardization—general

One of the most frequently occurring problems in cancer epidemiology involves comparison of incidence rates for a particular cancer between two different populations, or for the same population over time. Comparison of simple crude rates can frequently give a false picture because of differences in the age structure of the populations to be compared. If one population is on average younger than the other, then even if the age-specific rates were the same in both populations, more cases would appear in the older population than in the younger. Notice from Table 2 how quickly the age-specific rates increase with age.

Thus, when comparing cancer levels between two areas, or when investigating the pattern of cancer over time for the same area, it is important to allow for the changing

Example 1. Calculation of crude and age-specific rates

Table 2 presents data on the incidence of cancer of the trachea, bronchus and lung (International Classification of Diseases (ICD-9) 162) in males in Scotland. Cases and populations have been aggregated between 1980 and 1984.

Table 2. Data on the incidence of lung cancer in males in Scotland aggregated over the period 1980–84

Age class index (i)	Age class	Number of incident cases (r_i)	Person-years of observation (n_i)	Age-specific rate per 100 000 (r_i/n_i)
1	0–4	0	827 400	0.00
2	5–9	0	856 500	0.00
3	10–14	0	1 061 500	0.00
4	15–19	0	1 157 400	0.00
5	20–24	4	1 074 900	0.37
6	25–29	3	917 700	0.33
7	30–34	29	890 300	3.26
8	35–39	61	816 000	7.48
9	40–44	153	724 400	21.12
10	45–49	376	706 800	53.20
11	50–54	902	703 800	128.16
12	55–59	1 819	691 200	263.17
13	60–64	2 581	610 900	422.49
14	65–69	3 071	511 800	600.04
15	70–74	3 322	425 600	780.55
16	75–79	2 452	266 800	919.04
17	80–84	1 202	122 500	981.22
18	85+	429	54 700	784.28
		16 404	12 420 200	

Age-specific rates can be calculated by applying formula (11.5). For example, for age class 40–44 ($i = 9$),

$$a_9 = \frac{r_9}{n_9} \times 100\,000$$

$$= \frac{153}{724\,400} \times 100\,000$$

$$= 21.1$$

Thus, in the age class 40–44, the average, annual age-specific incidence rate is 21.1 per 100 000. Other age-specific rates calculated in a similar fashion are listed in Table 2.

The crude rate, C, is calculated using formula (11.3) by observing that R, the total number of cases, is 16 404, and N, the total person-years of observation, is 12 420 200.

$$C = \frac{16\,404}{12\,420\,200} \times 100\,000$$

$$= 132.1$$

Hence, the average, annual, all-ages incidence rate of lung cancer in males in Scotland over the period 1980–84 is 132.1 per 100 000.

or differing population age-structure. This is accomplished by age-standardization. *It must be emphasized, however, that the difficulty in comparing rates between populations with different age distributions can be overcome completely only if comparisons are limited to individual age-specific rates* (Doll & Smith, 1982). This point cannot be stressed too much. A summary measure such as that produced by an age-adjustment technique is not a replacement for examination of age-specific rates. However, it is very useful, particularly when comparing many sets of incidence rates, to have available a summary measure of the age-standardized rate.

There are two methods of age-standardization in widespread use which are known as the direct and indirect methods. The direct method is described first, since it has considerable interpretative advantages over the indirect method (for a full discussion, see, for example, Rothman, 1986), and is generally to be preferred whenever possible. (Further information is given in Breslow & Day (1987), pp. 72–75.)

Age-standardization—direct method

An age-standardized rate is the theoretical rate which would have occurred if the observed age-specific rates applied in a reference population: this population is commonly referred to as the Standard Population.

The populations in each age class of the Standard Population are known as the weights to be used in the standardization process. Many possible sets of weights, w_i, can be used. Use of different sets of weights (i.e., use of different standard populations) will produce different values for the standardized rate. The most frequently used is the World Standard Population (see Table 3), modified by Doll *et*

Table 3. The world standard population
(After Doll *et al.*, 1966)

Age class index (i)	Age class	Population (w_i)
1	0–4	12 000
2	5–9	10 000
3	10–14	9000
4	15–19	9000
5	20–24	8000
6	25–29	8000
7	30–34	6000
8	35–39	6000
9	40–44	6000
10	45–49	6000
11	50–54	5000
12	55–59	4000
13	60–64	4000
14	65–69	3000
15	70–74	2000
16	75–79	1000
17	80–84	500
18	85+	500
		100 000

Example 2. Calculation of age-standardized rates by the direct method

Table 4 reiterates the age-specific rates of lung cancer in males in Scotland calculated earlier in this chapter. These age-specific rates (a_i) are multiplied by the weights from the Standard Population (w_i) to give the products $a_i w_i$, whose sum is found to be 9 062 410, i.e.,

$$\sum_{i=1}^{A} a_i w_i = 9\,062\,410$$

Table 4. Calculation of the age-standardized incidence rate of lung cancer in males in Scotland, aggregated over the period 1980–84 by the direct method

Age class index (i)	Age class	Age-specific rate per 100 000 (a_i)	World standard population (w_i)	($a_i \times w_i$)
1	0–4	0.00	12 000	0
2	5–9	0.00	10 000	0
3	10–14	0.00	9000	0
4	15–19	0.00	9000	0
5	20–24	0.37	8000	2960
6	25–29	0.33	8000	2640
7	30–34	3.26	6000	19 560
8	35–39	7.48	6000	44 880
9	40–44	21.12	6000	126 720
10	45–49	53.20	6000	319 200
11	50–54	128.16	5000	640 800
12	55–59	263.17	4000	1 052 680
13	60–64	422.49	4000	1 689 960
14	65–69	600.04	3000	1 800 120
15	70–74	780.55	2000	1 561 100
16	75–79	919.04	1000	919 040
17	80–84	981.22	500	490 610
18	85 +	784.28	500	392 140
			100 000	9 062 410

The standard population weights used are those of the World Standard Population whose sum is, conveniently, 100 000, i.e.,

$$\sum_{i=1}^{A} w_i = 100\,000$$

The average, annual, age-standardized incidence rate (ASR) per 100 000 of lung cancer in males in Scotland during 1980–84 is then calculated as follows:

$$\text{ASR} = \frac{\sum_{i=1}^{A} a_i w_i}{\sum_{i=1}^{A} w_i} = \frac{9\,062\,410}{100\,000} = 90.62410$$

i.e., 90.6 per 100 000 per annum. (The units of the ASR, per 100 000 per annum, are those of the age-specific rates, a_i, used in the calculations.)

al. (1966) from that proposed by Segi (1960) and used in the published volumes of the series *Cancer Incidence in Five Continents*. Its widespread use greatly facilitates the comparison of cancer levels between areas.

By denoting w_i as the population present in the ith age class of the Standard Population, where, as above, $i = 1, 2, \ldots A$ and letting a_i again represent the age-specific rate in the ith age class, the age-standardized rate (ASR) is calculated from

$$\text{ASR} = \frac{\sum_{i=1}^{A} a_i w_i}{\sum_{i=1}^{A} w_i} \tag{11.6}$$

Cases of cancer of unknown age may be included in a series. This means that equation (11.1) is no longer valid, since the total number of cases (R) is greater than the sum of cases in individual age groups ($\sum r_i$), so that the ASR, derived from age-specific rates (equation 11.5), will be an underestimate of the true value.

Doll and Smith (1982) propose that a correction is applied, by multiplying the ASR (calculated as in 11.6) by

$$\frac{R}{\sum_{i=1}^{A} r_i}$$

Use this adjustment implies that the distribution by age of the cases of unknown age is the same as that for cases of known age. Though this assumption may often not be justified, because it is more often among the elderly that age is not recorded, the effect is not usually large, as long as the proportion of cases of unknown age is small (<5%).

Truncated rates

Doll and Cook (1967) proposed calculation of rates over the truncated age-range 35–64, mainly because of doubts about the accuracy of age-specific rates in the elderly when diagnosis and recording of cancer may be much less certain. Several authors continue to present data using truncated rates, although it is debatable whether the extra accuracy offsets the somewhat increased complexity of calculations and interpretation, and the wastage of much collected data. In effect, the calculation merely limits consideration to part of the data contained in Table 4.

The truncated age-standardized rate (TASR) can be written as follows

$$\text{TASR} = \frac{\sum_{i=8}^{13} a_i w_i}{\sum_{i=8}^{13} w_i} \tag{11.7}$$

Example 3. Calculation of truncated, age-standardized rate by the direct method

Table 5 contains that part of Table 4 which is relevant for the calculation of the truncated age-standardized rate: the truncated age range is 35 to 64.

Table 5. Calculation of truncated (35–64) age-standardized incidence rate of lung cancer in males in Scotland, aggregated over the period 1980–84

Age class index (i)	Age class	Age-specific rate per 100 000 (a_i)	World standard population (w_i)	($a_i \times w_i$)
1	0–4			
2	5–9			
3	10–14			
4	15–19			
5	20–24			
6	25–29			
7	30–34			
8	35–39	7.48	6000	44 880
9	40–44	21.12	6000	126 720
10	45–49	53.20	6000	319 200
11	50–54	128.16	5000	640 800
12	55–59	263.17	4000	1 052 680
13	60–64	422.49	4000	1 689 960
14	65–69			
15	70–74			
16	75–79			
17	80–84			
18	85+			
			31 000	3 874 240

Notice in this example that

$$\sum_{i=8}^{13} w_i = 31\,000$$

and

$$\sum_{i=8}^{13} a_i w_i = 3\,874\,240$$

It is essential to remember that the weights (w_i) are only summed over the same truncated range as the $a_i w_i$. Therefore,

$$\text{TASR} = \frac{\sum_{i=8}^{13} a_i w_i}{\sum_{i=8}^{13} w_i} = \frac{3\,874\,240}{31\,000} = 124.97548$$

i.e., the average, annual truncated (35–64) age-standardized incidence rate per 100 000 of lung cancer in males in Scotland during 1980–84 is calculated to be 125.0 per 100 000.

It is clear that expression (11.7) is a special case of expression (11.6) with summation starting at age class 8 (corresponding to 35–39) and finishing with age class 13 (corresponding to 60–64). Similarly, for comparison of incidence rates in childhood, the truncated age range 0–14 has been used, with the appropriate portion of the standard population (Parkin *et al.*, 1988).

Standard error of age-standardized rates—direct method

An age-standardized incidence rate calculated from real data is taken to be, in statistical theory, an estimate of some true parameter value (which could be known only if the units of observation were infinitely large). It is usual to present, therefore, some measure of precision of the estimated rate, such as the standard error of the rate.

The standard error can also be used to calculate confidence intervals for the rate, which are intuitively rather easier to interpret. The 95% confidence interval represents a range of values within which it is 95% certain that the true value of the incidence rate lies (that is, only five estimates out of 100 would have confidence limits that did not include the true value). Alternatively, 99% confidence intervals may be presented which, because they imply a greater degree of certainty, mean that their range will be wider than the 95% interval.

In general, the $(100(1 - \alpha))$ % confidence interval of an age-standardized rate, ASR, with standard error s.e.(ASR) can be expressed as:

$$\text{ASR} \pm Z_{\alpha/2} \times (\text{s.e.(ASR)}) \tag{11.8}$$

where $Z_{\alpha/2}$ is a standardized normal deviate (see Armitage and Berry (1987) for discussion of general principles). For example, the 95% confidence interval can be calculated by selecting $Z_{\alpha/2}$ as 1.96, the 97.5 percentile of the Normal distribution. For a 99% confidence interval, $Z_{\alpha/2}$ is 2.58.

There are two methods for calculating the standard error of a directly age-adjusted rate, the binomial and the Poisson approximation, which are illustrated below. They give similar results, and either can be used.

The age-standardized incidence rate (ASR) can be computed from formula (11.6). The variance of the ASR can be shown to be

$$\text{Var (ASR)} = \frac{\sum_{i=1}^{A} [a_i w_i^2 (100\,000 - a_i)/n_i]}{\left(\sum_{i=1}^{A} w_i\right)^2} \tag{11.9}$$

The standard error of ASR (s.e.(ASR)) can be simply calculated as

$$\text{s.e.(ASR)} = \sqrt{\text{Var (ASR)}} \tag{11.10}$$

The 95% confidence interval for the ASR calculated in Example 2 is given by formula (11.8):

$$\begin{aligned} \text{ASR} \pm Z_{\alpha/2} \times (\text{s.e.(ASR)}) &= 90.62 \pm 1.96 \times 0.73 \\ &= 89.19 \text{ to } 92.05 \end{aligned}$$

Example 4. Calculation of the standard error of an age-standardized rate (binomial approximation)

Table 6 contains the data for calculation of the standard error of the age-standardized rate of lung cancer in males in Scotland in 1980–84.

Table 6. Calculation of standard error of average, annual, all-ages, age-standardized incidence rate per 100 000 of lung cancer in males in Scotland over the period 1980–84 by the binomial method

Age class	Age specific rate per 100 000 (a_i)	World standard population (w_i)	Person-years (n_i)	$a_i w_i^2(100\,000 - a_i)/n_i$
0–4	0.00	12 000	827 400	0
5–9	0.00	10 000	856 500	0
10–14	0.00	9000	1 061 500	0
15–19	0.00	9000	1 157 400	0
20–24	0.37	8000	1 074 900	2 202 988
25–29	0.33	8000	917 700	2 301 398
30–34	3.26	6000	890 300	13 181 644
35–39	7.48	6000	816 000	32 997 532
40–44	21.12	6000	724 400	104 936 424
45–49	53.20	6000	706 800	270 823 584
50–54	128.16	5000	703 800	454 659 552
55–59	263.17	4000	691 200	607 586 624
60–64	422.49	4000	610 900	1 101 862 784
65–69	600.04	3000	511 800	1 048 838 592
70–74	780.55	2000	425 600	727 873 536
75–79	919.04	1000	266 800	341 301 952
80–84	981.22	500	122 500	198 284 096
85+	784.28	500	54 700	355 634 848

$$\sum_{i=1}^{18} [a_i w_i^2 (100\,000 - a_i)/n_i] = 5\,262\,486\,016$$

and

$$\left(\sum_{i=1}^{18} w_i \right)^2 = 10\,000\,000\,000$$

Thus, from expression (11.9)

$$\text{Var (ASR)} = \frac{5\,262\,486\,016}{10\,000\,000\,000}$$

$$= 0.526249$$

and, from (11.10), s.e.(ASR) = 0.73

Thus the standard error of the average, annual, age-standardized incidence rate of lung cancer in males in Scotland in 1980–84 is 0.73.

Example 5. Calculation of standard error of an age-standardized rate (Poisson approximation)

Table 7 contains the data for calculation of the standard error of the age-standardized rate using this second method.

Table 7. Standard error of age-standardized rate (Poisson approximation)

Age class	Age specific rate per 100 000 (a_i)	World standard population (w_i)	Person-years (n_i)	$\frac{a_i w_i^2 \times 100\,000}{n_i}$
0–4	0.00	12 000	827 400	0
5–9	0.00	10 000	856 500	0
10–14	0.00	9000	1 061 500	0
15–19	0.00	9000	1 157 400	0
20–24	0.37	8000	1 074 900	2 202 996
25–29	0.33	8000	917 700	2 301 406
30–34	3.26	6000	890 300	13 182 074
35–39	7.48	6000	816 000	33 000 000
40–44	21.12	6000	724 400	104 958 592
45–49	53.20	6000	706 800	270 967 744
50–54	128.16	5000	703 800	455 242 976
55–59	263.17	4000	691 200	609 189 824
60–64	422.49	4000	610 900	1 106 537 856
65–69	600.04	3000	511 800	1 055 169 984
70–74	780.55	2000	425 600	733 599 616
75–79	919.04	1000	266 800	344 467 776
80–84	981.22	500	122 500	200 248 976
85+	784.28	500	54 700	358 446 048

$$\sum_{i=1}^{18} (a_i w_i^2 \times 100\,000/n_i) = 5\,289\,515\,520$$

and

$$\left(\sum_{i=1}^{18} w_i\right)^2 = 10\,000\,000\,000$$

Thus, from expression (11.11)

$$\text{Var (ASR)} = \frac{5\,289\,515\,520}{10\,000\,000\,000}$$

$$= 0.52895$$

Hence, from (11.10)

$$\text{s.e.(ASR)} = 0.73$$

In this example, the result is the same to two decimal places as that found by the previous method, indicating the similarity of the two approaches.

An alternative expression can be obtained, as outlined by Armitage and Berry (1987), when the a_i are small (as is generally the case) by making a Poisson approximation to the binomial variance of the a_i. This results in an expression for the variance of the age-standardized rate (Var (ASR))

$$\text{Var (ASR)} = \frac{\sum_{i=1}^{A} (a_i w_i^2 \times 100\,000/n_i)}{\left(\sum_{i=1}^{A} w_i\right)^2} \qquad (11.11)$$

and the standard error of the age-standardized rate (s.e.(ASR)) is the square root of the variance, as before (expression 11.10).

Comparison of two age-standardized rates calculated by the direct method

It is frequently of interest to study the ratio of directly age-standardized rates from different population groups, for example from two different areas, or ethnic groups, or from different time periods. The ratio between two directly age-standardized rates, ASR_1/ASR_2, is called the standardized rate ratio (SRR), and represents the relative risk of disease in population 1 compared to population 2. It is usual to calculate also the statistical significance of the standardized rate ratio (as an indication of whether the observed ratio is significantly different from unity). Several methods are available for calculating the exact confidence interval of the standardized rate ratio (Breslow & Day, 1987 (p. 64); Rothman, 1986; Checkoway *et al.*, 1989); an approximation may be obtained with the following formula (Smith, 1987):

$$(ASR_1/ASR_2)^{1 \pm (Z_{\alpha/2}/X)} \qquad (11.12)$$

where $$X = \frac{(ASR_1 - ASR_2)}{\sqrt{(\text{s.e.}(ASR_1)^2 + \text{s.e.}(ASR_2)^2)}}$$

and $Z_{\alpha/2} = 1.96$ (at the 95% level)

or $Z_{\alpha/2} = 2.58$ (at the 99% level)

If this interval includes 1.0, the standardized rates ASR_1 and ASR_2 are not significantly different (at the 5% level if $Z_{\alpha/2} = 1.96$ has been used, or at the 1% level if $Z_{\alpha/2} = 2.58$ has been used).

When the comparisons involve age-standardized rates from many subpopulations, a logical way to proceed is to compare the standardized rate for each subpopulation with that for the population as a whole, instead of undertaking all possible paired comparisons. For example, in preparing the cancer incidence atlas of Scotland, Kemp *et al.* (1985) obtained numerator and denominator information for 56 local authority districts of Scotland covering the six-year period 1975–80. For each site of cancer and separately for each sex, an average, annual, age-standardized incidence rate per 100 000 person-years was calculated by the direct method using the

Example 6. Calculation of the confidence interval for the standardized rate ratio

The age-standardized rate of lung cancer in males in 1980–84 in Scotland was found to be 90.6 with a standard error of 0.73. In 1960–64, the corresponding rate was 68.3 with a standard error of 0.67.

The standardized rate ratio, $ASR_1/ASR_2 = 90.6/68.3 = 1.3265$

To obtain the confidence interval for the standardized rate ratio, expression (11.12) is used

$$X = \frac{ASR_1 - ASR_2}{\sqrt{(\text{s.e.}(ASR_1)^2 + \text{s.e.}(ASR_2)^2)}}$$

$$= \frac{90.6 - 68.3}{\sqrt{(0.73)^2 + (0.67)^2}} = \frac{22.3}{\sqrt{0.9818}} = 22.51$$

The 95% confidence interval is obtained by letting $Z_{\alpha/2} = 1.96$ and so

Lower bound $= (ASR_1/ASR_2)^{1-(Z_{\alpha/2}/X)} = (1.3265)^{1-(1.96/22.51)} = 1.29$

Upper bound $= (ASR_1/ASR_2)^{1+(Z_{\alpha/2}/X)} = (1.3265)^{1+(1.96/22.51)} = 1.36$

If the rates in the two time periods were the same, the ratio (ASR_1/ASR_2) would be 1.0. However, the estimated 95% confidence interval for this ratio (1.29, 1.36) does not contain this value and so it can be concluded that the rates are significantly different at the 5% level.

World Standard Population (as described above). Similarly, the standard error was calculated providing for each region and for Scotland as a whole a summary comparison statistic. To avoid the effect of comparing heavily populated districts (e.g., Glasgow, with 17% of the total population of Scotland), with the rate for Scotland, which is itself affected by their contribution, the rate for each district was compared with the rate in the rest of Scotland (e.g., Glasgow with Scotland-minus-Glasgow). The method of comparison was that for directly age-standardized rates described above and the ratios were reported as: significantly high at 1% level (++); (2) significantly high at 5% level (+); (3) not significantly high or low; (4) significantly low at 5% level (−); or (5) significantly low at 1% level (−−).

Table 8 lists lung cancer incidence rates from the atlas of Scotland (Kemp *et al.*, 1985). Among males, the highest rate reported was from district 33—Glasgow City (130.6 per 100 000, standard error 2.01) which was significantly different at the 1% level from the rate for the rest of Scotland. Neighbouring Inverclyde (109.9, 5.35) also reported a significantly high rate at this level of statistical significance, as did Edinburgh City (103.2, 2.32). It is worth noting the effect of population size on statistical significance levels. Although Edinburgh City ranked only seventh in terms of male lung cancer incidence rates, it has a large population, and was one of only three districts in the highest significance group.

A similar pattern is exhibited in females, with Glasgow City (33.3, 0.90) having the highest rate. However, the second highest rate was reported from Badenoch

Table 8. Indicence rates of lung cancer in selected districts of Scotland, 1975–80 (From Kemp *et al.*, 1985)

District		Male				Female			
No.	Name	Cases	ASR	SE	Rank	Cases	ASR	SE	Rank
7	Badenoch	21	55.0^{--}	12.58	47	14	31.8	9.36	2
21	Edinburgh	2087	103.2^{++}	2.32	7	734	25.9^{++}	1.05	8
24	Tweeddale	49	73.6	11.03	28	28	29.1	6.24	3
33	Glasgow	4579	130.6^{++}	2.01	1	1802	33.3^{++}	0.90	1
37	Cumbernauld	145	109.1	9.25	3	36	21.8	3.58	18
45	Inverclyde	438	109.9^{++}	5.35	2	137	27.5	2.46	5
54	Orkney	36	40.2^{--}	6.97	56	12	13.6^{-}	4.07	48
55	Shetland	39	46.1	7.70	53	7	5.8^{--}	2.33	56
All Scotland		19 239	91.4	0.67		6136	23.1	0.31	

ASR, Age standardized rate per 100 000 (direct method, world standard population)
SE, Standard error

$^{++}$, Significantly higher than for rest of Scotland, $p<0.01$
$^{--}$, Significantly lower than for rest of Scotland, $p<0.01$
$^{-}$, Significantly lower than for rest of Scotland, $p<0.05$

(31.8, 9.36), which did not differ significantly from the rest of Scotland, because of the sparse population of the latter district.

Testing for trend in age-standardized rates

As an extension to the testing of differences between pairs of age-standardized rates described above, sometimes a set of age-standardized rates is available from populations which are ordered according to some sort of scale. The categories of this scale may be related to the degree of exposure, to an etiological factor or simply to time. Simple examples are age-standardized rates from different time periods or from different socioeconomic classes. One might also order sets of age-standardized rates from different geographical areas (provinces, perhaps) according to, for example, the average rainfall, altitude, or level of atmospheric pollution.

In these circumstances, the investigator is interested not only in comparing pairs of age-standardized rates, but also in whether the incidence rates follow some sort of trend in relation to the exposure categories. Fitting a straight line regression equation is the simplest method of expressing a linear trend.

As an example, the annual age-standardized incidence rates of lung cancer in males in Scotland will be used for the years 1960–70, inclusive. To estimate the temporal trend, the actual year can be used to order the rates; however, to simplify the calculations, 1959 can be subtracted from each year, so that 1960 becomes 1, 1961 becomes 2, . . . and 1970 becomes 11. The same results for the trend can be obtained using either set of values.

In simple regression[1] there are two kinds of variable: the predictor variable (in this case year, denoted by x) and the outcome variable (in this case the age-standardized rate, denoted by y); the linear regression equation can be written as

$$y = a + bx \tag{11.13}$$

where y = age-standardized lung cancer incidence rate
x = year number (year minus 1959)
a = intercept
b = slope of regression line

Expressions for a, b, and the corresponding standard errors are derived in Bland (1987). For example,

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

which can be rewritten as

$$b = \frac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}} \tag{11.14}$$

where n = number of pairs of observations
and $\bar{y} = \sum y/n$ and $\bar{x} = \sum x/n$

The standard error of the slope, b, is given by

$$\text{s.e.}(b) = \sqrt{\frac{\dfrac{1}{n-2}\left\{\sum (y_i - \bar{y})^2 - b^2 \sum (x_i - \bar{x})^2\right\}}{\sum (x_i - \bar{x})^2}} \tag{11.15}$$

The intercept, a, can be calculated from

$$a = \bar{y} - b\bar{x} \tag{11.16}$$

---

[1] On many occasions weighted regression may be more appropriate, where each point does not contribute the same amount of information to fitting the regression line. It is common to use weights $w_i = 1/\text{Var}\,(y_i)$: see Armitage and Berry (1987).

The calculated slope (b) indicates the average increase in the age-standardized incidence rate with each unit increase in the predictor variable, i.e., in this example, the average increase from one year to the next. The standard error of the slope (s.e.(b)) can be used to calculate confidence intervals for the slope, in a manner analogous to that using expression (11.8).

A formal test that the slope is significantly different from 1.0 can be made by calculating of the ratio of the slope to its standard error (b/s.e.(b)), which will follow a t-distribution with $n - 2$ degrees of freedom. (See Armitage and Berry (1987) for further information.)

Age-standardization—indirect method

An alternative, and frequently used, method of age-standardization is commonly referred to as indirect age-standardization. It is convenient to think of this method in terms of a comparison between observed and expected numbers of cases. The expected number of cases is calculated by applying a standard set of age-specific rates (a_i) to the population of interest:

$$\sum_{i=1}^{A} e_i = \sum_{i=1}^{A} a_i n_i / 100\,000 \tag{11.17}$$

where e_i, the number of cases expected in age class i, is the product of the 'standard rate' and the number of persons in age class i in the population of interest.

The standardized ratio (M) can now be calculated by comparing the observed number of cases ($\sum r_i$) with that expected

$$M = \frac{\sum_{i=1}^{A} r_i}{\sum_{i=1}^{A} e_i} = \frac{\sum_{i=1}^{A} r_i}{\sum_{i=1}^{A} a_i n_i / 100\,000} \tag{11.18}$$

This is generally expressed as a percentage by multiplying by 100. When applied to incidence data it is commonly known as the standardized incidence ratio (SIR): when applied to mortality data it is known as the standardized mortality ratio (SMR).

Standard error of standardized ratio

The standardized ratio (M) is derived from formula (11.18) and its variance, Var (M), is given by

$$\text{Var}\,(M) = \frac{\sum_{i=1}^{A} r_i}{\left(\sum_{i=1}^{A} a_i n_i / 100\,000\right)^2} \tag{11.19}$$

Example 7. Calculation of the average annual change in the age-standardized incidence rate for lung cancer in males in Scotland and testing for significance of the trend

Between 1960 and 1970 the annual, all-ages incidence rates per 100 000 of lung cancer in males in Scotland were 77.05, 81.78, 87.78, 89.05, 85.68, 87.04, 89.97, 100.50, 104.85, 104.77 and 107.57 respectively.

In this example, the predictor variable (x) is (year − 1959). In other words, 1960 becomes 1, 1961 becomes 2, through to 1970 which becomes 11.

$$n = 11$$
$$\sum x_i = 66$$
$$\sum x_i^2 = 506$$
$$\sum y_i = 1016.0$$
$$\sum y_i^2 = 94913.0$$
$$\sum x_i y_i = 6419.2$$

Therefore, from (11.14)

$$b = \frac{6419.2 - ((66 \times 1016)/11)}{506 - ((66 \times 66)/11)}$$
$$= \frac{323.2}{110}$$
$$= 2.938$$

and

$$\bar{x} = \sum x_i/n = 6$$
$$\bar{y} = \sum y_i/n = 92.36$$

Hence, from expression (11.15)

$$\text{s.e.}(b) = 0.351$$

The 95% confidence interval for the slope (b) is then calculated as:

$$\text{lower limit} = 2.938 - (1.96 \times 0.351)$$
$$= (2.938 - 0.688)$$
$$= 2.250$$
$$\text{upper limit} = 2.938 + (1.96 \times 0.351)$$
$$= 3.626$$

So the 95% confidence interval for b is (2.25, 3.63). To test formally whether the slope, b, differs significantly from zero, we calculate the quantity

$$\frac{b}{\text{s.e.}(b)} = \frac{2.938}{0.351} = 8.375$$

which is compared to the critical values of the t distribution, with $(n - 2)$ degrees of freedom. Here the slope is significant at the 1% level ($p < 0.01$), which is highly significant.

It can be concluded that there is evidence of a significant increase in the incidence of lung cancer in males in Scotland between 1960 and 1970, with the standardized rate increasing by an average of approximately 2.9 cases per 100 000 per annum.

Example 8. Calculation of standardized incidence ratio by the indirect method

Table 9 contains data for calculating the SIR of lung cancer in Scotland in 1980–84, using the rates in 1960–64 as standard.

Table 9. Calculation of the age-standardized incidence ratio of lung cancer in males in Scotland in 1980–84 using the rates of 1960–64 as standard

Age class	1960–64 rates per 100 000 (a_i)	Person years of observation (n_i)	Expected no. of deaths ($e_i = a_i n_i / 100\,000$)	Actual no. of deaths (r_i)
0–4	0.00	827 400	0.00	0
5–9	0.00	856 500	0.00	0
10–14	0.00	1 061 500	0.00	0
15–19	0.09	1 157 400	1.04	0
20–24	1.12	1 074 900	12.04	4
25–29	1.70	917 700	15.60	3
30–34	4.91	890 300	43.71	29
35–39	16.25	816 000	132.60	61
40–44	29.38	724 400	212.83	153
45–49	79.92	706 800	564.87	376
50–54	151.07	703 800	1063.23	902
55–59	269.58	691 200	1863.34	1819
60–64	391.41	610 900	2391.12	2581
65–69	459.74	511 800	2352.95	3071
70–74	400.46	425 600	1704.36	3322
75–79	285.21	266 800	760.94	2452
80–84	207.49	122 500	254.18	1202
85+	100.84	54 700	55.16	429
			11427.97	16404

$$\sum e_i = \sum_{i=1}^{A} a_i n_i / 100\,000 = 11\,427.97$$

and

$$\sum r_i = 16\,404$$

Therefore, the standardized incidence ratio, $M \times 100$, for the period 1980–84 given by formula (11.18), is

$$\frac{16\,404}{11\,427.97} \times 100 = 144$$

In other words, lung cancer in males was 44% higher in 1980–84 than in 1960–64, after the different age structure had been taken into account.

and the standard error of the indirect ratio, s.e.(M), is the square root of the variance, as before (expression 11.10).

$$\text{s.e.}(M) = \frac{\sqrt{\sum_{i=1}^{A} r_i}}{\sum_{i=1}^{A} a_i n_i / 100\,000} \tag{11.20}$$

Vandenbroucke (1982) has proposed a short-cut method for calculating the $(100(1 - \alpha))\%$ confidence interval of a standardized ratio, involving a two-step procedure. First, the lower and upper limits for the observed number of events are calculated:

$$\text{Lower limit} = [\sqrt{\text{observed events}} - (Z_{\alpha/2} \times 0.5)]^2$$

$$\text{Upper limit} = [\sqrt{\text{observed events}} + (Z_{\alpha/2} \times 0.5)]^2$$

Example 9. Calculation of standard error of indirectly standardized ratio

Table 9 contains the data for calculating the standard error of the standardized incidence ratio (SIR) for males in Scotland in 1980–84 relative to 1960–64.

Recalling that SIR = $M \times 100$,

$$\text{Var (SIR)} = \text{Var}(M \times 100) = 10\,000\ \text{Var}(M)$$

Now, from (11.19),

$$\text{Var (SIR)} = 10\,000 \frac{\sum_{i=1}^{A} r_i}{\left(\sum_{i=1}^{A} a_i n_i\right)^2}$$

$$= \frac{10\,000 \times 16\,404}{(11\,427.97)^2}$$

$$= 1.2561$$

and

$$\text{s.e.(SIR)} = \sqrt{\text{Var (SIR)}} = 1.12$$

Thus, in this example, the SIR is 144 and the corresponding standard error is 1.12. The 95% confidence interval for the SIR is, therefore,

$$\text{SIR} \pm (Z_{\alpha/2} \times (\text{s.e.(SIR)})) = 144 \pm (1.96 \times 1.12)$$

i.e., (141.8, 146.2)

Division of these limits for the observed number by the expected number of events yields the approximate 95% (or 99%) confidence interval for the SIR.

$$\text{Lower limit of SIR} = \frac{[\sqrt{\text{observed events}} - (Z_{\alpha/2} \times 0.5)]^2}{\text{expected events}}$$

$$= \frac{\left\{\sqrt{\sum_{i=1}^{A} r_i} - (Z_{\alpha/2} \times 0.5)\right\}^2}{\sum_{i=1}^{A} a_i n_i / 100\,000} \tag{11.21}$$

$$\text{Upper limit of SIR} = \frac{\left\{\sqrt{\sum_{i=1}^{A} r_i} + (Z_{\alpha/2} \times 0.5)\right\}^2}{\sum_{i=1}^{A} a_i n_i / 100\,000} \tag{11.22}$$

Testing whether the standardized ratio differs from the expected value

This can be achieved simply by calculating the appropriate confidence intervals, so that it can be seen whether the value of 100 is included or excluded.

Example 10. Calculation of approximate 95% confidence interval of the standardized incidence ratio

The data for this calculation have already been presented in Table 9.

No. of observed events = 16 404

No. of expected events = 11 427.97

For 95% confidence interval, $Z_{\alpha/2}$ is 1.96

$$\text{Lower limit} = \frac{[\sqrt{16\,404} - (1.96 \times 0.5)]^2}{11\,427.97}$$

$$= 1.41353$$

$$\text{Upper limit} = \frac{[\sqrt{16\,404} + (1.96 \times 0.5)]^2}{11\,427.97}$$

$$= 1.45747$$

Since the SIR was expressed as a percentage, these approximate limits become 141.4 and 145.7.

These limits are quite close to the more precise values obtained in Example 9, yet have the advantage of being simpler to calculate.

Example 11. Test of significance of indirectly adjusted ratios

In Example 9, the SIR in Scotland in 1980–84 was calculated to be 144% with a standard error of 1.12%, and thus a 95% confidence interval of 141.8 to 146.2, which does not include 100.

Similarly, the 99% confidence interval for the SIR is 144 ± (2.58 × 1.12), i.e., 141.1, 146.9), which, again, does not include 100.

Thus, it can be concluded that the lung cancer rate observed in 1980–84 was significantly higher than that in 1960–64 at the 1% level of significance.

It should be noted that, with indirect standardization, the population weights which are used in the standardization procedure are the age-specific populations in the subgroup under study. Thus if SIRs are calculated for many population subgroups (e.g., different provinces, ethnic groups) with different population structures, the different SIRs can only be related to the standard population (as in Example 11) and not to each other. Thus, if the SIR for lung cancer in males in Scotland in 1970–74, using the incidence rates of 1960–64 as our standard, is calculated to be 1.22 (or 122 as a percentage), it cannot be deduced that the relative risk in 1980–84 compared to 1970–74 is 144/122 or 1.18.

Cumulative rate and cumulative risk

Day (1987) proposed the cumulative rate as another age-standardized incidence rate. In Volume IV of the series *Cancer Incidence in Five Continents*, this measure replaced the European and African standard population calculations (Waterhouse *et al.*, 1982).

The *cumulative risk* is the risk which an individual would have of developing the cancer in question during a certain age span if no other causes of death were in operation. It is essential to specify the age period over which the risk is accumulated: usually this is 0–74, representing the whole life span. For childhood cancers, 0–14 can be used.

The *cumulative rate* is the sum over each year of age of the age-specific incidence rates, taken from birth to age 74 for the 0–74 rate. It can be interpreted either as a directly age-standardized rate with the same population size in each age group, or as an approximation to the cumulative risk.

It will be recalled that a_i is the age-specific incidence rate in the ith age class which is t_i years long. In other words if the age classes used are 0, 1–4, 5–9 . . . then t_1 will be 1, t_2 will be 4, t_3 will be 5 etc. The cumulative rate can be expressed as

$$\text{Cum. rate} = \sum_{i=1}^{A} a_i t_i \qquad (11.23)$$

where the sum is until age class A. Assuming five-year age classes have been used throughout in the calculation of age-specific rates, for the cumulative rate 0–74, $A = 15$ and

$$\text{Cum. rate (0–74)} = \sum_{i=1}^{15} 5a_i$$

It is more common to express this quantity as a percentage rather than per 100 000.

The cumulative risk has been shown by Day (1987) to be

$$\text{Cum. risk} = 100 \times [1 - \exp(-\text{cum. rate}/100)] \qquad (11.24)$$

Example 12. Calculation of cumulative rate and cumulative risk

Table 10 contains data for calculations of the cumulative rate and cumulative risk for lung cancer in Scotland among males in 1980–84. Only equal age classes are used in the example; all the t_i are 5 years long.

Table 10. Calculation of cumulative rate and cumulative risk (0–74) of lung cancer in males in Scotland, 1980–84

Age class	Age-specific rate per 100 000 (a_i)	Length of age class (t_i)	Age specific rate × (length of age class) per 100 000 ($a_i t_i$)
0–4	0.00	5	0
5–9	0.00	5	0
10–14	0.00	5	0
15–19	0.00	5	0
20–24	0.37	5	1.85
25–29	0.33	5	1.65
30–34	3.26	5	16.30
35–39	7.48	5	37.40
40–44	21.12	5	105.60
45–49	53.20	5	266.00
50–54	128.16	5	640.80
55–59	263.17	5	1315.85
60–64	422.49	5	2112.45
65–69	600.04	5	3000.20
70–74	780.55	5	3902.75
75–79	—	—	—
80–84	—	—	—
85+	—	—	—
			11 400.85

$$\text{Cum. rate} = \sum_{i=1}^{15} a_i t_i = 11\,400.85$$

The cumulative rate (0–74) is 11 400.9 per 100 000, or 11.4%.

The cumulative risk (0–74) is

$$100 \times [1 - \exp(-11.4/100)]$$
$$= 10.8\%$$

Thus, in the absence of other causes of death, a male in Scotland has an estimated 10.8% risk of developing lung cancer before the age of 75.

Standard error of cumulative rate

The variance and standard error of the cumulative rate can be derived from the expressions for the variance and standard error of a directly adjusted rate (11.10 and 11.11) using the appropriate weights (i.e., the lengths of the age-intervals, t_i) and the Poisson approximation:

$$\text{Var (cum. rate)} = \sum_{i=1}^{A} (a_i t_i^2 / n_i) \qquad (11.25)$$

Example 13. Calculation of standard error of cumulative rate

Table 11 contains the calculations necessary to compute the standard error of the cumulative rate.

Table 11. Calculation of standard error of cumulative (0–74) rate of lung cancer in males in Scotland 1980–1984

Age class	Age-specific rate per 100 000 (a_i)	Length of age class (t_i)	Person-years (n_i)	$a_i t_i^2/n_i$
0–4	0.00	5	827 400	0
5–9	0.00	5	856 500	0
10–14	0.00	5	1 061 500	0
15–19	0.00	5	1 157 400	0
20–24	0.37	5	1 074 900	0.00001
25–29	0.33	5	917 700	0.00001
30–34	3.26	5	890 300	0.00009
35–39	7.48	5	816 000	0.00023
40–44	21.12	5	724 400	0.00073
45–49	53.20	5	706 800	0.00188
50–54	128.16	5	703 800	0.00455
55–59	263.17	5	691 200	0.00952
60–64	422.49	5	610 900	0.01729
65–69	600.04	5	511 800	0.02931
70–74	780.55	5	425 600	0.04584
				0.10947

$$\text{Var (cum. rate)} = \sum_{i=1}^{15} a_i t_i^2/n_i = 0.10947 \text{ per } 100\,000$$

$$= 0.00010947\%$$

The standard error of the cumulative (0–74) rate, s.e.(cum. rate) is obtained by taking the square root of this expression.

$$\text{s.e.(cum. rate)} = \sqrt{0.00010947} \text{ per } \sqrt{100}$$

$$= 0.105\%$$

i.e., the cumulative (0–74) rate of lung cancer in males in Scotland was found to be 11.4% and the standard error was 0.1%.

and hence the standard error of the cumulative rate, s.e.(cum. rate) can be expressed as

$$\text{s.e.(cum. rate)} = \sqrt{\sum_{i=1}^{A} (a_i t_i^2 / n_i)} \tag{11.26}$$

A 95% confidence interval for the cumulative rate is readily obtained by using equation (11.8):

$$11.4 \pm (1.96 \times 0.105)$$

i.e. 11.6, 11.2

PART II. PROPORTIONATE METHODS

Percentage (relative) frequency

If the population from which the cases registered are drawn is unknown, it is not possible to calculate incidence rates. In these circumstances, different case series must be compared in terms of the proportionate distribution of different types of cancer. The usual procedure is to calculate the percentage frequency (or relative frequency) of each cancer relative to the total:

$$\text{relative frequency} = \frac{R}{T} \tag{11.27}$$

where R = number of cases of the cancer of interest in the study group

T = number of cases of cancer (all sites) in the study group

An alternative is the ratio frequency (Doll, 1968) where each cancer is expressed as a proportion of all other cancers, rather than as a proportion of the total:

$$\text{ratio frequency} = \frac{R}{T - R} \tag{11.28}$$

This may have advantages in certain circumstances (for example, when dealing with a cancer that constitutes a large proportion of the total series), but there are disadvantages also, and it is not considered further here.

Comparisons of relative frequency may take place between registries, or within a registry, for example, between different geographical areas, different ethnic groups or different time periods. The problem with using relative frequency of different tumours in this way is that the comparison is often taken as an indication of the actual difference in risk between the different subgroups, which in fact can only be measured as the ratio between incidence rates. The ratio between two percentages will be equivalent to the relative risk only if the overall rates (for all cancers) are the same.

In the example shown in Figure 1, the ratio between the incidence rates (rate ratio) of liver cancer in Cali and Singapore Chinese, which have similar overall rates of incidence, is 6.9. This is well approximated by the ratio between the percentage frequencies of liver cancer in the two populations (7.3). However, although the rate

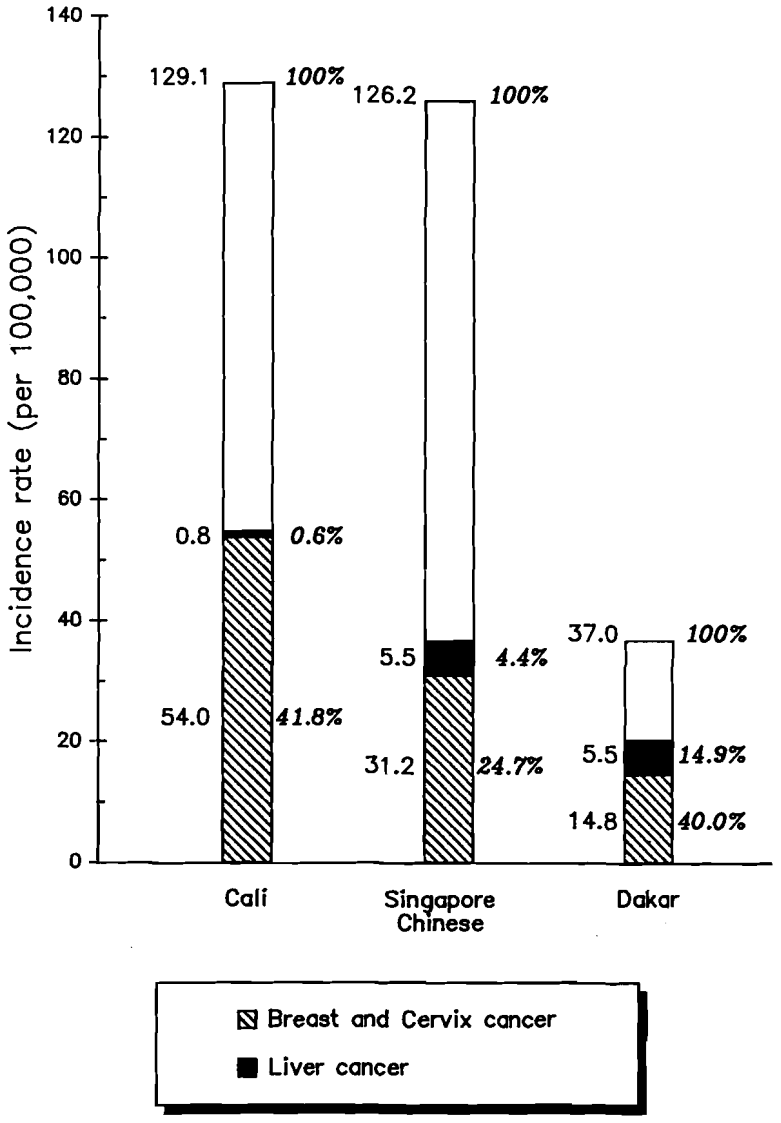

Figure 1. Incidence rates (per 100 000) and percentage frequencies of cancers in females in three three registries

Breast + cervix cancer (ICD 174 + 180); liver cancer (ICD 155). For liver cancer, ratio of incidence rates Singapore Chinese:Cali = 5.5/0.8 = 6.9, Singapore Chinese:Dakar = 5.5/5.5 = 1.0; Ratio of percentages Singapore Chinese:Cali = 4.4/0.6 = 7.3, Singapore Chinese:Dakar = 4.4/14.9 = 0.3.

ratio (relative risk) of liver cancer in Singapore Chinese and Dakar is 1.0, the ratio between the two percentages is 0.30. This is because the overall incidence rate in Dakar (37.0 per 100 000) is only 29% of that in Singapore Chinese (126.2 per 100 000) because cancers other than liver cancer are less frequent there.

An analogous problem is encountered in comparing percentage frequencies of cancers in males and females from the same centre. In practically all case series, the incidence of female-specific cancers (breast, uterus, ovary) will be considerably greater than for male-specific cancers (prostate, testis, penis). However, because in comparisons of relative frequency the total percentage must always be 100, the frequency of those cancers which are common to both sexes will always be lower in females.

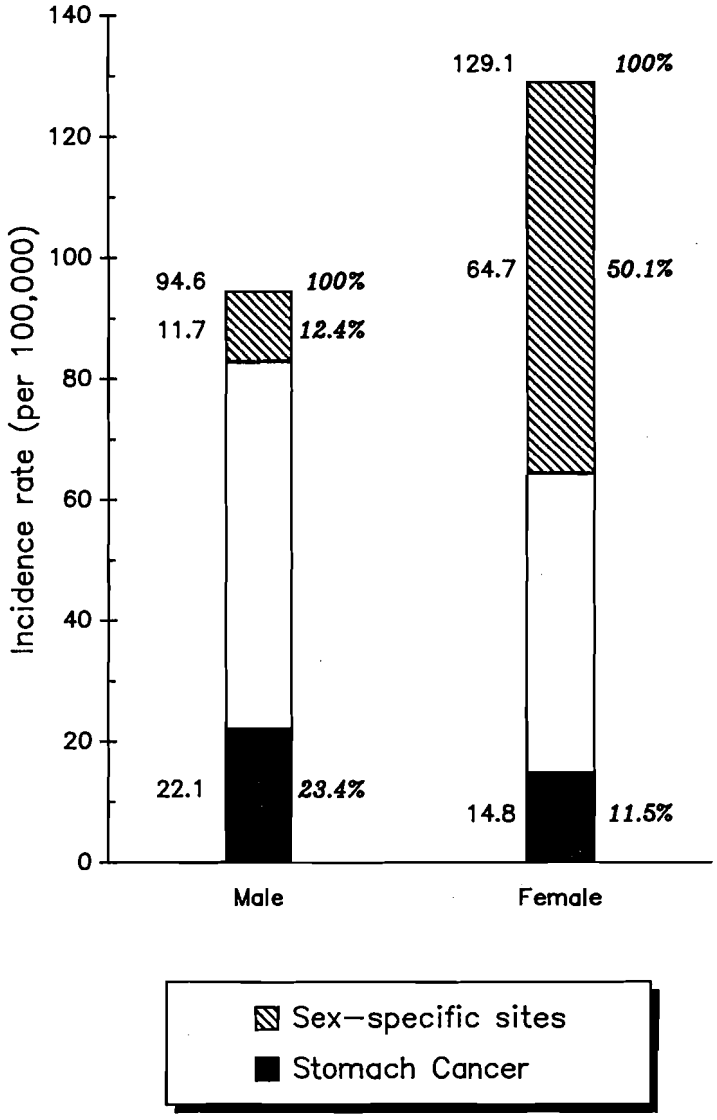

Figure 2. Incidence rates (per 100 000) and percentage frequencies of stomach cancer and sex-specific cancers in males and females, Cali, Colombia, 1972–1976
Sex-specific sites (ICD 174-183 females, ICD 185-187 males); Stomach cancer (ICD 151). Sex ratio of stomach cancer: ratio of incidence rates, M:F = 22.1/14.8 = 1.49; ratio of percentages, M:F = 23.4/11.5 = 2.03; ratio of percentages excluding sex-specific sites, M:F = 26.7/23.0 = 1.16.

In the example shown, the risk of stomach cancer in males relative to females in Cali, comparing incidence rates, is 1.49 (Figure 2). However, the ratio of the relative frequencies is 2.03, because sex-specific cancers are responsible for about half of the tumours in females, whereas they account for only 12% in males. Comparisons of relative frequencies within a single sex do, of course, give the same results as comparisons of incidence rates.

One solution to the problem of comparing relative frequencies between different centres where the occurrence of certain common tumours is highly variable is to calculate residual frequencies, that is the percentage frequency of a particular cancer after removing tumours occurring at the most variable rates from the series. This procedure may be useful for comparing series where the differences in total incidence

rates are largely due to a few variable tumours—it has been used, for example, for comparing series from Africa by Cook and Burkitt (1971). However, it does somewhat complicate interpretation, and the results may be no clearer than using the simple relative frequency. Thus, in the example in Figure 2, removing sex-specific sites from the denominator means that total incidence becomes higher in males than females, so that the ratio of residual frequencies for stomach cancer (1.16) becomes an under-estimate rather than over-estimate of the true relative risk (1.49).

In the example already presented in Figure 1, cervix plus breast cancer constitutes 40% of cancers in Dakar but only 24.7% in Singapore Chinese. If these variable tumours are excluded from the denominator, the residual frequencies of liver cancer are 5.8% (4.4/100 − 24.7) in Singapore Chinese and 24.8% (14.9/100 − 40.0) in Dakar. The estimate of relative risk obtained by comparing these residual frequencies is 0.23 (5.8/24.8), which is further from the true value (1.0) than the estimate obtained by comparing crude percentages (0.30).

Age-standardization

As in the case of comparisons of incidence rates, comparison of proportions is complicated by differences in the age structure of the populations being compared.

The relative frequency of different cancer types varies considerably with age; for example, certain tumours, such as acute leukaemia, are commoner in childhood whilst others, which form a large proportion of cancers in the elderly (such as carcinomas of the respiratory and gastrointestinal tract) are very rare. Thus the proportion of different cancers in a case series is strongly influenced by its age composition, and some form of standardization for age is necessary when making comparisons between them.

Two methods have been used for age-standardization, the age-standardized cancer ratio (ASCAR), which is analogous to direct age standardization (Tuyns, 1968), and the standardized proportional incidence ratio (SPIR or PIR), which is an indirect standardization. Of these, the PIR has considerable advantages, the ASCAR being really of value only when data sets from completely different sources are compared, where there is no obvious standard for comparison.

The age-standardized cancer ratio (ASCAR)

The ASCAR is a direct standardization, which requires the selection of a set of standard age-specific proportions to which the series to be compared will be standardized. The choice is quite arbitrary, but a standard which is somewhat similar to the age-distribution of all cancers in the case series being compared will lead to the ASCAR being relatively close to the crude relative frequency. The proportions used for comparing frequencies of cancers in different developing countries (Parkin, 1986) are shown in Table 12.

The ASCAR is calculated as

$$\text{ASCAR} = \sum_{i=1}^{A} (r_i/t_i)\, w_i \qquad (11.29)$$

where

r_i = number of cases of the cancer of interest in the study group in age class i

t_i = number of cases of cancer of all sites in the study group in age class i

w_i = standard proportion for age class i

Table 12. Standard age distribution of cancer cases for developing countries[a]

Age range	%
0–14	5
15–24	5
25–34	5
35–44	10
45–54	20
55–64	25
65–74	20
75+	10
All	100

[a] From Parkin (1986)

Example 14. Calculation of the age-standardized cancer ratio

Table 13 contains data for the calculation of the ASCAR of nasopharyngeal cancer in Tunisian males. By equation 11.29, ASCAR = 10.98, which may be compared with the crude relative frequency (from equation 11.27) of

$$\frac{344}{3073} \times 100 = 11.19\%$$

Table 13. Calculation of age-standardized cancer ratio (ASCAR) for nasopharyngeal cancer in Tunisian males, 1976–80

Age class	No. of cases		Nasopharyngeal as proportion of all cancers	Standard proportion %	Expected %
	Nasopharyngeal (r_i)	All cancers (t_i)	(r_i/t_i)	(w_i)	$(r_i/t_i)w_i$
0–14	16	257	0.062	5	0.31
15–24	37	239	0.155	5	0.78
25–34	22	132	0.167	5	0.84
35–44	60	292	0.205	10	2.05
45–54	88	612	0.144	20	2.88
55–64	76	744	0.102	25	2.55
65–74	40	619	0.065	20	1.30
75+	5	178	0.028	10	0.28
	344	3073	0.112	100	10.98

[a] From Parkin (1986)

The ASCAR is interpreted as being the percentage frequency of a cancer which would have been observed if the observed age-specific proportions applied to the percentage age-distribution of all cancers in the standard population. It must be stressed that the problems of making comparisons between data sets with different overall incidence rates remain the same and are not corrected by standardization.

The statistical problems of comparing ASCAR scores have not been investigated and there appears to be no formula available for calculating a standard error.

The proportional incidence ratio (PIR)

The proportional incidence ratio is the method of choice for comparing data sets where a standard set of age-specific proportions is available for each cancer type (analogous to indirect age standardization, which requires a set of standard age-specific incidence rates). The usual circumstance is when a registry wishes to compare different sub-classes of the cases within it—defined, for example, by place of residence, ethnic group, occupation etc. In this case a convenient standard is provided by the age-specific proportions of each cancer for the registry as a whole. (Actually, an external standard is preferable, since the total for the registry will also include the sub-group under study. In practice, unless any one subgroup forms a large percentage (30% or more) of the total, this is relatively unimportant.)

In the proportional incidence ratio, the expected number of cases in the study group due to a specific cancer is calculated, and the PIR is the ratio of the cases observed to those expected—just like the SIR—and it is likewise usually expressed as a percentage.

The expected number of cases of a particular cancer is obtained by multiplying the total cancers in each age group in the data set under study, by the corresponding age–cause-specific proportions in the standard. Expressed symbolically,

$$\text{PIR} = (R/E) \times 100 \tag{11.30}$$

$$E = \sum_{i=1}^{A} t_i(r_i^*/t_i^*) \tag{11.31}$$

where

R = observed cases at the site of interest in the group under study

E = expected cases at the site of interest in the group under study

r_i^* = number of cases of the cancer of interest in the age group i in the standard population

t_i^* = number of cases of cancer (all sites) in the age group i in the standard population

t_i = number of cases of cancer (all sites) in the age group i in the study group

Breslow and Day (1987) give a formula for the standard error of the log PIR as follows:

$$\text{s.e.(log PIR)} = \frac{\left[\sum_{i=1}^{A} r_i(t_i - r_i)/t_i\right]^{1/2}}{R} \tag{11.32}$$

Example 15. Calculation of the proportional incidence ratio

The data given in Table 14 allow the calculation of the PIR for liver cancer in one region of Thailand, using as a standard the age-specific proportions of liver cancers in Thailand as a whole.

Table 14. Data for calculation of PIR for liver cancer in males in one region of Thailand[a]

Age	Thailand			Region 4		
	Liver cancer (r_i^*)	All cancers (t_i^*)	Proportion (r_i^*/t_i^*)	Liver cancer (r_i)	All cancers (t_i)	Expected liver cancer $t_i(r_i^*/t_i^*)$
0–4	2	210	0.010	0	9	0.090
5–9	1	143	0.007	0	5	0.035
10–14	4	145	0.027	0	4	0.108
15–19	7	230	0.030	1	12	0.360
20–24	23	265	0.087	2	23	2.001
25–29	50	368	0.136	11	37	5.032
30–34	120	492	0.244	22	57	13.908
35–39	169	685	0.247	31	84	20.748
40–44	314	1077	0.292	52	123	35.916
45–49	383	1540	0.249	107	213	53.037
50–54	470	2155	0.218	95	220	47.960
55–59	388	2093	0.185	66	182	33.670
60–64	323	2161	0.150	74	174	26.100
65–69	230	1910	0.120	41	152	18.240
70–74	148	1631	0.091	27	90	8.190
75–79	69	980	0.070	12	35	2.450
80–84	21	426	0.049	4	15	0.735
85+	5	172	0.029	0	8	0.232
	2727	16 683		545	1443	268.812

[a] From Srivatanakul *et al.* (1988)

From expression (11.31),

$$E = \sum_{i=1}^{A} t_i(r_i^*/t_i^*) = 268.812$$

From expression (11.30)

$$\text{PIR} = (R/E) \times 100 = 545/268.812 \times 100 = 203\%$$

where

r_i = number of cases of the cancer of interest in the age group i in the study group

A simpler formula may be used as a conservative approximation to formula (11.32), provided that the fraction of cases due to the cause of interest is quite small:

$$\text{s.e.}(\log \text{PIR}) = 1/\sqrt{R} \qquad (11.33)$$

From the data in Table 14, using expression (11.32), the standard error can thus be calculated as:

$$\text{s.e.(log PIR)} = \frac{\sqrt{325.03}}{545} = 0.033$$

and using the approximate formula (11.33)

$$\text{s.e.(log PIR)} = \frac{1}{\sqrt{545}} = 0.043$$

Breslow and Day (1987) do not recommend that statistical inference procedures be conducted on the PIR; questions of statistical significance of observed differences can be evaluated with the confidence interval.

To obtain 95% confidence interval for a PIR of 2.03 (Example 15), and using the s.e.(log PIR) calculated by using expression (11.32)

$$\text{PIR} = 2.03$$

$$\log \text{PIR} = 0.708$$

$$95\% \text{ confidence interval for } \log \text{PIR} = 0.708 \pm (1.96 \times 0.033)$$

$$= 0.643, 0.773$$

$$95\% \text{ confidence interval for PIR} = 1.90, 2.17$$

Relationships between the PIR and SIR

Because calculation of the PIR does not require information on the population at risk, a raised PIR does not necessarily mean that the risk of the disease is raised, merely that there is a higher proportion of cases due to that cause than in the reference population.

The relationship between the PIR and the SIR has been studied empirically by several groups (Decouflé *et al.*, 1980; Kupper *et al.*, 1978; McDowall, 1983; Roman *et al.*, 1984).

In practice, it is found that for any study group

$$\text{PIR} = \frac{\text{SIR}}{\text{SIR (all cancers)}}$$

The ratio SIR/SIR (all cancers) is termed the relative SIR. Thus, a relative SIR of greater than 100 suggests that the cause-specific incidence rate in the study population is greater than would have been expected on the basis of the incidence rate for all cancers. A consequence of this is that the PIR can be greater than 100 whilst the SIR is less, or vice versa.

Table 15 shows an example from the Israel cancer registry (Steinitz *et al.*, 1989). In this example, Asian-born males have a lower incidence of cancer (all sites) than the reference population (here 'all Jewish males'), resulting in an SIR (all cancers) of 77%. They also have a lower SIR for lung cancer than all Jewish males (86%). However,

because lung cancer is proportionately more important in Asian males than in Jewish males as a whole, the PIR exceeds 100.

Table 15. Relationship between PIR and SIR. Cancer incidence in Jews in Israel: males born in Asia relative to all Jewish males

Cause	Observed cases	SIR (%)	PIR (%)	Relative SIR (%)
All cancers	6771	77	100	100
Oesophagus	114	105	139	136
Stomach	693	76	100	99
Liver	125	110	140	143
Lung	1062	86	112	112