Chapter 8. Age-standardization

F. Bray

Age is a major determinant of cancer incidence. The risk of epithelial cancers, which comprise 90% of all cancers worldwide, increases approximately as a fifth power of age (Armitage & Doll, 1954). There are major differences in the underlying age structures of the registry populations compiled in *Cancer Incidence in Five Continents*, and it is therefore essential that comparisons of cancer risk be made independent of the age profile of each population. This chapter briefly discusses the age-corrected rates used in this volume to allow summary comparisons of incidence across populations and over time.

It has been stressed in each consecutive volume that the most suitable comparisons of cancer risk are those made using the age-specific rates directly. A graphical display of age-specific rates in this respect can be particularly informative. Such exploratory analyses can be realized using the CD-ROM that complements this volume.

To facilitate comparisons, however, a summary rate is required that absorbs each registry's schedule of age-specific rates. The crude rate does not meet this objective, as it depends on the age structure of the population under study. The two standardized measures that appear in the main tables in this and the three previous volumes are the age-standardized rate and the cumulative rate. Both aim to provide a single summary statistic that is independent of the effects of age, thus allowing comparisons of cancer risk between registries.

What follows is a concise account of the properties of the measures, and the principal calculations involved. Readers wishing to explore the topic in more depth may consult the chapters by Smith (1992) and Day (1992) in volume VI. Other publications by Jensen *et al.* (1991) and Dos Santos Silva (1999) include a practical discussion of age-standardization of cancer data, while Estève *et al.* (1994) give a more theoretical account of the methodology.

Other issues related to achieving data comparability between registries in this volume are discussed in Chapter 5.

Standardized incidence rate

The age-standardized incidence rate is the summary rate that would have been observed, given the schedule of age-specific rates, in a population with the age composition of some reference population, often called the *standard*. The calculation of the standardized rate is an example of *direct standardization*, whereby the observed age-specific rates are applied to a standard

Table 8.1. Computation of age-standardized incidence rates (stomach cancer, Denmark, males, 1993–97)

Age group i	No. of cases d_i	Person-years at risk y_i	Age-specific incidence (per 10^5 years) $10^5(d_i/y_i)$	Standard world population w_i	Expected cases in standard population $d_i w_i/y_i$
0–4	0	863 799	0.00	12 000	0.00
5–9	0	765 274	0.00	10 000	0.00
10–14	0	710 183	0.00	9 000	0.00
15–19	1	824 459	0.12	9 000	0.01
20–24	1	946 843	0.11	8 000	0.01
25–29	4	1 034 977	0.39	8 000	0.03
30–34	10	1 040 497	0.96	6 000	0.06
35–39	16	958 241	1.67	6 000	0.10
40–44	36	942 690	3.82	6 000	0.23
45–49	67	1 013 068	6.61	6 000	0.40
50–54	113	902 693	12.52	5 000	0.63
55–59	142	691 196	20.54	4 000	0.82
60–64	197	583 101	33.78	4 000	1.35
65–69	238	525 292	45.31	3 000	1.36
70–74	307	455 623	67.38	2 000	1.35
75–79	262	330 900	79.18	1 000	0.79
80–84	235	209 558	112.14	500	0.56
85+	126	126 826	99.35	500	0.30
Total	1755	12 925 220	13.58	100 000	8.19

population. The calculation is illustrated in Table 8.1 for stomach cancer incidence among males in Denmark 1993–97, using the world population modified by Doll *et al.* (1966), after Segi (1960), as the reference population. Age groups are indexed by the subscript i, d_i is the number of cases, y_i is the number of person-years at risk (obtained by multiplying the number of males in the Danish population by the observation period of five years) and w_i is the number of persons (or *weight*) in age group i in the world standard population. The crude rate per 100 000 per annum is:

$$10^5 \left(\sum_i y_i\right) \Big/ \left(\sum_i d_i\right) = 5 \times 1755/12925220$$

$$= 13.58$$

The age-standardized rate is given by

$$\sum_i d_i \, w_i / y_i = 8.19$$

In this example, the age-standardized rate is 40% lower than the crude rate in the five-year period. This is because the standard world population has proportionally fewer individuals in the older age groups than the corresponding Danish population, and the risk of disease (age-specific rates) is highest in the oldest age groups.

A note on the choice of standard population

The main criticism levelled at age-standardized rates is the need to select an arbitrary standard population. Age-standardized rates can be meaningfully compared only if they refer to the same standard. In the last half century, the most widely used reference population for global comparisons has been the world standard, as proposed by Segi (1960) on the basis of the pooled population of 46 countries, and modified for the first volume of this series by Doll *et al.* (1966).

It is clear that this age composition is representative of neither the present nor the future age-specific global population (United Nations, 1999). This has recently led the World Health Organization (WHO) to propose a standard based on the mean world population age structure projected for the period 2000–25, for its disease comparisons (Ahmad *et al.*, 2000).

While this reasoning is sound, switching to another standard in this volume would require clear benefits that outweigh the drawbacks of rendering the Segi world-standardized rates in the previous seven volumes no longer useful. With this in mind, a recent study (Bray *et al.*, 2002) examined the validity of the ratio of age-standardized rates using the WHO and Segi standards as estimators of relative risk against methods based on Mantel and Haenszel and maximum likelihood methodologies. Geographical and temporal risk differences were evaluated using data for four cancers exhibiting differing risk patterns by age.

While the age-standardized rates calculated using the Segi and WHO standards produced very different absolute values, the estimates of relative risk were similar using either standard, and in accordance with the relative risks estimated using the other stratified methods. This result held whether risk was compared between registries in a fixed period, or in a fixed population, between periods. The authors concluded that there was nothing to gain by changing the standard population for routine comparisons of cancer data worldwide, other than the inconsequential property that the value of the standardized rate would be closer to the crude rate.

Accordingly, for these theoretical and practical reasons, the age-standardized rates in this volume are calculated as previously, using the age composition of the Segi world standard given in Table 8.1. The statistic thus continues to provide a means to rapidly examine geographical and temporal variations in cancer risk across all eight consecutive volumes of the series.

Cumulative rate and cumulative risk

To better understand the properties of the cumulative rate, the concept of cumulative risk will be first introduced. The cumulative risk is defined as the probability that an individual will develop the disease in question during a certain age span, in the absence of other competing causes of death. The age span over which the risk is accumulated must be specified. In this volume, the age ranges 0–64 and 0–74 years are used, to give two representations of the lifetime risk of developing the disease. Other age ranges may be appropriate for more specific needs, such as investigating childhood diseases. If the cumulative risk using the above age ranges is less than 10%, as is the case for most tumours, it can be approximated very well by the cumulative rate.

The cumulative rate is the summation of the age-specific rates over each year of age from birth to a defined upper age limit. As age-specific incidence rates are usually computed for five-year age intervals, the cumulative rate is five times the sum of the age-specific rates calculated over the five-year age groups, assuming the age-specific rates are the same for all ages within the five-year age stratum.

In the example of stomach cancer incidence in males in Denmark, all the age groups are of five years, so the cumulative rate from 0 to 74 is given by:

$$5\sum_i d_i / y_i = 5 \times 193.2 \times 10^{-5}$$

$$= 0.0097$$

The cumulative rate is not in fact a rate, but a dimensionless quantity. In other words, it is not expressed in units of 'per annum' but simply as a number. It is most conveniently expressed as a percentage, so the cumulative rate up to age 74 in the above example would be given as 0.97%.

The precise mathematical relationship between the cumulative rate and the cumulative risk is:

cumulative risk = 1 – exp (– cumulative rate)

Table 8.2 shows the correction needed to convert the cumulative rate into the cumulative risk. For values under 10%, the difference is small.

The cumulative rate has several advantages over age-standardized rates. Firstly, as a form of direct standardization, the problem of choosing an arbitrary reference population is eliminated. Secondly, as an approximation to the cumulative risk, it has a greater intuitive appeal, and is more directly interpretable as a measurement of lifetime risk, assuming no other causes of death are in operation.

Table 8.2. Conversion of cumulative rates (100x) into the corresponding cumulative risk 100 ($1–e^{-x}$)

100x	**0.1**	**0.5**	**1.0**	**5.0**	**7.0**	**10.0**	**15.0**	**20.0**	**30.0**	**40.0**	**50.0**
100 ($1–e^{-x}$)	**0.1**	**0.499**	**0.995**	**4.88**	**6.76**	**9.52**	**13.93**	**18.13**	**25.92**	**32.97**	**39.35**

Calculation of the standard error

Both the standardized and the cumulative rate are weighted sums of the age-specific rates, so the standard error can be derived in both cases from the same formula. If the age-specific rate in age group i is estimated from d_i cases and y_i person-years, the age-standardized rate (with w_i representing the standardization weights), given by

$$\sum_i w_i d_i / y_i$$

has an estimated variance (based on the Poisson distribution) of

$$\sum_i d_i (w_i / y_i)^2$$

and an estimated standard error of

$$\sqrt{\sum_i d_i (w_i / y_i)^2}$$

For the age-standardized rate, the weights are given by the number of persons in each age group per 100 000 in the standard population. For the cumulative rate, the weights are equal to the widths of the age groups. When all the age groups are five years across, the expression for the standard error of the cumulative rate (expressed as a percentage) reduces to

$$\text{S.E. (cumulative rate)} = \sqrt{\sum_i d_i / y^2_i}$$

For the example of stomach cancer among males in Denmark, the estimates with standard errors are 8.19 (0.20) for the rate standardized to the world population and 0.97 (0.03) for the cumulative rate (ages 0–74).

Cases of unknown age

As in previous volumes, the standardized rate and the cumulative rate have been corrected for cases of unknown age. The procedure involves simply multiplying either summary measure, based on cases of known age, by T/K, where T is the *total* number of cases of cancer of the same type in persons of the same sex and K is the number occurring in persons of *known* age. The standard errors were also multiplied by the same correction factor (T/K).

The correction relies on the assumption that the 'missing' cases are randomly distributed, and therefore have the same age distribution as the known cases. In other words, the probability that the age of a case is unknown does not depend on the age of the case. Although this assumption probably does not hold—it is more likely that age is not recorded in older cases—it is nevertheless important that *all* registered cases are accounted for, so that the summary statistics are not underestimated.

Comment

It cannot be stressed sufficiently that neither the age-standardized rate nor the cumulative rate are alternatives to the age-specific incidence rates, which should always be the starting point and foundation of any thorough analysis of the incidence data.

References

Ahmad, O.E., Boschi-Pinto, C., Lopez, A.D., Murray, C.J.L., Lozano, R. & Inoue M. (2000) *Age Standardization of Rates: a New WHO Standard* (GPE Discussion Paper Series: No. 31), Geneva: World Health Organization

Armitage P. & Doll R. (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer*, **8**, 1–12

Bray, F., Guilloux, A., Sankila, R. & Parkin, D.M. (2002) Practical implications of imposing a new World standard population. *Cancer Causes Control*, **13**, 175–182

Day, N.E. (1992) Cumulative rate and cumulative risk. In: Parkin, D.M., Muir, C.S., Whelan, S.L., Gao, Y.-T., Ferlay, J. & Powell, J. (eds), *Cancer Incidence in Five Continents, Volume VI* (IARC Scientific Publications No. 120), Lyon, IARC

Dos Santos Silva, I. (1999) *Cancer Epidemiology: Principles and Methods*, Lyon, IARC

Doll, R., Payne, P. & Waterhouse, J. (eds) (1966) *Cancer Incidence in Five Continents: A Technical Report*, Berlin, Springer-Verlag (for UICC)

Estève, J., Benhamou, E. & Raymond, L (1994) *Statistical Methods in Cancer Research, Vol. IV. Descriptive Epidemiology* (IARC Scientific Publications No. 128), Lyon, IARC

Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S. & Skeet, R.G. (eds) (1991) *Cancer Registration: Principles and Methods* (IARC Scientific Publications No. 95), Lyon, IARC

Segi, M. (1960) *Cancer Mortality for Selected Sites in 24 Countries (1950–57)*, Sendai, Tohoku University School of Public Health

Smith, P.G. (1992) Comparison between registries: age-standardized rates. In: Parkin, D.M., Muir, C.S., Whelan, S.L., Gao, Y.-T., Ferlay, J. & Powell, J. (eds) *Cancer Incidence in Five Continents, Volume VI* (IARC Scientific Publications No. 120), Lyon, IARC

United Nations (1999) *World Population Prospects: the 1998 Revision*. Volume 1: Comprehensive Tables. New York: United Nations