

# Chapter 7: Age standardization

Freddie Bray and Jacques Ferlay

## INTRODUCTION

For every type of cancer in every population worldwide, age is a key determinant of the risk of developing the disease (Armitage and Doll, 1954). Furthermore, the age structures of the registry catchment populations included in this volume vary considerably. For these two reasons, any global comparisons of incidence rates must be made independent of the effects of age in order to be meaningful. This chapter briefly discusses the summary measures used for this purpose in the preparation of *Cancer Incidence in Five Continents (CI5)*, allowing for comparisons of incidence across populations (within Volume XI) and within the same population over time (across consecutive volumes).

In each of the previous 10 volumes of CI5, we have emphasized that an age-specific examination of cancer incidence rates is an essential exploratory step in any comparative analysis, and Volume XI is no exception. Graphical displays of age-specific rates may reveal possible artefacts in the data or provide insight into the biology and etiology of cancer in the population or populations under study. The online resources available at <https://ci5.iarc.fr> (and the graphical facilities within) allow for a visual inspection that complements the tabulated summaries in the published volume.

To compare incidence across multiple populations or time points, we require a summary measure that absorbs the schedule of age-specific rates in different registries and across time. The crude rate does not take into account the varying age structures in the underlying populations. Therefore, as in previous volumes, two standardized measures – the age-standardized rate and the cumulative rate – are presented; this enables comparisons of cancer risk between registries, independent of the effects of age.

A brief description of the properties of these summary measures is given below, together with the principal calculations involved. Readers who wish to explore the topic in more depth should consult the chapters by Smith (1992) and Day (1992) in CI5 Volume VI. The books by Jensen et al. (1991), Dos Santos Silva (1999), and Kirkwood and Sterne (2003) include chapters that discuss in practical terms the age standardization of cancer data, and Estève et al. (1994) give a more theoretical account of the underlying methodology.

## CALCULATING AGE-STANDARDIZED RATES

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 population. Calculation of the standardized rate is an example of direct standardization, whereby the observed age-specific rates are applied (directly) to a theoretical standard population. The calculation is illustrated in Table 7.1, for stomach cancer incidence among females in Yanting County, China (2008–2012), using the world standard population first introduced by Segi (1960), drawn from a pooled population of 46 countries and modified for the first volume of this series by Doll et al. (1966). The Segi–Doll world standard is used in this and all previous volumes of CI5 as a means to rapidly examine geographical and temporal variations in cancer risk.

The main criticism levelled at age-standardized rates concerns the need to select an arbitrary standard population; age-standardized rates can only be meaningfully compared if they refer to the same standard. Although it is clear that the age composition of the world standard is representative of neither the present nor the future age-specific global population, switching to a more up-to-date standard would be worthwhile only if it could provide benefits that outweighed the drawbacks of rendering the Segi–Doll standardized rates obsolete. However, it has been shown that although very different absolute values of age-standardized rates are obtained using different standards, the estimates of age-adjusted relative risk (the main purpose of such routine comparisons) are quite similar (Bray et al., 2002). Therefore, the main effect of changing the standard to a more up-to-date representation of the global population would be the inconsequential result that the value of the age-standardized rate would be closer to that of the crude rate. Accordingly, for both theoretical and practical reasons, the age-standardized rates in this volume are calculated as previously, using the age composition of the Segi–Doll world standard, shown in Table 7.1.

Age groups are indexed by the subscript  $i$ ,  $d_i$  is the number of cases in age group  $i$ ,  $y_i$  is the number of person-years at risk in age group  $i$ , and  $w_i$  is the number of individuals in (or the weight of) age group  $i$  in the world standard population. The crude rate per 100 000 per year is therefore calculated as follows (using the data presented in Table 7.1 as an example):

$$10^5 \left( \frac{\sum_i d_i}{\sum_i y_i} \right) = 10^5 \times 1270 / 1451646 = 87.49$$

**Table 7.1. Calculation of age-standardized incidence rates: stomach cancer (ICD-10 C16), China, Yanting County, females (2008–2012)**

Age group (years) <i>i</i>	No. of cases <i>d<sub>i</sub></i>	Person-years at risk <i>y<sub>i</sub></i>	Age-specific incidence (per 10 <sup>5</sup> years) $10^5(d_i / y_i)$	Standard world population <i>w<sub>i</sub></i>	Expected no. of cases in standard world population <i>d<sub>i</sub> w<sub>i</sub> / y<sub>i</sub></i>
0–4	0	84 468	0.00	12 000	0.00
5–9	0	81 670	0.00	10 000	0.00
10–14	0	72 068	0.00	9 000	0.00
15–19	0	106 093	0.00	9 000	0.00
20–24	0	146 968	0.00	8 000	0.00
25–29	0	149 108	0.00	8 000	0.00
30–34	6	133 780	4.48	6 000	0.27
35–39	35	124 408	28.13	6 000	1.69
40–44	58	102 229	56.74	6 000	3.40
45–49	90	92 643	97.15	6 000	5.83
50–54	117	83 769	139.67	5 000	6.98
55–59	131	75 003	174.66	4 000	6.99
60–64	175	66 514	263.10	4 000	10.52
65–69	171	51 164	334.22	3 000	10.03
70–74	162	37 469	432.36	2 000	8.65
75–79	148	23 505	629.65	1 000	6.30
80–84	95	13 259	716.49	500	3.58
85+	82	7 528	1089.27	500	5.45
Total	1 270	1 451 646	87.49	100 000	69.68

And the age-standardized rate is calculated as follows:

$$\sum_i d_i w_i / y_i = 69.68$$

In this example, the age-standardized rate is 20% less than the crude rate in the same 5-year period. This is because the world standard population has proportionally fewer individuals in the older age groups than does the actual registry population, and the risk of disease (according to the age-specific rates) is higher at older ages.

#### **CUMULATIVE RATE AND CUMULATIVE RISK**

The cumulative rate is the sum 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 5-year age intervals, the cumulative rate is 5 times the sum of the age-specific rates calculated over the 5-year age groups, assuming that the age-specific rates are the same for all ages within a 5-year age group. In the example of stomach cancer incidence in females in the China, Yanting County registry population, all the age groups have a width of 5 years, so the cumulative rate for ages 0 to 74 years is calculated as follows:

$$\begin{aligned} 5 \sum_i d_i / y_i &= 5 \times 1330.51 \times 10^{-5} \\ &= 0.0765 \end{aligned}$$

The cumulative rate is in fact not a rate at all but a dimensionless quantity most conveniently expressed as a percentage. In the example above, the cumulative rate up to the age of 74 years would be given as 7.65%.

The age span over which a rate is accumulated must be specified. In this volume, the age ranges 0–64 years and 0–74 years are used to provide two representations of the lifetime risk of developing the disease. Other age ranges may be more appropriate for more specific purposes, such as investigating childhood cancers.

The cumulative rate's companion – 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 causes of death. If the cumulative risk in an age range is less than 10%, as is the case with most tumours, it can be approximated very well by the cumulative rate. Table 7.2 shows the correction needed to convert the cumulative rate into the cumulative risk. For values of less than 10%, the difference between the two is small. The precise mathematical relationship between the cumulative rate and the cumulative risk is as follows:

$$\text{cumulative risk} = 1 - \exp(-\text{cumulative rate})$$

Continuing with our example, the cumulative risk is 7.37%, and thus the estimated risk for a female in Yanting County of developing stomach cancer between the ages of 0 and 74 years is estimated as 7.4%. In other words, approximately 1 in 14 females

**Table 7.2. Conversion of cumulative rates,  $100x$ , into the corresponding cumulative risks,  $100(1-e^{-x})$**

<b>100x</b>	0.1	0.5	1.0	5.0	7.0	10.0	15.0	20.0	30.0	40.0	50.0
<b>100(1-e<sup>-x</sup>)</b>	0.1	0.499	0.995	4.88	6.76	9.52	13.93	18.13	25.92	32.97	39.35

would develop stomach cancer in the defined region within the specified lifetime, assuming no other causes of death.

The cumulative rate has two principal advantages over the age-standardized rate. First, because it is a form of direct standardization, the problem of choosing an arbitrary reference population is eliminated. Second, as an approximation of 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.

#### **CALCULATING THE STANDARD ERROR**

Both the age-standardized rate 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, then the age-standardized rate (with  $w_i$  representing the standardization weights) is

$$\sum_i d_i w_i / y_i$$

with 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 individuals 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 age groups are 5 years wide, the expression for the standard error (SE) of the cumulative rate (expressed as a percentage) reduces to:

$$SE(\text{cumulative rate}) = 500 \sqrt{\sum_i d_i / y_i^2}$$

For the example of stomach cancer in females in the China, Yanting County registry population, the estimates (with standard errors) are 69.7 (1.98) for the rate standardized to the world population and 7.65 (0.264) for the cumulative rate (ages 0–74 years).

#### **CASES OF UNKNOWN AGE**

The age-standardized rate and the cumulative rate are necessarily corrected for cases of unknown age that occur within the CI5 registry datasets. The correction procedure involves multiplying the 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 individuals of the same sex, and  $K$  is the number of cancers occurring in individuals of known age. The standard errors are also multiplied by the same correction factor ( $T/K$ ).

This correction relies on the assumption that the cases of unknown age are randomly distributed, and therefore have the same age distribution as the cases of known age. In other words, the correction assumes that 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 among older patients – it is nevertheless important that all registered cases are accounted for, so that the summary statistics are not underestimated.

#### **AVAILABILITY OF DATA FOR UPPER AGE GROUPS**

The precision of rates calculated using direct standardization varies according to the number and width of the age groups used. Age-standardized rates, for example, yield slightly different values depending on the extent to which population data are available for the oldest age groups (aged 75 years or more). For most registries, census data are available for the same 18 age groups shown in Table 7.1, enabling the calculation of rates for the upper age ranges of 75–79, 80–84, and 85+. However, some registries can only obtain population data for 16 or 17 age groups, and thus the oldest age groups instead correspond to individuals aged 75+, or 80+, respectively. As an example of the effect of this difference, the age-standardized rate of 69.7 for stomach cancer incidence in females in the China, Yanting County registry population changes to 69.0 when only 16 age groups are used in the calculation instead of all 18 age groups.

**REFERENCES**

- Armitage P, Doll R (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer*. 8(1):1–12. <https://doi.org/10.1038/bjc.1954.1> PMID:13172380
- Bray F, Guilloux A, Sankila R, Parkin DM (2002). Practical implications of imposing a new world standard population. *Cancer Causes Control*. 13(2):175–82. <https://doi.org/10.1023/A:1014344519276> PMID:11936824
- Day NE (1992). Cumulative rate and cumulative risk. In: Parkin DM, Muir CS, Whelan SL, Gao Y-T, Ferlay J, Powell J, editors. *Cancer Incidence in Five Continents*, Vol. VI. IARC Scientific Publication No. 120. Lyon: International Agency for Research on Cancer.
- Doll R, Payne P, Waterhouse J, editors (1966). *Cancer Incidence in Five Continents: A Technical Report*. Berlin: Springer-Verlag (for UICC).
- Dos Santos Silva I, editor (1999). *Cancer Epidemiology: Principles and Methods*. IARC Non-Serial Publication. Lyon: International Agency for Research on Cancer. Available from: <https://publications.iarc.fr/421>.
- Estève J, Benhamou E, Raymond L (1994). *Statistical Methods in Cancer Research*, Vol. IV. Descriptive Epidemiology. IARC Scientific Publication No. 128. Lyon: International Agency for Research on Cancer.
- Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors (1991). *Cancer Registration: Principles and Methods*. IARC Scientific Publication No. 95. Lyon: IARC Press. Available from: <https://publications.iarc.fr/238>.
- Kirkwood BR, Sterne JAC (2003). *Essential Medical Statistics*. 2nd ed. Oxford: Blackwell Science.
- Segi M (1960). *Cancer mortality for selected sites in 24 countries (1950–57)*. Sendai, Japan: Department of Public Health, Tohoku University School of Medicine.
- Smith PG (1992). Comparison between registries: age-standardized rates. In: Parkin DM, Muir CS, Whelan SL, Gao Y-T, Ferlay J, Powell J, editors. *Cancer Incidence in Five Continents*, Vol. VI. IARC Scientific Publication No. 120. Lyon: International Agency for Research on Cancer.