

Appendix B: Statistical methods and technical notes

Age-specific rates

Age-specific rates provide information on the incidence of a particular event in an age group relative to the total number of people at risk of that event in the same age group. It is calculated by dividing the number of events occurring in each specified age group by the corresponding 'at risk' population in the same age group and then multiplying the result by a constant (e.g. 100,000) to derive the rate. Age-specific rates are often expressed per 100,000 population.

Age-standardised rates

A crude rate provides information on the number of, for example, new cases of cancer or deaths from cancer by the population at risk in a specified period. No age adjustments are made when calculating a crude rate. Since the risk of cancer is heavily dependent on age, crude rates are not suitable for looking at trends or making comparisons across groups in cancer incidence and mortality.

More meaningful comparisons can be made by the use of age-standardised rates, with such rates adjusted for age in order to facilitate comparisons between populations that have different age structures, for example, between Indigenous peoples and other Australians.

There are two methods commonly used to adjust for age: direct and indirect standardisation. In this report, the direct standardisation approach presented by Jensen and colleagues (1991) is used. To calculate age-standardised rates, age-specific rates (usually grouped in 5-year intervals) are multiplied against a constant population – either the Australian population as at 30 June 2001 or the World Health Organization (WHO) 2000 World Standard Population. This effectively removes the influence of age structure on the summary rate and it is described as the age-standardised rate.

Confidence intervals

An observed value of a rate may vary due to chance, even where there is no variation in the underlying value of the rate. A confidence interval provides a range of values that has a specified probability of containing the true rate or trend. The 95% (p -value = 0.05) confidence interval is used in this report; thus, there is a 95% likelihood that the true value of the rate is somewhere within the stated range. Confidence intervals can be used as a guide to whether or not differences are consistent with chance variation. In cases where no values within the confidence intervals overlap, the difference between rates is greater than that which could be explained by chance and is regarded as statistically significant. Note, however, that overlapping confidence intervals do not necessarily mean that the difference between two rates is definitely due to chance. Instead, an overlapping confidence interval represents a

difference in rates which is too small to allow differentiation between a real difference and one which is due to chance variation. It can, therefore, only be stated that no statistically significant differences were found, and not that no differences exist. The approximate comparisons presented might understate the statistical significance of some differences, but they are sufficiently accurate for the purposes of this report.

As with all statistical comparisons, care should be exercised in interpreting the results of the comparison of rates. If two rates are statistically significantly different from each other, this means that the difference is unlikely to have arisen by chance. Judgement should, however, be exercised in deciding whether or not the difference is of any practical significance.

With one exception, the confidence intervals presented in this report were calculated using a method developed by Dobson and associates (1991). This method calculates approximate confidence intervals for a weighted sum of Poisson parameters.

The one exception applies to the confidence intervals that were calculated for the international comparisons of incidence and mortality data using GLOBOCAN data, as shown in Figures 2.5 and 3.5. For those data, the lack of the required data meant that the Dobson method could not be used and the AIHW approximated the confidence intervals using the following formula:

$$95\% \text{ CI approximation} = \text{AS rate} \pm 1.96 \times \frac{\text{AS rate}}{\sqrt{\text{Number of cases}}}$$

Since the GLOBOCAN data are based on the estimates of the number of new cases and deaths from breast cancer, the associated confidence intervals indicate the range of random variation that might be expected, should those estimates be 100% accurate.

Note that statistical independence of observations is assumed in the calculations of the confidence intervals for this report. This assumption may not always be valid for episode-based data (such as data from the National Hospital Morbidity Database and Medicare Australia).

Incidence projections

To calculate the incidence projections shown in Chapter 2, breast cancer incidence data for females for the 10-year period from 1997 to 2006 were divided into 18 series – one for each 5-year age group. The incidence numbers were divided by the age-specific mid-year populations to obtain the age-specific incidence rates. Least squares linear regression was used to find the straight line of best fit through the 1997 to 2006 rates and to compute the various quantities needed for the 95% prediction intervals. The projected incidence rates were then multiplied by the estimated resident population to obtain the projected incidence numbers. The populations used were the Australian Bureau of Statistics (ABS) projected populations from Series 29(B) (ABS 2008b).

Mortality-to-incidence ratio

Both mortality-to-incidence ratios (MIRs) and relative survival ratios can be used to estimate survival from a particular disease, such as breast cancer, for a population. Although MIRs are the cruder of the two ratios, deriving MIRs is far less complicated. Thus, the MIR is considered to be a better measure when comparing survival between countries.

The MIR is defined as the age-standardised mortality rate divided by the age-standardised incidence rate. For example, an MIR of 0.42 in a given year for all types of cancers means that for every 100 new cancer cases diagnosed that year, there were 42 deaths due to cancer in the same year (though the deaths need not be of the same people as the cases). If people tend to die relatively soon after diagnosis from a particular cancer (that is, the death rate is nearly as high as the incidence rate for that cancer), then the MIR will be close to 1.00. In contrast, if people tend to survive a long time after being diagnosed, then the MIR will be close to zero.

The MIR only gives a valid measure of the survival experience in a population if:

- cancer registration and death registration are complete or nearly so, and
- the incidence rate, mortality rate and survival proportion are not undergoing rapid change.

The incidence and mortality data used to calculate the MIRs in Chapter 4 were extracted from the 2002 GLOBOCAN database (Ferlay et al. 2004).

Relative survival analysis

Relative survival estimates compare the survival of persons diagnosed with breast cancer (i.e. the observed survival) with the survival of the entire Australian population of the same sex and age in the same calendar year as the cancer cohort (i.e. the expected survival). Note that the actual cause of death (whether it is from breast cancer or another cause) is not of importance in these analyses. Thus, relative survival is defined as follows:

$$\text{relative survival} = \frac{\text{observed survival for cancer cohort}}{\text{expected survival for 'matched' population}}$$

The resulting value is usually given as a proportion. For example, if the observed 5-year survival of a particular cohort diagnosed with breast cancer was 0.80 (that is, 80% of them were still alive 5 years after diagnosis) and their expected survival, based on Australian life-tables, was 0.90 (that is, 90% of people with the same age- and sex-profile as the cohort would be expected to be alive 5 years later), then the 5-year relative survival would be $0.8/0.9 = 0.89$ or 89%. One way to interpret this figure is that the 'average' person in the cancer cohort has an 89% chance of being alive 5 years after diagnosis *relative to others of the same sex and age*.

In order for the relative survival estimate to be a valid approximation of the probability that a person will not die of their diagnosed cancer within the given time interval, the presence of the cancer is assumed to be the only factor that distinguishes the cancer cohort from the general population (Ries et al. 2008). The degree to which this is true is not known.

Relative survival proportions have traditionally been calculated using the 'cohort method' and National Breast and Ovarian Cancer Centre preferred the use of that method for this report. In the cohort method, a cohort of people diagnosed with cancer is followed over time to estimate the proportion surviving for a selected time frame (e.g. 1, 5 or 10 years). An alternative approach to calculating relative survival is the period method which was developed by Brenner and Gefeller (1996). This method examines the survival experience of people who were alive at the beginning of a particular recent calendar period and who were diagnosed with cancer before this period. Therefore, the period method might provide more up-to-date estimates of survival, especially in the presence of temporal trends affected by improvements in cancer detection and treatment. However, the cohort method is thought to provide more precise estimates (i.e. estimates with narrower confidence intervals).

An alternative to the calculation of relative survival proportions is to use the 'cause-specific model' to derive survival estimates. This model calculates survival based on deaths due to cancer-related causes alone. There are various advantages and disadvantages to using the cause-specific model (Le Teuff et al. 2005). Because the 2006 version of the Australian Cancer Database (ACD) that was utilised for this report included a limited amount of cause of death information, this approach could not be used to calculate survival estimates.

Data from the ACD on the incidence of breast cancer were used to calculate observed survival proportions. These incidence data were linked to the National Death Index in order to obtain information on those people with breast cancer who died and the date on which this occurred (see Appendix C for more information on these data sources). In order to calculate the expected survival belonging to the age-, sex- and calendar-year matched population, life tables for the population under study were used. These life tables were obtained from the Australian Bureau of Statistics (ABS 2009b).

The software used to calculate the relative survival proportions was written by Dickman (2004). It uses the Ederer II method of calculating the interval-specific expected survivals. Further details on the approach used to calculate the relative survival estimates, including rules which were applied during data preparation, can be found in the 2008 report prepared by the AIHW on cancer survival and prevalence (AIHW, CA & AACR 2008).

Risk to age 75 and 85 years

The calculations of risk shown in this report are measures that approximate the risk of developing (or dying from) breast cancer before a given age, assuming that the risks at the time of estimation remained throughout life. It is based on a mathematical relationship with the cumulative rate. Note that in these risk factors, no account is taken of specific breast cancer risk factors. Further details on how the risks were calculated can be found in the 2008 *Cancer in Australia* report (AIHW & AACR 2008).