Chapter 2

Stastistical methods for cancer survival analysis

Swaminathan R and Brenner H

Abstract

Adequate and complete follow-up is a prerequisite for the conduct of any survival study. Passive follow-up relies on routine availability of mortality data through unique data linkage possibilities, while active follow-up supplements mortality ascertainment, for which there are a variety of methods. Cox proportional-hazard model was employed to test whether censoring was random in presence of loss to follow-up. Absolute survival probability was estimated by the actuarial method following semi-complete approach for all registries, and the period approach was also used wherever possible. Expected survival probability for registries was estimated from the respective country-, age- and sex-specific abridged life tables. Relative survival, as the ratio of absolute to expected survival, was calculated to exclude the effect arising from different background mortalities. To account for the differences in the age structure of the cancer cases, relative survival was adjusted for age and reported as age-standardized relative survival. Estimated incident cancer cases from less-developed countries together for every classified cancer site served as the standard population. Weights were assigned to individual patients, depending on their age, and standardization was carried out using weighted individual data. Analyses were done using the publicly available macros in SAS software.

Introduction and background

The life table, one of the basic tools in the description of mortality experience of a population, was first developed as early as 1693 by E. Halley in England. It forms the basis for calculation of the life table estimate of the survivor function, which is still widely used today in the analysis of data from epidemiological studies. Information on survival has long been recognized as an important component in monitoring cancer control activities [1]. Like all other health indices, survival statistics are useful primarily as comparative measures. It is these comparisons that help us to suggest possible reasons for the variations and provide targets for improvement and a means of monitoring progress towards them [2]. Survival data obtained from a population-based cancer registry ideally portrays the average outcome of the disease in the pertaining region covered since it is based on an unselected series of incident cancer cases [3].

Follow-up

Adequate and complete follow-up is a prerequisite to conducting a survival study. Lengthy periods of time may be required until the event of interest (any death is the outcome studied in this publication) occurs in all cases studied and maintenance of surveillance on patients may be extremely difficult. Hence, a closing date for follow-up is typically imposed keeping in mind the adequacy of follow-up information needed to estimate the survival at a specified time. Complete follow-up is deemed to have been achieved when the vital status (alive/dead) at closing date is known for an individual. If not known, then the follow-up is incomplete.

With passive follow-up, information on deaths is routinely received either by-law or via an arrangement with the vital statistics division. Using this procedure, those patients for whom no information of death has been received may be considered to be "alive" until that point of time. The main requirement for this method to work efficiently is that there is a high quality of registration of mortality data and unique data linkage possibilities which ensure the follow-up of cases to be complete with the exception of migration or rare losses. A few of the registries contributing data to this scientific publication have relied almost entirely on this means of obtaining follow-up information.

Active follow-up is necessary in the absence of a reliable health information system, and it may supplement the latter in case of incomplete passive follow-up. Most registries that contributed data to

this scientific publication generally resorted to this method after the routine matching of the incident cancer cases with the available mortality information was completed. The different ways by which this is accomplished are by repeated scrutiny of medical records in hospitals, enquiries with attending physicians, scanning the population registers (city directories), health registers of national health services, health insurance registers, electoral lists, postal/telephone enquiries and visits to the homes of the cases or persons known to them.

Censoring

It is impractical to continue follow-up until all cases under study are dead. With a closing date of follow-up in place, for the subjects who are withdrawn wilfully, drop out or are lost from the study before this date and for those who are still alive at this date, only a lower limit on lifetime is available. This is not to conclude that no information is available on them, but that the information is partial. This unique feature in lifetime data analysis, which occurs when exact lifetimes until death are known for only a portion of the individuals in the study and known to exceed certain values in the remainder, is called "censoring".

When censoring occurs, either due to the termination of study at the closing date which is solely technical or due to loss to follow-up that is 'unrelated' to the outcome studied, e.g. death, it is said to be random or non-informative censoring. When censoring occurs due to loss of follow-up which is 'related' to death, it is known as non-random or informative censoring.

Test for random censoring

Little reliance can be placed on the estimated survival assuming random censoring when the magnitude of loss to follow-up is high. In such instances, it is desirable to investigate deviation from randomness of censoring. In this publication, the Cox proportional-hazard model [4] was used whenever the censoring before closure of study or loss to follow-up exceeded 10% of total cases. For this purpose, the outcome studied is the "loss to follow-up" within a specified time from the index date. Since the survival is estimated at five years for the majority of cancer registries in this publication, the time is fixed as five years. All cases censored before closure of the study and having had a follow-up of less than five years constitute the loss to follow-up group, and the rest of the cases who are either dead or known to be alive on the closing date of follow-up are treated as censored for this analysis to detect the presence of informative censoring. Since the Cox model deals with survival time dynamically, the varying patterns of every loss to follow-up at different intervals on the survival time scale are well accounted for. Based on the general availability, the variables or determinants that are tested for association with loss to follow-up are age at diagnosis, sex and extent of disease. An example of this type of analysis is given in Table 1, where the proportion of patients lost to follow-up ranges between 7–16% among categories of age at diagnosis and 0–27% among categories of extent of disease. A statistically significant differential risk of loss to follow-up is observed. This suggests the presence of non-randomness of loss to follow-up and, therefore,

Table 1. Example of test for randomness of loss to follow-up: Cox proportional-hazards model

Registry	:	**Mumbai**
Site of cancer	:	**Female breast**
Period of registration of cases	:	**1992–1994**
Period of follow-up	:	**1992–1999**
Event studied	:	**Lost to follow-up before 31st December 1999 and having a follow-up of <5 years**
% Loss to follow-up	:	**10.9%**

Determinants of loss to follow-up	Lost to follow-up		Relative hazard of loss to follow-up[$]	
	Number	%	Hazard ratio	95% CI
Age at diagnosis				
≤ 44 years	53	6.9	1.00	-
45–54	75	10.4	1.63	1.14–2.31*
55–64	89	16.2	2.47	1.76–3.47*
65–74	47	13.7	2.25	1.52–3.33*
75+	9	7.4	1.26	0.62–2.56
Extent of disease				
Localized	129	14.6	1.00	-
Regional	98	8.2	0.54	0.41–0.71*
Distant metastasis	1	0.4	0.05	0.01–0.35*
Unknown	45	26.9	2.40	1.70–3.38*

[$] *Each factor is adjusted for the other in the table; CI: Confidence interval; * p ≤0.05.*

the survival estimates assuming random censoring should be interpreted with caution.

Actuarial method of estimation of absolute survival probability

It is rare to find a closed group of subjects in a survival study without censoring, except possibly in an artificial situation such as the construction of a life table. The actuarial method of estimating survival probability [5] handles censoring by assuming it to be random. This method involves the construction of a life table that permits the calculation of the cumulative probability of survival at time t_{i+1} from the conditional probabilities of survival during consecutive intervals of follow-up time up to < t_{i+1}. This method has been used in this publication to estimate the absolute survival probability. The layout and method of calculation of the elements of a life table are illustrated in Table 2 [6].

For each time period t_i to t_{i+1}, n_i is the number of subjects at risk of outcome at the beginning of the time interval. The number of cases censored during the interval, because they are lost to follow- up or withdrawn alive at the end of the follow-up period, is shown as w_i. The symbol d_i denotes subjects who experienced the outcome during each interval. The effective number of subjects at risk during each interval is calculated as:

$$N_i = n_i - (\frac{w_i}{2})$$

In this way, subjects who are alive and at risk of experiencing the outcome during the interval t_i to t_{i+1}, but who are censored at some point of time during the interval, are assumed to have been followed up for, on average, half of the interval. This actuarial assumption is based on the censorings being independent of the outcome studied (i.e., any death, in this publication). The probability of occurrence of the outcome during the interval is given by

$$q_i = \frac{d_i}{N_i}.$$

The probability of survival during the interval beginning t_i is then calculated as

$$p_i = 1 - q_i$$

from which the cumulative probability of survival up to time t_{i+1} is derived from the product of the p_i's

$$P_{i+1} = \prod_{j=0}^{i} p_j$$

This quantity p_{i+1} is often multiplied by 100 to give the "percentage survival" at time t_{i+1}.

Different approaches

There are several approaches to estimating the absolute survival at a given time by varying the registration and follow-up periods of time. These are discussed below and illustrated in Figure 1.

Cohort analysis

The simplest way of computing survival probability is to compute the ratio or percentage of the number of subjects alive at the end of, e.g., 5 years from the index date by the total number of subjects in the study at the beginning of the study, excluding those who did not have a chance to be followed for 5 years

Table 2. Illustration of the layout of the life table and calculation of cumulative survival probability by the actuarial method

Interval	Alive at beginning of interval	Last known alive during interval (consored)	No. of deaths during interval	Effective number at risk	Conditional probability of death	Conditional probability of survival	Cumulative probability of survival (to end of interval)
t_i - t_{i+1}	n_i	w_i	d_i	N_i	q_i	p_i	P_{i+1}
0–1	3289	166	365	3206.0	0.114	0.886	0.886
1–2	2758	275	301	2620.5	0.115	0.885	0.784
2–3	2182	37	278	2163.5	0.128	0.872	0.683
3–4	1867	30	191	1852.0	0.103	0.897	0.613
4–5	1646	20	106	1636.0	0.065	0.935	0.573

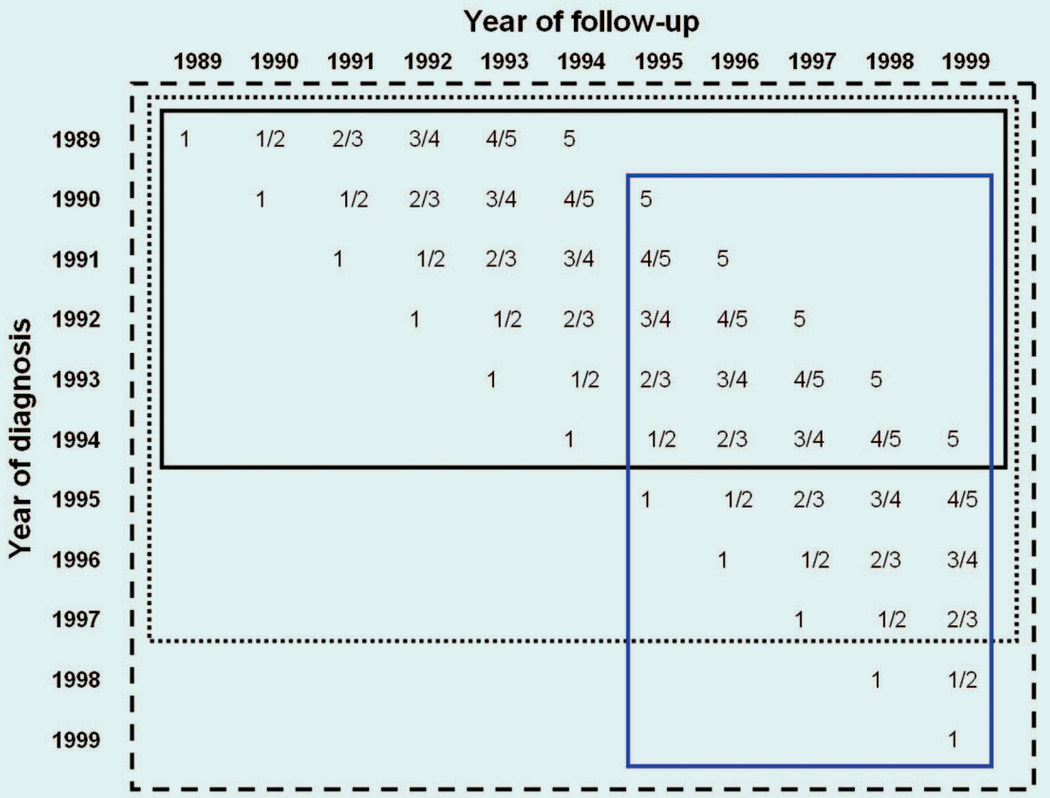
Source: Black and Swaminathan (1998)

after diagnosis. For this purpose, only subjects potentially under observation for at least 5 years and having a potentially complete follow-up of five years are taken into consideration. This approach, which has been called cohort analysis [7] has the disadvantage that even the most recent survival estimates are exclusively based on patients diagnosed many years ago. For example, with a database that includes patients diagnosed between 1989 and 1999 with a closing date of follow-up at the end of 1999, a cohort estimate of 5-year survival could be obtained from patients diagnosed in 1994 at the latest, because patients diagnosed in later years could not possibly have 5-year follow-up by the end of 1999. This approach is illustrated by the solid black frame in Figure 1.

Complete analysis

This is the approach to be used when there is no restriction on the potential follow-up time to equal, e.g., five years from the index date for which the survival is estimated. Rather, all subjects who are diagnosed as incident cancers until the closing date of the follow-up period qualify for inclusion in the analysis. Apart from the subjects with a complete follow-up of five years, those under observation for a variable period of time and having an incomplete follow-up of less than five years are included [7]. In the example given above, all patients diagnosed in 1995–1999 could be included in addition to those diagnosed in earlier years for the derivation of a complete estimate of 5-year survival. This approach is illustrated by the dashed black frame in Figure 1.

Figure 1. Types of analysis to derive up-to-date 5-year survival estimates based on data of patients diagnosed in 1989–1999 and followed until the end of 1999



Cohort analysis: solid black frame;
Complete analysis: dashed black frame;
Semi or partially complete analysis: dotted black frame;
Period analysis: solid blue frame.
The numbers within the cells indicate the years since diagnosis.

Semi or partially complete analysis

This approach is widely practised in the estimation of survival by cancer registries. It was adopted in the previous publication on cancer survival [6], and is used for most analyses in this publication as well. Here, not all patients diagnosed until the closing date of follow-up are included. Rather, only patients who have had some minimum potential follow-up time at the closing date of follow-up, such as two or three years, are included. In our example, a partially complete estimate of 5-year survival may be obtained from patients diagnosed in or before 1997 and who have had a minimum of two potential years of follow-up at the end of 1999. This approach, which is in between the pure cohort and pure complete analysis, is illustrated by the dotted black frame in Figure 1.

Period analysis

This is an alternative approach [8] to deriving more up-to-date estimates of cancer patient survival by exclusively utilising the survival information pertaining to the most recent incidence and follow-up periods. The period of interest could be a single calendar year or more. Period analysis exclusively reflects the survival experience of subjects within the most recent calendar period for which the follow-up is available. This is achieved by left truncation of observations at the beginning of this period in addition to censoring at its end [9].

In our example, assume that a period estimate of 5-year survival is to be derived for the 1995–1999 period, the most recent period for which pertinent data are available, then all observations are left truncated at the beginning of 1995 in addition to being censored at the end of 1999. The 5-year period estimate of survival would be obtained from patients diagnosed in 1990–1999 for whom some proportion of 5-year follow-up might have fallen in the 1995–1999 period. With this approach, illustrated by the solid blue frame in Figure 1, different parts of the survival function would be derived from patients diagnosed in various calendar years. Survival during the first year following diagnosis would be estimated for patients diagnosed in 1994–1999, survival during the second year following diagnosis would be estimated for patients diagnosed in 1993–1998, and so on, until survival experience during the fifth year following diagnosis which would be obtained for patients diagnosed in 1990–1995. These conditional survival probabilities are then combined in the usual way to generate 5-year cumulative survival estimates for the 1995–1999 period. It has been shown that period analysis is the approach that clearly provides the most up-to-date estimates of cancer patient survival, and that period estimates of survival for some given period quite closely predict survival experience of patients diagnosed during that period [10]. In this publication, however, period analysis could not routinely be used because incidence data had not been collected up to the closing date of follow-up by most registries. That said, period analysis was used with data from registries in Qidong and Tianjin, China, and Singapore. A comparison of the survival estimates by cohort and period approaches has been done and the trends over calendar time were depicted.

Relative survival

Berkson [11] in 1942 introduced the concept of relative survival. The relative survival (R_i) for a group of patients at the end of an interval beginning at time t_i is defined as

$$R_i = \frac{S_i}{S_i^*}$$

where S_i is the absolute survival for subjects with a particular cancer and S_i^* is the expected survival of a group of individuals with the same demographic characteristics (age, sex, etc.) who are at risk of death only from causes other than the cancer under study [12]. Berkson and Gage [13] suggested that the observed proportion of survivors of cancer can be compared with an expected proportion of survivors derived from similar people from the general population, most of whom do not have the disease under study. The concept of relative survival methodology has primarily been designed for cancer survival studies to exclude the effect arising from different background mortalities.

Estimation of expected survival probabilities

Expected survival probabilities are usually estimated from age- and sex-specific (sometimes also race-specific) life tables of the general population for the registry area. At least three different methods have been proposed to estimate expected survival, the so-called Ederer I [12], Ederer II [14] and Hakulinen [15] methods. For follow-up times up to 5 years (as reported in this publication) they generally give very similar results. In this study, expected survival probabilities are estimated from country-, age- and sex-specific abridged life tables [16] according to the Ederer II method [14] (for 5-year survival) and the Hakulinen method [15] (for 10- and 15-year survival), the latter of which corrects for potential heterogeneity in patient withdrawal over long potential follow-up times. The estimation of expected survival for earlier calendar periods is done using the country-, age- and sex-specific life tables of the respective calendar periods.

Age-standardization of survival

Most biological phenomena are related to age; there is no reason to expect that survival is not. It is important to note that use of relative rather than absolute survival does not make age-standardization unnecessary. For many types of cancer, the risk of dying as a result of the cancer itself is clearly associated with a subject's age at diagnosis. The ages at diagnosis of cases of any cancer in the developing and developed countries are vastly different [17]. When comparing survival in different groups of patients from different regions, there is a definite need to standardize both absolute and relative survival estimates for age.

For this purpose, direct standardization of survival estimates has been advocated [18]. This is commonly done by using direct standardization of age-specific survival estimates to derive summary statistics called age-standardized absolute survival (ASAS) or age-standardized relative survival (ASRS). For example, ASAS at the end of some follow-up period i is given by

$$ASAS_i = \frac{\sum_x a_{ix} st_x}{\sum_x st_x}$$

where the a_{iX} are age-specific (x:0–4;5–9; etc.) absolute survival estimates at the end of follow-up period ti and stx are the age-specific proportions used as "standard or weight" for standardization. The st_x could be arbitrary. Traditionally, the weights have been chosen to reflect the age distribution at diagnosis of some standard cancer population, such as the world standard cancer population [19].

However, for relative survival, the traditional age-standardization, as outlined above, provides results that are conceptually different from crude survival data [20]. Furthermore, traditional age-standardization is often difficult if not impossible to carry out in the presence of sparse and censored data. Hence, in this publication, an alternative approach to age-standardization [21] has been adopted. In this approach, one first assigns the weights to the individual patients depending on their age and then carries out conventional survival analyses using the "weighted individual data". The weights are defined as the ratio of the proportion of patients in the respective age group (x) in the standard population (st) divided by the proportion of patients in the respective age group in the study population. Whereas in the unadjusted (crude analyses), each patient in the study population and her/his contributions to the numbers of persons at risk and deaths are (implicitly) entered with a weight of 1, the proposed form of age-adjustment gives weights higher (lower) than 1 to patients in age groups which are under-represented (over-represented) in the study population compared to the standard population. The advantages of doing this type of adjustment are: (i) it remains feasible with sparse data, even in situations where survival estimates cannot be derived for certain age groups, and (ii) it provides age-adjusted estimates of relative survival that are conceptually consistent with the crude estimates. In particular, age-adjustment to the study population's own age structure yields a standardized relative survival that is identical to the crude one.

In this study, the weights are defined as the ratio of the proportion of patients in the respective age group in the standard population as summarized in GLOBOCAN 2002 [22], divided by the proportion of patients in the respective age group in the study population registry for every classified cancer site/type.

Software used

While absolute survival can be estimated with any of a large number of commercially available statistical software packages, there are only few specialized programs for relative survival analysis. In this study, analyses are done using the publicly available SAS macros "period" or "periodh" (age-specific and crude analysis, [10]) or "adperiod" or "adperiodh" (age-adjusted analysis, [21]) , which can be used to calculate both absolute and relative survival (Ederer and Hakulinen methods) with either the cohort, semi-complete or period approach.

References

1. WHO/IARC. *Cancer Statistics: Report of a WHO/IARC Expert Committee, WHO Technical Report Series No. 632*. World Health Organization, Geneva, 1979.

2. Black RJ, Sankaranarayanan R and Parkin DM. Interpretation of population based cancer survival data. In: *Cancer Survival in Developing Countries*. (eds) R. Sankaranarayanan, RJ Black and DM Parkin. IARC Scientific Publications No. 145. IARCPress, Lyon, 1998, pp 13–17.

3. Sankaranarayanan R, Black RJ and Parkin DM. Eds. *Cancer Survival in Developing Countries*. IARC Scientific Publications No. 145. IARCPress, Lyon, 1998.

4. Cox DR. Regression models and life tables. *J R Statistic Soc (B)* 1972;34: 187–220.

5. Cutler SJ, Ederer F. Maximum utilization of the life table method in analysing survival. *J Chron Dis*. 1958;8: 699–712.

6. Black RJ and Swaminathan R. Statistical methods for the analysis of cancer survival data. In: *Cancer Survival in Developing Countries* (eds) Sankaranarayanan R, Black RJ and Parkin DM. IARC Scientific Publications No. 145. IARCPress, Lyon, 1998; pp 3–7.

7. Brenner H, Gefeller O. Deriving more up-to-date estimates of long term patient survival. *J Clin Epidemiol* 1997; 50: 211–216.

8. Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996; 78: 2004–2010.

9. Brenner H, Gefeller O, Hakulinen T. A computer program for period analysis of cancer patient survival. *Eur J Cancer* 2002; 38: 690–695.

10. Brenner H, Hakulinen T. Up-to-date long-term survival curves of patients with cancer by period analysis. *J Clin Oncol* 2002; 20: 826–832.

11. Berkson J. The calculation of survival rates. In: *Carcinoma and other malignant lesions of the stomach*. (eds) W. Wlaters, HK Gray and JT Priestly, 467–484. Philadelphia: Sanders, 1942.

12. Ederer F, Axtell LM and Cutler SJ. The relative survival rate: a statistical methodology. *Monogr Natl Cancer Inst* 1961; 6: 101–121.

13. Berkson J, Gage RP. Calculation of survival rates for cancer. *Proc Mayo Clinic* 1950; 25: 270–286.

14. Ederer F, Heise H. *Instructions to IBM 650 programmers in processing survival computations. Methodological note No. 10*. Bethesda, National Cancer Institute, 1959.

15. Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982; 38: 933–942.

16. Lopez AD, Ahmad OB, Guillot M, Inoue M, Ferguson BD, Salomon JA. *Life tables for 191 countries for 2000: data, methods, results (GPE discussion paper No. 40). Health Systems Performance Assessments Peer Review, Technical Documentation. IV Outcomes: Population Health. Evidence and information for Policy (EIP)*. WHO, October 2001.

17. Parkin DM, Whelan SL, Ferlay J, Teppo L and Thomas DB. Eds. *Cancer Incidence in Five Continents, Vol VIII. IARC Scientific Publications No. 155*. IARCPress, Lyon, 2002.

18. Parkin DM and Hakulinen T. Analysis of survival. In: Jensen OM, Parkin DM, MacLennan R, Muir CS and Skeet RG (eds) *Cancer Registration, Principles and Methods. IARC Scientific Publications No. 95*. IARCPress, Lyon, 1991; pp 159–176.

19. Black RJ and Bashir SA. World standard cancer patient populations: A resource for comparative analysis of survival data. In: *Cancer Survival in Developing Countries*. (eds) Sankaranarayanan R, Black RJ and Parkin DM. IARC Scientific Publications No. 145. IARCPress, Lyon, 1998; pp 9–11.

20. Brenner H, Hakulinen T. On crude and age-adjusted relative survival rates. *J Clin Epidemiol* 2003; 56: 1185–1191.

21. Brenner H, Arndt V, Gefeller O, Hakulinen T. An alternative approach to age-adjustment of cancer survival rates. *Eur J Cancer* 2004; 40: 2317–2322.

22. Ferlay J, Bray F, Pisani P and Parkin DM. *GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide. IARC CancerBase No. 5*, version 2.0. IARCPress, Lyon, 2004.