Chapter 12

Introduction to survival analysis

12.1 Introduction

In this chapter we examine another method used in the analysis of intervention trials, cohort studies and data routinely collected by hospital and population-based cancer registries. Consider the following example:

Example 12.1*. A cohort of 40 women diagnosed with breast cancer in a particular hospital during the years 1989–93 were followed up from diagnosis to the end of 1995 to assess their survival experience. Table 12.1 gives the dates of diagnosis and death (or of last contact) for each of the study subjects.*

In Example 12.1, the patients entered and left the study at different points in time (Table 12.1). We discussed in previous chapters (Chapters 4, 7 and 8) one way of analysing data of this type which takes into account the varying individual lengths of follow-up. That approach involves the calculation of rates based on *person-time at risk*. These calculations are based on the assumption that the rate under consideration remains approximately constant over time, so that 100 person-years of observation are treated identically, whether they involve 100 subjects followed over one year or 50 subjects followed over two years.

In many situations, however, the rate of occurrence of the event under study does not remain constant over time. For instance, the probability of dying may rise suddenly with the onset of disease and then decline gradually as time since diagnosis increases. The most appropriate approach in these situations is to conduct *survival analysis*.

12.2 Estimation of survival

The first requirement for the estimation of survival is a *clear and well defined case definition*. For cancer patients, this should specify the site of the cancer, histology, stage, and the sex of the patients. In Example 12.1, all histologically confirmed female breast cancer cases were included in the analysis.

The second requirement is a *clear and well defined starting point*. The dates of the first diagnosis, the initiation of therapy, or the admission to a hospital are frequently used. Although date of onset of the clinical phase of the disease would seem more appropriate, this is generally difficult to define. In clinical trials, the appropriate starting point is the time of randomization,

Table 12.1.
Follow-up of 40 women diagnosed with breast cancer in a certain hospital during the years 1989–93: hypothetical data.

Patient study number	Age (years)	Stage[a]	Date of diagnosis	Date of last contact or death	Vital status at last contact[b]	Cause of death[c]	No. of complete years of observation from diagnosis to last contact or death	No. of days from diagnosis to last contact or death
1	39	1	01/02/1989	23/10/1992	A	–	3	1360
2	55	1	22/03/1989	12/02/1995	A	–	5	2153
3	56	2	16/04/1989	05/09/1989	D	BC	0	142
4	63	1	23/05/1989	20/12/1992	D	BC	3	1307
5	62	2	12/06/1989	28/12/1995	A	–	6	2390
6	42	2	05/09/1989	17/12/1990	A	–	1	468
7	45	1	05/10/1989	04/08/1995	A	–	5	2129
8	38	2	30/11/1989	11/10/1991	D	BC	1	680
9	53	2	07/01/1990	25/10/1990	D	BC	0	291
10	55	1	03/02/1990	31/01/1991	D	BC	0	362
11	49	2	23/03/1990	29/08/1992	A	–	2	890
12	61	1	28/04/1990	13/05/1994	A	–	4	1476
13	58	1	14/05/1990	01/06/1990	A	–	1	383
14	45	2	15/07/1990	10/09/1993	D	BC	3	1153
15	60	2	03/08/1990	27/11/1994	A	–	4	1577
16	69	1	31/08/1990	06/10/1995	D	O	5	1862
17	58	2	18/09/1990	02/01/1993	D	BC	2	837
18	54	2	09/11/1990	18/06/1995	A	–	4	1682
19	56	2	28/11/1990	27/06/1995	D	BC	4	1702
20	52	1	12/12/1990	13/05/1995	D	O	4	1613
21	67	2	24/01/1991	23/12/1994	D	BC	3	1429
22	64	2	17/02/1991	06/09/1994	D	O	3	1297
23	73	1	21/04/1991	24/12/1993	A	–	2	978
24	48	2	09/06/1991	26/06/1994	A	–	3	1113
25	42	2	20/06/1991	15/03/1992	D	BC	0	269
26	56	2	25/08/1991	19/08/1994	A	–	2	1090
27	43	1	01/03/1992	06/06/1994	D	BC	2	827
28	64	2	12/04/1992	13/02/1995	D	O	2	1037
29	35	2	13/04/1992	15/04/1994	D	BC	2	732
30	77	1	05/05/1992	10/05/1995	A	–	3	1100
31	59	2	10/08/1992	08/11/1992	D	BC	0	90
32	68	1	13/10/1992	21/10/1993	D	BC	1	373
33	70	1	19/11/1992	20/12/1995	A	–	3	1126
34	58	1	17/01/1993	29/10/1994	A	–	1	650
35	75	2	02/02/1993	10/03/1994	D	BC	1	401
36	55	2	02/05/1993	29/09/1993	D	BC	0	150
37	45	1	11/05/1993	07/02/1994	D	BC	0	272
38	69	1	09/11/1993	26/05/1995	A	–	1	563
39	70	1	07/12/1993	27/05/1995	A	–	1	536
40	27	1	31/12/1993	03/06/1995	A	–	1	519

[a] Stage: 1 = absence of regional lymph node involvement and metastases
2 = involvement of regional lymph node and/or presence of regional or distant metastases

[b] A=alive; D=dead

[c] BC=breast cancer; O=causes other than breast cancer

because this is the point when the treatment groups are comparable. In Example 12.1, the date of diagnosis was taken as the starting point.

The third requirement is a *clear and well defined outcome*. Often the outcome of interest is death, but it need not be so. It can be recurrence of

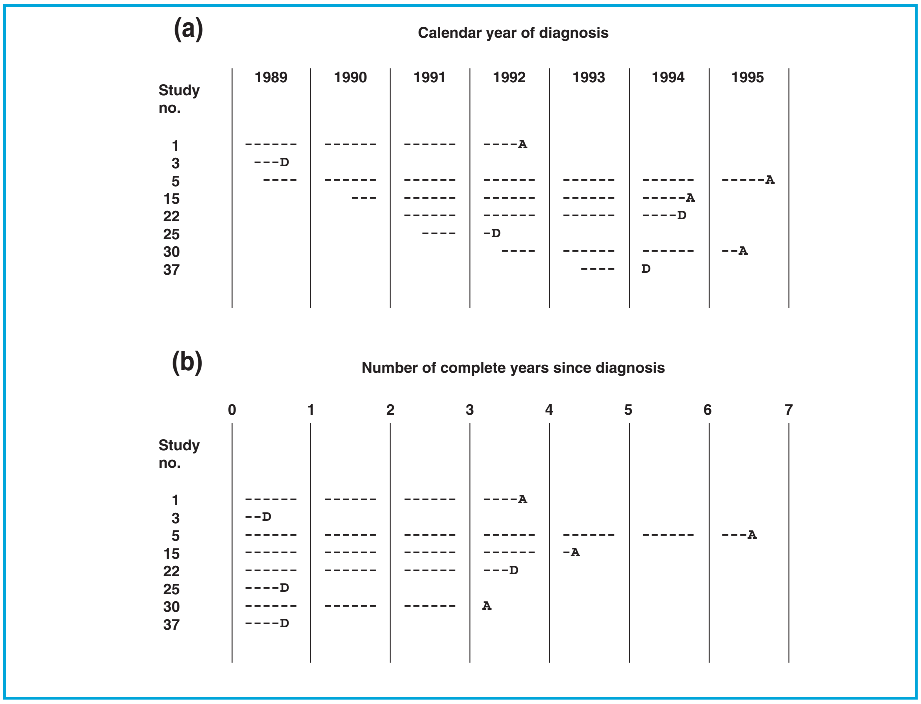

(a)
Calendar year of diagnosis
Study no.
1989
1990
1991
1992
1993
1994
1995
1 ------ ------ ------ ----A
3 ---D
5 ---- ------ ------ ------ ------ ------ -----A
15 --- ------ ------ ------ -----A
22 ------ ------ ------ ----D
25 ---- -D
30 ---- ------ ------ --A
37 ---- D
(b)
Number of complete years since diagnosis
Study no.
0
1
2
3
4
5
6
7
1 ------ ------ ------ ----A
3 --D
5 ------ ------ ------ ------ ------ ------ ---A
15 ------ ------ ------ ------ -A
22 ------ ------ ------ ---D
25 ----D
30 ------ ------ ------ A
37 ----D

Figure 12.1.
Diagram illustrating how follow-up data from 8 of the 40 women with breast cancer (see Table 12.1) can be presented (*a*) by calendar year of diagnosis and (*b*) by time since entry into the study (A=alive; D=dead).

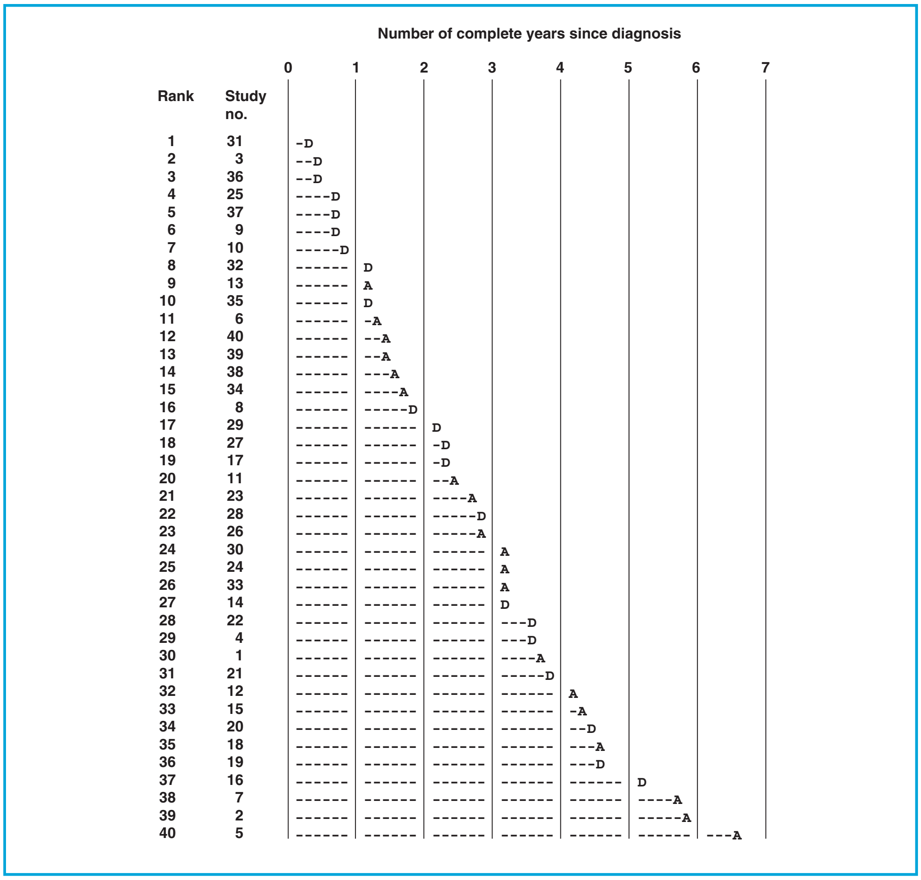

Number of complete years since diagnosis
0
1
2
3
4
5
6
7
Rank
Study no.
1 31 -D
2 3 --D
3 36 --D
4 25 ----D
5 37 ----D
6 9 ----D
7 10 -----D
8 32 ------ D
9 13 ------ A
10 35 ------ D
11 6 ------ -A
12 40 ------ --A
13 39 ------ --A
14 38 ------ ---A
15 34 ------ ----A
16 8 ------ -----D
17 29 ------ ------ D
18 27 ------ ------ -D
19 17 ------ ------ -D
20 11 ------ ------ --A
21 23 ------ ------ ----A
22 28 ------ ------ -----D
23 26 ------ ------ -----A
24 30 ------ ------ ------ A
25 24 ------ ------ ------ A
26 33 ------ ------ ------ A
27 14 ------ ------ ------ D
28 22 ------ ------ ------ ---D
29 4 ------ ------ ------ ---D
30 1 ------ ------ ------ ----A
31 21 ------ ------ ------ -----D
32 12 ------ ------ ------ ------ A
33 15 ------ ------ ------ ------ -A
34 20 ------ ------ ------ ------ --D
35 18 ------ ------ ------ ------ ---A
36 19 ------ ------ ------ ------ ---D
37 16 ------ ------ ------ ------ ------ D
38 7 ------ ------ ------ ------ ------ ----A
39 2 ------ ------ ------ ------ ------ -----A
40 5 ------ ------ ------ ------ ------ ------ ---A

Figure 12.2.
The data of Table 12.1 ordered by length of observed survival time, with (D) representing dead and (A) alive at the end of the follow-up period.

tumour, first occurrence of a particular complication, etc. The only requirement is that the endpoint is a binary variable (e.g., being alive versus being dead) and that each subject can have one and only one endpoint. In our example (12.1), death was considered the outcome of interest.

The time between the starting point and the occurrence of the outcome of interest (or the date of the last contact) is known as *survival time*. The calculation of survival time for some of the patients in Table 12.1 is illustrated in Figure 12.1(*b*). Note that subjects may have different dates of diagnosis but still have the same survival time. For instance, patients No. 25 and 37 had similar survival time, despite differing dates of entry (20/06/1991 and 11/05/1993, respectively; Figure 12.1(*a*); Table 12.1). Figure 12.2 shows the individual survival times for the 40 breast cancer women of Example 12.1 ranked by increasing duration.

The interpretation of the results of a survival study depends greatly upon the length of time each person was followed up. A typical survival study involves a patient accrual period during which patients are recruited and their follow-up is initiated, a follow-up period during which patients are followed up but no further recruitments are made, and a closing date for the analysis. In Example 12.1, the recruitment period was from the start of 1989 until the end of 1993, the follow-up period continued from the beginning of 1994 to the end of 1995, and the closing date for the present analysis was the end of 1995.

One way of summarizing survival data is to report the proportion of patients still alive at a fixed point in time. In Example 12.1, we might initially restrict our analysis to patients for whom we have complete information on the first two years of follow-up. Figure 12.2 shows that six women (Nos 13, 6, 40, 39, 38 and 34) were lost to follow-up before completing a two-year period and should therefore be excluded from the analysis.

In summary, 34 patients completed a two-year follow-up, of whom 10 died and 24 were still alive (Figure 12.2). These results can be presented in a tree diagram (Figure 12.3), in which the upper branch of the tree corresponds to deaths and the lower branch to survivors.

On the basis of these results, we might estimate the *probability (or risk) of dying* in the first two years as 10/34 = 0.29 = 29%.

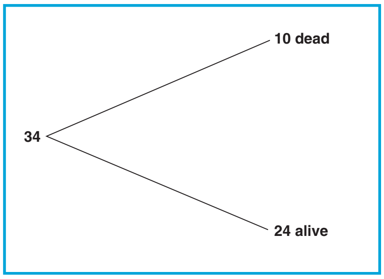


Figure 12.3.
Tree diagram illustrating the two possible outcomes for the 34 patients who completed a two-year follow-up period.

12.3 Censored observations

A closed group consists of a group of individuals in which there are only complete observations. In practice, it is rare to find a closed group, because there are almost always some subjects whose follow-up is incomplete. This occurs because they join the cohort too late to complete the appropriate follow-up before the study ends or because they are lost to follow-up (because of, for example, change of address or migration). Early termination of follow-up for any such reason is called *censoring*.

Our previous calculation of the probability of dying in the first two years excluded censored observations. However, the fact that censored subjects did not die during the time they were in the study provides some information about the probability of dying. Suppose we do not know the exact dates when censoring occurred and all we know is the number of patients who were unable to complete the defined follow-up period. If the time-interval is relatively short, we can make a simple estimate by assuming that on average we observed each censored patient for half the follow-up period without observing any deaths among them. Thus, for a cohort of size N with D observed deaths and L losses due to censoring, we estimate the probability of dying in the interval as

$$D/(N - 0.5L)$$

Thus censoring reduces the effective size of the cohort by half the size of the group lost to follow-up ($0.5L$). This rather crude way of taking account of censoring works adequately provided L is small compared with N.

We can now re-calculate the probability of dying in the first two years in Example 12.1. Thus, of the 40 breast cancer patients recruited into the study

10 died during the two-year follow-up period ($D = 10$)
24 were still alive at the end of the follow-up ($A = 24$)
6 survived but were lost to follow-up ($L = 6$)

These results can be presented in a tree diagram similar to the one shown in Figure 12.3, except that there is now an additional middle branch corresponding to the censored observations (Figure 12.4).

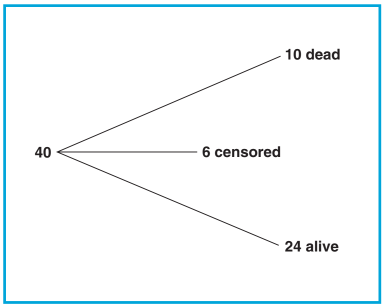


Figure 12.4.
Tree diagram illustrating the outcome of the 40 breast cancer patients from Example 12.1 at the end of a two-year follow-up period.

We have now included all 40 patients in our calculations. However, the effective size of the cohort is no longer 40 but 37 due to the six censored observations (= 40 – 0.5 × 6). The probability of dying is estimated as 10 / 37 = 0.27 = 27%.

Similarly, we can calculate the probability of dying during the first three years of diagnosis. Since the last attempt to contact patients was made in 1995, patients diagnosed after 31 December 1992 entered the study too late to have been able to complete a three-year follow-up. Thus, the observations for four patients (Nos 34, 38–40) were censored (Table 12.1). Five other women (Nos 13, 6, 11, 23, 26) did not complete the three-year observation period because they were lost to follow-up (Figure 12.2). Thus, of the 40 breast cancer patients recruited into the study:

14 died during the three-year follow-up period
17 were still alive at the end of the follow-up
9 were lost to follow-up or joined the cohort too late to complete three years of observation.

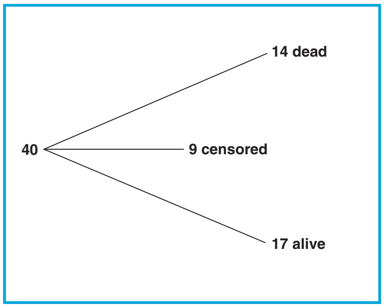


Figure 12.5.
Tree diagram illustrating the outcome of the 40 breast cancer patients from Example 12.1 at the end of a three-year follow-up period.

The corresponding tree diagram is shown in Figure 12.5. The probability of dying in the first three years after a diagnosis of breast cancer can be estimated as 14 / (40 – 0.5 × 9) = 39%.

12.4 Consecutive follow-up intervals

The use of a single interval of follow-up has several limitations. Firstly, it is a very crude way of summarizing the survival experience of a cohort, since it ignores any information about when the deaths and censoring took place. Only the total number of deaths and the total number of censored observations that occurred during the defined interval is required for the calculations. Secondly, it is possible to compare the survival experience of different cohorts only if the same follow-up interval is used. For instance, it is not possible to compare the survival experiences of two cohorts of breast cancer patients if the experience of one cohort is summarized as the probability of dying in the first two years after diagnosis and that of the second as the probability of dying in the first five years.

One way of overcoming these limitations is to use a number of shorter consecutive intervals of time, rather than just one long interval. The experience of the cohort during each of these intervals can then be used to build up the experience over the entire period. Instead of a single calculation of the probability of dying, there will be a sequence of calculations, one for each interval.

Consider again the three-year follow-up shown in Figure 12.5. This period can be divided into three one-year intervals. We can use the data shown in Figure 12.2 to present the number of patients who contribute to each of the three possible outcomes (i.e., death, censoring and survival) in each of the three consecutive years of follow-up. The resulting tree diagram is shown in Figure 12.6.

In this tree diagram, the survivors from one year go on to the start of the next year. In the first year, there were 40 breast cancer patients of whom seven died and none were censored, leaving 33 patients at the start of the second year. Of these 33 patients, three died and six were censored during the second year, leaving 24 at the beginning of the third year. During the third year, four women died, three were censored and 17 were known to be alive.

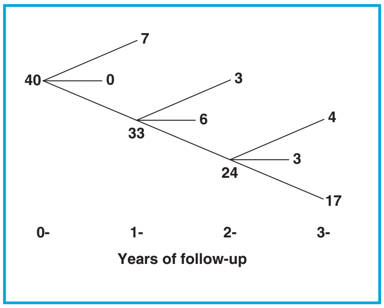


Figure 12.6.
Tree diagram showing the number of breast cancer patients from Example 12.1 who contributed to the different outcomes in each of the first three years of follow-up.

12.5 Estimation of probabilities

We can now replace the numbers of patients on our tree by the probabilities of dying and surviving in each of the intervals. The probability of dying in each interval can be calculated as before, taking account of the censored observations in that interval. The *probability of survival* in the interval is just one minus the probability of dying in the interval.

In the *first year* there were no censored observations, thus

the probability of dying in the year is 7/40 = 0.175;
the probability of surviving the year is 1 – 0.175 = 0.825.

In the *second year*, six women were censored. The effective size of the cohort in this year can then be estimated as 33 – (0.5 × 6) = 30. Thus

> the probability of a subject dying during the second year, given that the subject was alive at the start of the year, is estimated to be 3/30 = 0.10;
> the probability of surviving the year is estimated to be 1 – 0.10 = 0.90.

In the *third year*, three women were censored. The effective size of the cohort is 24–(0.5 × 3) = 22.5. Thus

> the probability of a subject dying during the third year, given that the subject was alive at the start of the year, is estimated to be 4/22.5 = 0.178;
> the probability of surviving the year is estimated to be 1 – 0.178 = 0.822.

The full tree with the branch (conditional) probabilities of dying in each year given that the subject survived the previous years is shown in Figure 12.7.

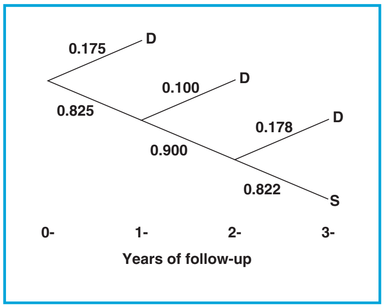


Figure 12.7.
Tree diagram showing the probabilities of each possible outcome in each of the first three years of follow-up (D = death; S = survival).

There are now four possible outcomes of interest, corresponding to the tips of the tree. The probability of each outcome can be calculated by multiplying down the branches of the tree. Therefore the probabilities for each outcome are:

1. Probability of dying during the first year = 0.175
2. Probability of dying during the second year (i.e., probability of surviving in year 1 × probability of dying in year 2) = 0.825 × 0.10 = 0.083
3. Probability of dying during the third year = 0.825 × 0.90 × 0.178 = 0.132
4. Probability of being alive by the end of the three years = 0.825 × 0.90 × 0.822 = 0.610

These probabilities will always add up to 1, since there are no other possible outcomes. The probability of dying at *some point* during the three-year interval is equal to 0.175 + 0.083 + 0.132 = 0.390. This probability can be found more conveniently by subtracting the probability of surviving the whole three-year period from 1, giving 1 – 0.610 = 0.390.

The final probability of surviving (0.610) is an example of a *cumulative survival probability* for the cohort, i.e., the probability of surviving three consecutive years.

12.6 Actuarial life-table

The data from the previous calculations are often presented in the form of an actuarial life table, which shows the numbers of deaths and censorings occurring in each consecutive interval. A life table for the 40 breast cancer patients from Example 12.1 is shown in Table 12.2.

In this table, the probability of dying during each year is calculated as $D/(N - 0.5L)$. Thus, the probability of surviving the year is equal to $1 - D/(N - 0.5L)$. The cumulative survival is found by multiplying the survival probabilities for each of the consecutive years to obtain the cumulative

Table 12.2.
Actuarial life table for the 40 breast cancer patients of Example 12.1.

Year	No. at start of interval (*N*)	No. of deaths (*D*)	No. of losses (*L*)	Effective denominator (*N*–0.5*L*)	Probability of dying during the year	Probability of surviving the year	Cumulative survival
0–	40	7	0	40	0.175	0.825	0.825
1–	33	3	6	30.0	0.100	0.900	0.743
2–	24	4	3	22.5	0.178	0.822	0.610
3–	17	4	4	15.0	0.267	0.733	0.447
4–	9	2	3	7.5	0.267	0.733	0.328
5–	4	1	2	3.0	0.333	0.667	0.219
6–	1	0	1	0.5	0.0	1.00	0.219
Total		**21**	**19**				

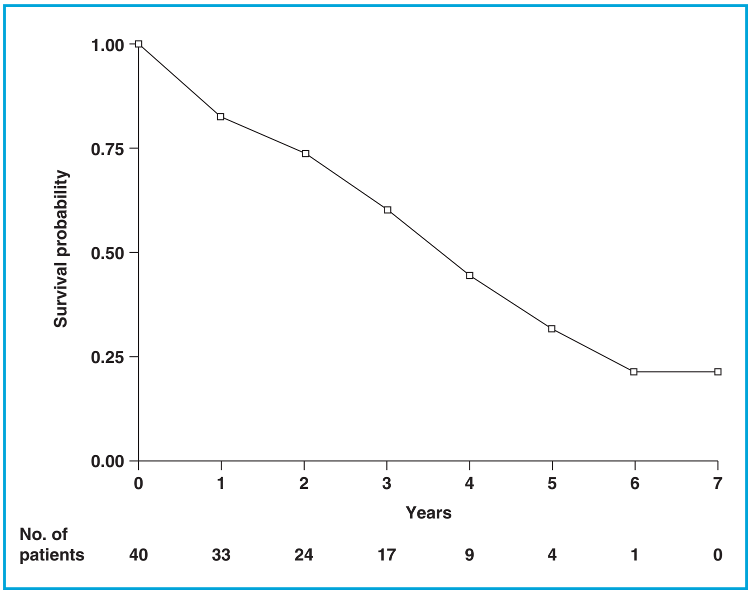


Figure 12.8.
Life-table (actuarial) survival curve for the 40 breast cancer patients of Example 12.1.

probabilities of surviving 1, 2, ..., 6 years. For example, the probability of surviving three years without dying is 0.825 × 0.90 × 0.822 = 0.610 (the same value we calculated before). The life table is therefore just a convenient way of displaying these probabilities which are derived in the same way as those on the tree diagram. Life-tables are useful to examine whether the probability of dying changes with follow-up time, and for presenting concisely summary measures for different intervals of follow-up.

The cumulative survival probabilities can also be displayed graphically as in Figure 12.8. This plot is called a *survival curve*. The curve starts at 1 (all patients alive) and with time progressively declines towards 0 (all patients have died).

12.7 Kaplan–Meier method

The actuarial life-table method described in Section 12.6 does not require information on the exact time when deaths or censoring occur. Only knowledge of the subjects' vital status at each of the limits of the intervals is required. If the *exact times* when deaths occur are known, survival probabilities can be estimated immediately after each individual death without any need to aggregate the data into intervals of one year (or of any other length). This method of estimating the cumulative survival probabilities is called the *Kaplan–Meier method* and it is the preferred approach whenever event and censoring times are available (see Estève *et al.* (1994) for a full description of the calculations).

Similarly to the life-table survival curve, the Kaplan–Meier estimates can be used to plot cumulative survival probabilities. In this instance, however, the plot is in the form of a *stepped line*, rather than a smooth curve, since the cumulative survival drops at the precise time that a death occurs and remains at a plateau between successive death times. For instance, the curve for the 40 breast cancer patients of Example 12.1 shown in Figure 12.9 starts at 1 and continues horizontally until the first death (patient number 31) at day 90; at this time it drops by a function of the estimated probability of dying. It then continues horizontally until the next death (patient 3) at day 142, and so on. The graph will reach zero only if the patient with the longest observed survival time has died. If this patient is still alive at the end of the follow-up period, the Kaplan–Meier curve has a plateau commencing at the time of the last death and continuing until the censored survival time of this longest surviving patient. In Figure 12.9, the survival time of each censored observation is marked in the curve by a cross. After the last death (patient 16, at day 1862 (5.1 years)), the curve remains flat until the longest censored survival time (patient 5, at day 2390 (6.5 years)).

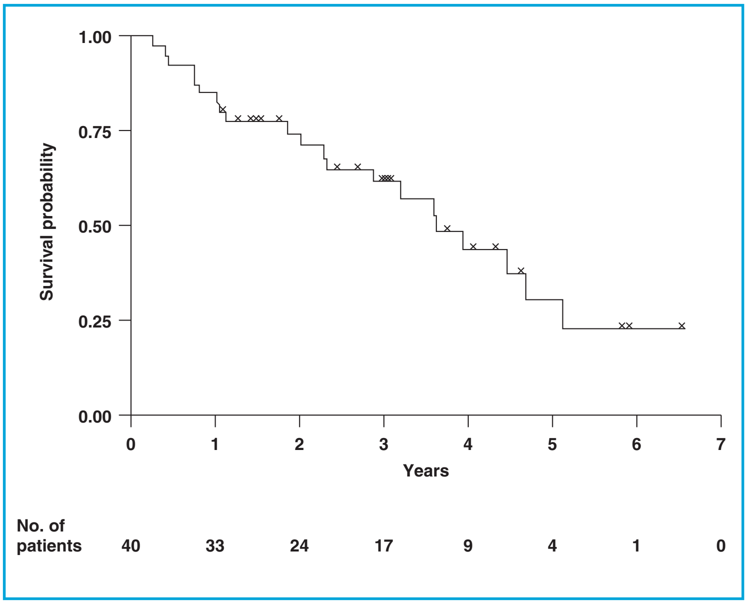


Figure 12.9. Survival curve produced by the Kaplan–Meier method for the 40 breast cancer patients of Example 12.1 (x indicates censoring times).

It is useful to give the number of patients at risk at selected time points (for example, at the start of each year) under the graph and/or to present confidence intervals around the survival probability estimates. This information is crucial for a sensible interpretation of any survival curve.

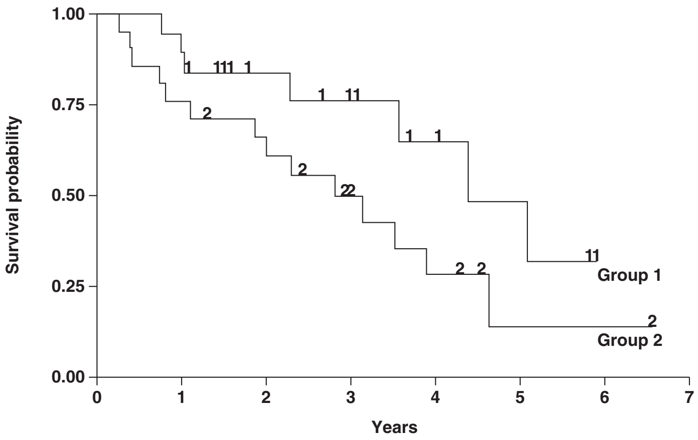
12.8 Comparison of survival curves

In many situations, the primary objective of the study is to compare the survival experience of different groups of patients. These groups may be defined according to sex, stage of the tumour at the time of diagnosis (as in Example 12.2), histological type, etc. In clinical trials, the groups will be defined on the basis of the treatment given. Cumulative survival probabilities are calculated separately for each group and the two curves plotted on the same graph for comparison (Figure 12.10).

A visual comparison of survival curves is extremely useful. Consider the graphs presented in Figure 12.11. In graph (*a*), the two curves overlap in the first two years of follow-up but diverge thereafter. In graph (*b*), group A initially has better survival than group B, but the curves cross after four years of follow-up and ultimately group A does worse than group B.

These patterns would be missed if the comparison was restricted to a specific follow-up period. For instance, if only two-year survival probabilities were calculated, they would indicate that there was no clear difference between the treatments in graph (*a*) and that treatment A was much superior to treatment B in graph (*b*). These two examples clearly illustrate that comparison of survival experiences should always be based on survival curves. Statistical tests for the formal comparison of two survival curves, such as the *logrank test*, can then be used to assess the statistical significance of any observed differences (see Estève *et al.*, 1994).

***Example 12.2**. In Example 12.1, the investigators also collected data on stage of the tumour at the time of the diagnosis (Table 12.1). Separate Kaplan–Meier curves were prepared for each stage (Figure 12.10).*



No. of patients								
Group 1:	19	17	11	9	5	3	0	0
Group 2:	21	16	13	8	4	1	1	0

Figure 12.10.
Kaplan–Meier survival curves for patients with breast cancer by stage of the tumour at the time of diagnosis (group 1 = tumour without lymph node involvement or metastasis; group 2 = tumour with lymph node involvement and/or regional or distant metastasis). The numbers on the survival curves represent censored observations.

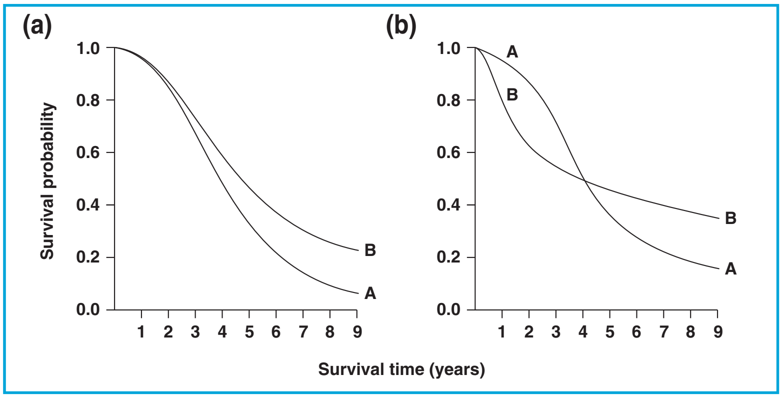


Figure 12.11.
Two examples of comparative survival curves (reproduced by permission of the BMJ Publishing Group, from Gore, 1981).

When comparing survival curves in relation to a particular prognostic (or therapeutic) factor, it is important to ensure that the groups are similar in relation to other prognostic factors. In Example 12.2, for instance, other characteristics such as age should have been taken into account. In randomized trials this is accomplished by the random allocation of the subjects to the various arms of the trial (provided the sample size is large). In observational studies, it is possible to obtain Kaplan–Meier curves adjusted for confounders such as age, sex, stage of the tumour, etc. (see Estève *et al.*, 1994) provided data on these variables are collected.

12.9 Overall survival and cause-specific survival

The first step in the analysis of the survival experience of a group of patients should be to examine their *overall survival*. In our breast cancer example, no distinction was made between deaths from breast cancer and deaths from other causes. However, a subject who dies in a traffic accident is no longer at risk of dying from breast cancer. One way of adjust-

Year	No. at start of interval (*N*)	No. of deaths (*D*)	No. of losses (*L*)	Effective denominator (*N*–0.5*L*)	Probability of dying during the year	Probability of surviving the year	Cumulative survival
0–	40	7	0	40	0.175	0.825	0.825
1–	33	3	6	30.0	0.100	0.900	0.743
2–	24	3	4	22.0	0.136	0.864	0.641
3–	17	3	5	14.5	0.207	0.793	0.509
4–	9	1	4	7.0	0.143	0.857	0.436
5–	4	0	3	2.5	0.0	1.0	0.436
6–	1	0	1	0.5	0.0	1.0	0.436
Total		**17**	**23**				

Table 12.3.
Life-table probabilities of dying from breast cancer for the 40 breast cancer patients of Example 12.1. In this table, deaths from causes other than breast cancer were considered as censored observations.

ing for these ‘competing’ causes of death is to treat patients who died from other causes as if they had been withdrawn alive (i.e., censored at the time of their death) and then carry out the life-table calculations as described above. The survival probabilities obtained by this method are *cause-specific survival probabilities*, since they take into account deaths due to causes other than the disease under study.

In Example 12.1, four patients died from causes other than breast cancer (see Table 12.1). A new actuarial life-table can then be constructed by considering these deaths as censored observations (Table 12.3). The total number of deaths is decreased by 4 (17 instead of 21) and the number of losses increased by 4 (23 instead of 19). Similarly, when the exact dates at which deaths occur are known, it is possible to use the Kaplan–Meier method to estimate these cause-specific survival probabilities (Figure 12.12).

The calculation of cause-specific survival probabilities requires information on cause-specific mortality. This information may not be easy to obtain. Deaths from other causes tend to be under-reported in cancer patients, as many of them will be entered in the death certificate simply as deaths from cancer. Even when other causes apart from cancer are reported, it is difficult to establish whether the cause of death was unrelated to the cancer of interest (e.g., cancer in adjacent organs).

If accurate cause-specific mortality data are not available, this method cannot be used. It may be possible, however, to compare the *observed survival* with what would have been *expected* for a group of people in the general population similar to the patient group with respect to race, sex, age and calendar period of observation. This expect-

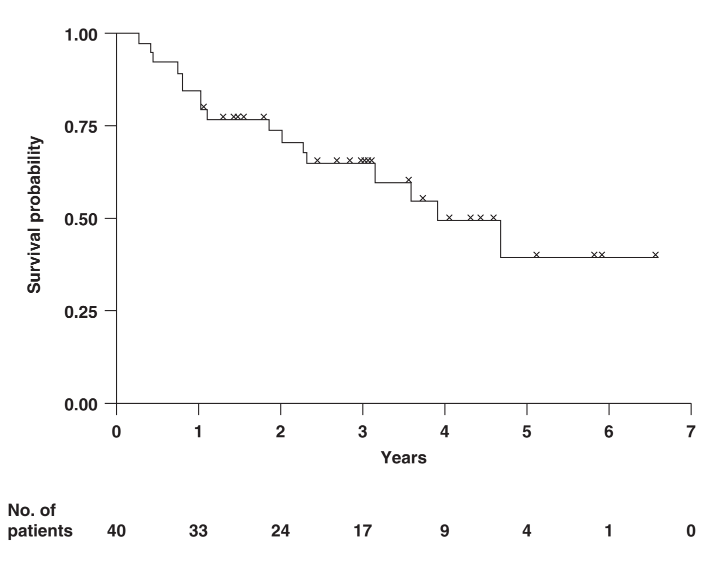


Figure 12.12.
Kaplan–Meier breast cancer-specific survival curve for the 40 breast cancer patients of Table 12.1.

ed survival can be derived from published demographic life tables (see below). The comparison yields *relative survival ratios* which are adjusted for the patients' probability of dying from a cause other than the one under study (see Parkin & Hakulinen (1991) for an illustration of these calculations). Thus, the relative survival ratios represent the survival experience of a group of patients adjusted for their probability of dying from causes other than the one under investigation. In practice, the 'all causes' and 'all causes minus cancer' demographic life tables are very similar and since the former are more readily available, these are generally used in the calculations.

12.10 Demographic life tables

All the above discussion of life tables relates to data derived from real cohorts, i.e., from groups of people who were actually followed up in time.

Demographic life tables, computed on the basis of national (or regional or specific for a particular ethnic or socioeconomic group) mortality data, can be obtained by applying the currently observed mortality risks at various ages to an imaginary cohort. Thus the life expectancy of women at birth in England and Wales, which was 77 years in 1981 (Bone *et al*., 1995), depends on the assumption that baby girls born in 1981 will be exposed to 1981 age-specific risks of dying as they go through life (e.g., when they are age 30 in the year 2011, they will experience the 1981 mortality risks for 30-year-olds). Although taken literally, this assumption is unrealistic, demographic life tables are a good way of summarizing current mortality risks. These demographic life tables are usually prepared and published by governmental statistical offices.

12.11 Other outcomes

The methods described in this chapter are part of a group of statistical techniques used in 'survival analysis'. The term 'survival' comes from the fact that the first use of such techniques arose in the insurance industry, which was particularly interested in developing methods of costing insurance premiums. For this purpose, they needed to know the average life expectancy for different types of customer.

The use of survival analysis techniques is, however, by no means restricted to studies where death is the outcome of interest. It has also been widely used to study other outcomes such as fertility, compliance with treatment, recurrence of disease, occurrence of complications, etc.

The trial described in Example 12.3 had more than one outcome of interest. The results in Figure 12.13 show little evidence of a difference in overall survival ($P = 0.5$) or survival free from regional or distant recurrence ($P = 0.19$). However, the trial provided moderate evidence in favour of the hypothesis that women who received radiation had a lower risk of developing local recurrences ($P = 0.06$).

***Example 12.3**. A total of 381 women with invasive breast cancer in histopathological stage I had sector resection with meticulous control for complete excision of local disease plus axillary dissection. After this surgery, 187 were randomized to receive postoperative radiotherapy to the breast and 194 women to receive no further treatment. The outcomes of interest were overall survival and time from treatment to local recurrence and to regional or distant metastasis. The Kaplan–Meier method was used in the analysis (Uppsala-Örebro Breast Cancer Study Group, 1990). The main results from this trial are shown in Figure 12.13.*

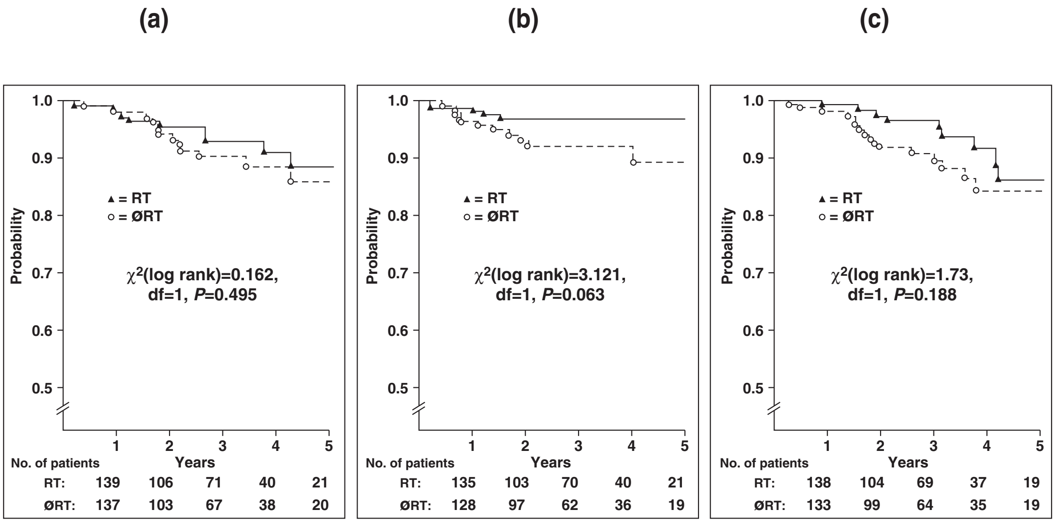


Figure 12.13.
Probability of (*a*) overall survival; (*b*) of remaining free from local recurrence; and (*c*) of remaining free from regional or distant metastasis for 381 breast cancer patients according to type of postoperative treatment (RT = postoperative radiotherapy to the breast; ØRT = no further treatment) (reproduced, by permission of Oxford University Press, from Uppsala-Örebro Breast Cancer Study Group, 1990).

12.12 Final notes

Survival analysis can be carried out easily with many statistical computer packages such as STATA, EGRET, SAS or SPSS. Unfortunately, it is not possible to conduct this type of analysis in EPI INFO.

The application of survival analysis to data collected by cancer registries is discussed in Section 17.6.2.

Further reading

* The use of probability trees in this chapter was based on the approach suggested by Clayton & Hills (1993).

* A more elaborate presentation of the general statistical concepts underlying survival analysis and their application to routinely collected data is given in Estève *et al.* (1994).

* A guide to the use of survival curves in cancer trials is given by Peto *et al*. (1976, 1977).

Box 12.1. Key issues

- Survival analysis is another method used in the analysis of data from intervention trials, cohort studies and data routinely collected by cancer registries. It is particularly useful when the probability of occurrence of the event under study changes with time since entry into the study.

- The survival experience of a group of people may be summarized by reporting the proportion still alive at a particular point in time (e.g., at the end of a two-year follow-up). This approach has several limitations, however. First, no account is taken of the time when deaths and censoring took place. Second, it is possible to compare the survival experience between groups of people only if the same follow-up period is used. Third, it does not provide any indication of changes in survival with follow-up time.

- All the above limitations can be overcome by calculating *cumulative survival probabilities* for consecutive follow-up intervals. These probabilities can then be displayed graphically in a plot called a *survival curve.*

- Cumulative survival probabilities can be calculated by using either the *actuarial life-table* method or the *Kaplan–Meier* method. The two methods are basically similar, but the shape of the resulting survival curve is slightly different. The actuarial life-table method produces a smooth curve because cumulative survival probabilities are calculated only at the end of each of the consecutive follow-up intervals, whereas the Kaplan–Meier method produces a stepped line because these probabilities are calculated immediately after each death takes place.

- The first step in survival analysis should be to estimate the *overall survival* experience of the entire cohort. Sometimes it may be useful to proceed to estimate *cause-specific survival*. This can be easily done if accurate cause-specific mortality data are available for the study subjects. If these data are not available, it is still possible to look at cause-specific survival by using information from demographic life-tables.