European Journal of Surgical Oncology 1999; **25**: 321–329

Survival analysis: caveats and pitfalls

A. Mathew*, M. Pandey† and N. S. Murthy‡

*Divisions of *Epidemiology and Clinical Research and †Surgical Oncology, Regional Cancer Centre, Trivandrum, Kerala 695 011, and ‡Division of Biostatistics, Institute of Cytology and Preventive Oncology (ICMR), Bahadur Shah Zafar Marg, New Delhi 110 002, India*

Background: Survival analysis in clinical studies is important to assess the effectiveness of a given treatment and to understand the effect of various disease characteristics. A number of methods exist to estimate the survival rate and its standard error. However, one cannot be certain that these methods have been handled appropriately. The widespread use of computers has made it possible to carry out survival analysis without expert guidance, but using inappropriate methods can give rise to erroneous conclusions. The majority of the biomedical journals now recommend that a statistical review of each manuscript should be carried out by an experienced bio-statistician, in addition to obtaining expert referees' comments on the article. The problem is compounded in papers from third-world countries where bio-statisticians may not be available in all institutions to guide clinicians as to the selection of proper techniques.

Methods: The present paper deals with the various techniques of survival analysis and their interpretation, using a modal data set of malignant upper-aerodigestive tract melanoma patients treated in the Regional Cancer Centre, Trivandrum since 1982.

Results: The Kaplan–Meier method was found to be the most suitable for survival analysis. The median survival time is a better method of summarizing data than the mean. Rothman's method of estimation of the confidence limit is better than Peto's method as the confidence limit for survival probability tends to go beyond the range of 0–1.0 when calculated by Peto's method, especially when the sample size is small.

Conclusion: The results from the present study suggest that survival analysis should be carried out by the Kaplan–Meier method. The median survival time should be provided wherever possible, rather than relying on mean survival. Confidence limits should be calculated as a measure of variability. A suitable rank test should be used to compare two or more survival curves, rather than a Z-test. Stratified analysis and Cox's model, when stratified analysis fails, can be used to define the impact of prognostic factors on survival.

Key words: **actuarial method; Kaplan–Meier method; median survival time; Cox's model.**

Introduction

Survival is one of the major end-points in clinical audits. Survival analysis describes the outcome, with time, of a given disease in a group of patients. An estimation of the survival associated with a disease is essential for setting public health priorities. It provides information on the probability of survival following the onset of disease for a newly diagnosed patient and also on the prognosis according to patient characteristics (sex, age at diagnosis, etc.), disease characteristics (site, histology, etc.) and methods of treatment.

There are a number of methods to estimate survival rates, such as the direct method,[1] the actuarial (life-table) method,[1–3] and the Kaplan–Meier (product-limit) method.[4] Similarly, there are several ways to estimate variability (standard error) for survival rates.[5–7] The difference between two or more survival curves can be tested by using rank tests.[6,8] Stratified rank tests can assess the survival experience of different levels of one prognostic factor after adjusting for other factors.[9] However, when the number of prognostic factors increases the strata soon become too small. In such situations the effectiveness of stratified analysis becomes limited. To overcome this, Cox's regression model[10] can be used with a number of prognostic factors.

Here, we describe the various methods currently used for estimating survival rates and standard error. These methods are illustrated using a data set of melanoma patients treated at the Regional Cancer Centre, Trivandrum, Kerala, India.[11] The results obtained are compared critically and the limitations of stratified analysis and the usefulness of Cox's regression model are highlighted.

Fundamentals of survival rates

In survival studies, a group of individuals with some common morbidity experience is followed-up from a well-defined starting point. In the estimation of survival rates

Correspondence to: Dr Aleyamma Mathew, Assistant Professor in Statistics and Epidemiology, Regional Cancer Centre, Trivandrum, 695 011 Kerala, India. Fax: 91-471-550 782; E-mail: rcctvm@md2.vsnl.net.in

0748–7983/99/030321 + 09 $12.00/0

the fundamentals such as the start date and common closing date of the study and the status of all study patients at the closing date should be defined before beginning the study.[3]

Start date

This is defined as the date at which the observation period of a group of individuals in the study commences. This varies according to the purpose of the study. Commonly used start dates include the date of first symptoms, of diagnosis, of the first visit to the physician or the clinic, of hospital registration or admission, or the date of beginning treatment.

Common closing date

This refers to the date of termination of the study. This should be set prior to the initiation of the study. Each individual must be followed-up until the closing date.

Status at common closing date

The status of each individual who participated in the study must be assessed at the closing date. In survival studies, when the period of observation ends, some patients will still be alive with or without disease (disease-free). Some patients will have died either due to their disease under study or to other causes; and, finally, some patients would have been lost from the study.

Any patient who dies due to the disease under study is termed a 'failure' and only these such observations have a complete follow-up. All other patients have incomplete follow-up and are termed 'censored' observations.

Length of follow-up

The length of follow-up is an important component in survival analysis, defined as the time between the date of the beginning of the observation period and the date at which the observation ends.[3] It is expressed in arbitrary units of days, months or years. The unit is selected based on the natural history of the disease under consideration. In diseases with a long natural history, the unit could be years. If the disease being studied has a very short natural course, the unit could be expressed in terms of months or days.[12]

Estimation methods

Direct method

This is a simple method of summarizing patient survival. The survival probability at a given time is estimated by the ratio between the number of survivors at the end of a specified interval and the number of observations at the beginning of the study.[1] This approach is termed the 'direct' method by some authors.[13] Not all the censored observations are included in the survival rate calculation, but only those who have completed the specified follow-up time.

Actuarial (life-table) method

In 1958 Cutler and Ederer devised a method of calculating survival by incorporating two types of censored observations: (a) those patients who are lost to follow-up (LFU) during the observation period, and (b) those who are known to be alive but have withdrawn from observations due to the closure of the study (withdrawn alive). This method is frequently referred to as the 'actuarial' method because the techniques used are similar to those employed by actuaries. The method relies on the assumption that censored observations are subject to the same force of mortality as patients with complete follow-up.[2]

Table 1 shows the computation of survival rates by the actuarial method. The first step is the division of the maximum follow-up time into intervals (i); the length of the intervals is set *a priori* based on the distribution of deaths over a period of time. If necessary, the intervals can be of unequal sizes. The second step is the calculation of number of observations exposed to the risk of dying (N_i) which is obtained by subtracting half the number of subjects who were withdrawn alive (w_i) or LFU (l_i) during the interval from the number at the beginning (n_i) of the interval [$N_i = n_i - (w_i + l_i)/2$]. The figure thus obtained is termed the 'effective number at risk'. The next step is the computation of the conditional probabilities of death (q_i) in each interval, which are calculated by the ratio of the effective number at risk (N_i) to the number who died (d_i) during each interval

Table 1. Calculation of survival rates by the actuarial method

(1) Interval	(2) No. at beginning of interval	(3) No. died during interval	(4) LFU* during interval	(5) Withdrawn alive during interval	(6) Effective no. exposed to risk of dying $\left[n_i - \frac{(w_i + l_i)}{N_i^2}\right]$	(7) Conditional probability of death d_i/N_i	(8) Conditional probability of survival $1 - d_i/N_i$	(9) Survival rate $P_i = p_1 \times p_2 \times \ldots\ldots \times p_i$
i	n_i	d_i	l_i	w_i		q_i	p_i	P_i
1	n_1	d_1	l_1	w_1	N_1	q_1	p_1	P_1
2	n_2	d_2	l_2	w_2	N_2	q_2	p_2	P_2
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

* Lost to follow-up.

($q_i = d_i/N_i$). The conditional probability of survival (p_i) during each interval is then calculated by subtracting the conditional probability of death from 1 ($p_i = 1 - q_i$).

The survival rate at the end of a given interval (i) is obtained by multiplying the conditional probabilities of survival over all the intervals up to this time point ($P_i = p_1 \times p_2 \times ..... \times p_i$), and hence this is called the 'cumulative survival rate'. Cumulative survival rates calculated at the end of each interval are used to plot a survival curve, providing a pictoral description of the survival pattern.

Kaplan–Meier (K–M) or product–limit (P–L) method

In 1958, Kaplan and Meier[4] proposed a variation on the actuarial method in which the necessity of grouping the data in pre-fixed time intervals was removed. The difference is that the time intervals are determined by the occurrences of the outcome of interest (event), as and when they occur. The follow-up times of all observations are arranged in increasing order of magnitude. The procedure relies on the same principles as the actuarial method, but the conditional probabilities of survival are estimated every time an event occurs. These are estimated from the number of at-risk observations at the time of each event. The number of at-risk observations is obtained by subtracting all censored observations prior to the occurrence of the specified event. As in the actuarial method, the cumulative survival rate at any given time point is obtained by the product of all conditional probabilities of survival calculated at each event up to the time point under consideration. This method requires the calculation of as many survival rates as there are events, unless several events occur at the same time. Hence, the K–M method is also called the product–limit (P–L) method.

The survival curve as plotted by the K–M method consists of horizontal lines with vertical steps corresponding to each event (a step-ladder). The vertical axis of the graph represents the survival rate of members of the study group. The magnitude of the step is related to the number of at-risk observations and the number of events occurring. The points on the survival curve are the best estimate of the survival rate of members of the study group.

The survival rates at the beginning of the curve are more definite and reliable in both the actuarial and the K–M method, because a greater number of study individuals are at risk during this time interval. However, at the tail end of the curve, the number of at-risk observations is relatively small due to increased censored observations which result in fewer patients being followed-up for that length of time.

Mean survival time

The mean survival time is the average duration of the follow-up time of all observations in the study. To estimate mean survival time accurately, all individuals must have a 'complete follow-up time', i.e. all observations must be 'failures' by the closure of the study. If this condition is not met, the estimate of mean survival time will be based on incomplete observations (censored)[14] and will not reflect a true picture of the data.

Median survival time

The median survival time is an easily calculated location parameter which provides an interpretable summary of the data. It is defined as the survival time at which the survival rate is equal to 50%.[14] This value can be estimated from the survival curve plotted for survival rates estimated by using either the actuarial or the K–M method.

Confidence interval

The estimation of variability of a survival rate becomes especially essential when a calculation is carried out in small sample groups. Variability is depicted by estimating the standard error (SE), followed by its confidence interval (CI). Several formulae have been proposed for the calculation of the SE and CI for an estimated survival rate at a given time point.[5–7]

A rough estimate of the standard error (SE) of a survival rate (P_i) at the end of a given interval (i) calculated using the formula of Peto *et al.*[6] is:

$$SE(P_i) = P_i \sqrt{\frac{(1-P_i)}{N_i}},$$

where N_i is the number of observations at risk during the interval i. On the assumption that P_i will have an approximately normal sampling distribution, the 95% CI for P_i is $P_i \pm 1.96\ SE(P_i)$.

Another formula, suggested by Greenwood,[5] for the standard error of P_i is based on an estimate of the variance of P_i and is:

$$SE(P_i) = P_i X \sqrt{\sum_{j=1}^{i} [d_j/N_j(N_j - d_j)]},$$

where N_j is the number at risk and d_j is the number of deaths during the interval j. On the assumption that P_i will have an approximately normal sampling distribution, the 95% CI for P_i is $P_i \pm 1.96\ SE(P_i)$.

Rothman, in 1978,[7] suggested another formula to calculate CI in which the limits will always lie between 0 and 1:

$$\frac{n_0}{n_0 + z_{\alpha/2}^2}\left[P_i + \frac{z_{\alpha/2}^2}{2n_0} \pm z_{\alpha/2}\sqrt{\frac{P_i(1-P_i)}{n_0} + \frac{z_{\alpha/2}^2}{4n_0^2}}\right],$$

where $n_0 = P_i(1-P_i)/V_i$, V_i is the Greenwood's variance at time interval i and z_α is the standard value of the Z-statistic at α level of significance.

Comparing survival rates

Comparing two survival rates

A comparison of survival rates between two groups can be made at any time point on the survival curve. The standard Z-test provides a numerical estimate of the probability that a difference as large as that observed would have occurred if only chance were operating.[12]

If P_1 and P_2 are the survival rates at one time point for two groups, 1 and 2, and $SE(P_1)$ and $SE(P_2)$ are the standard errors for P_1 and P_2, respectively, then the Z statistic is calculated by the formula:

$$Z = \frac{|P_1 - P_2|}{\sqrt{[SE^2(P_1) + SE^2(P_2)]}},$$

where $|P_1 - P_2|$ is the absolute value of the difference in survival rates. If the null hypothesis is true, Z approximately follows a normal distribution, i.e. if the observed value of Z for a particular data set is $\geqslant 1.96$, then the probability that a difference as large as that observed occurred by chance is 5% or less.

Rank tests to compare survival curves

The difference between two or more survival curves can be tested using a rank test, by making optimum use of the available information about the survival of the patients in different groups. One commonly used rank test is the log-rank test.[6] This test is designed particularly to detect a difference between two or more survival curves which results when the mortality rate in one group is consistently higher than the corresponding rate in a second group and the ratio of these two rates is constant over time.

If O_1 and O_2 are the observed number of deaths in two groups, 1 and 2, and E_1 and E_2 are the corresponding expected number of deaths, then the log-rank statistic (T) used for comparison purposes is:

$$T = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2.$$

Based on the number of individuals in each group who are alive just before the observed death time and the total number of deaths observed at that time, the expected number of deaths in each group can be calculated. If the null hypothesis is true, T follows an approximately chi-squared distribution with one degree of freedom, i.e. if the observed value of T for a particular data set is $\geqslant 3.84$, then the probability that the difference observed occurred by chance is 5% or less.

There are a number of other tests for comparing the survival of two or more groups.[8,15–20] The generalization proposed by Breslow and Gehan[8,16] of the Mann–Whitney *U*-test (or the Kruskal–Wallis test for more than two populations) is similar to the log-rank test. The Gehan–Breslow test is based on a comparison of observed deaths with expected deaths in a group at each time point where at least one death is observed. The tests only differ in the weight given to the difference between the observed and expected deaths in each group. Another commonly used test is Peto's generalized Wilcoxon statistic.[18] This method differs from the log-rank test in that it attaches more importance to early deaths than to later deaths, whereas the log-rank test gives equal weight to all deaths.

Stratified comparison of survival

The survival of a group of patients is generally associated with many prognostic factors which are interrelated. The comparison should take these prognostic factors into account using a method based on an appropriate stratification, provided that the necessary information is available for each subject. By employing the stratified rank test, the survival experiences of different levels of one factor can be assessed after adjusting for other prognostic factors.[9]

Regression models

Several survival distributions, such as the exponential and Weibull distributions, were introduced to model the survival experience of an homogenous population.[21] However, failure time usually depends upon several prognostic factors.

Suppose a set of prognostic factors (or covariates), z, has been observed, which may include both quantitative and qualitative variables. Qualitative variables can be incorporated through the use of indicator variables.

Exponential and Weibull regression models

The exponential distribution can be generalized to obtain a regression model by allowing the failure rate to be a function of the covariate z. In failure time models, failures have been termed as 'hazards'. Thus, the hazard for a given z is a constant characterizing an exponential failure time distribution, but the failure rate depends on z. The exponential distribution model is:

$$h\,(t,z) = h\,e^{zB},$$

where h (t,z) is the hazard at time t for an individual with a given set of covariates, z; B is a vector of unknown regression coefficients; and h is a constant. The model assumes a multiplicative relationship between the hazard function and the effect of covariates and specifies that log failure rate is a linear function of the covariates.

The Weibull distribution can also be generalized in essentially the same way. The model is:

$$h\,(t,z) = hp(ht)^{p-1}e^{zB},$$

where h (t,z) is the hazard function of an individual with a given set of covariates, z; B is a vector of unknown regression coefficients; and h and p are constants where h, p>0. In this model also the effect of the covariates acts multiplicatively on the hazard function.

Table 2. Survival data of 11 patients with mucosal melanoma of the head and neck

Patient no.	Age (years)	Sex	Site*	Regional node†	Survival (months)	Vital status
1	35	M	1	1	14	Dead
2	50	M	1	1	96	Alive
3	45	F	1	1	5	LFU‡
4	40	F	2	0	1	Dead
5	62	M	1	1	3	LFU
6	56	M	1	1	15	LFU
7	30	F	2	1	18	Dead
8	28	M	1	1	10	Dead
9	60	M	2	1	8	Alive
10	55	M	1	0	12	Alive
11	60	M	1	0	1	Alive

* 1, Oral cavity; 2, pharynx.
† 0, Node absent; 1, node present.
‡ Lost to follow-up.

The proportional hazards model

The exponential and Weibull models involve stronger distributional assumptions than are suitable and the inference procedures may not be sufficiently robust to departures from these assumptions. The distribution of survival times must also be known to apply these models. In most studies, however, the distribution of survival times is unknown and can vary from one disease to another. Hence, these models are too rigid to take into account the diversity of situations which are encountered in practice.

Cox, in 1972,[10] developed a modelling procedure termed Cox's proportional hazards model for use when the number of prognostic factors is large. The model helps to assess the effect of various prognostic factors on survival after adjusting each for the other factors. In other words, using this approach it is possible to identify the independent prognostic factors associated with survival. The advantage of this model is that it does not assume a specific mathematical distribution for observed failures over time. The model measures the hazard ratio of the outcome under the assumption that the hazard is constant over the follow-up period.

The proportional hazards model is:

$$h(t, z) = h_0(t) \exp(B' z),$$

where h(t, z) is the hazard function of an individual with a set of covariates (prognostic factors) z; B is a vector of unknown base line regression coefficients; and $h_0(t)$ is an unknown hazard function for an individual with covariate vector $z = 0$. The model assumes a multiplicative relationship between the hazard function and the set of covariates. The second assumption is that the effect of the covariates on the hazard function is log-linear.

Detailed procedures for the estimation of the above regression coefficients and hazards ratio are not described here. Computer programmes for these models are available in all major statistical software packages, such as SPSS, SAS, BMDP, GLIM and EGRET, which makes the application of these models much easier than in the past.

Illustration

In order to compare the various methods discussed above, data relating to patients with a histopathological diagnosis of malignant mucosal melanoma of the upper aero-digestive tract treated between 1982 and 1996 at the Regional Cancer Centre, Trivandrum, were considered. The date of diagnosis was taken as the starting date and December 1997 was the common closing date of the study. A total of 11 patients were reported during this period. Four patients died, three were lost to follow-up before the closure of the study and four were withdrawn alive at the end of the study. Table 2 shows the individual survival data, along with other characteristics for these 11 patients. The variables sex, site at diagnosis (oral cavity/pharynx) and regional node involvement (absent/present) were used to investigate prognostic importance and for the application of Cox's proportional hazards model.

Table 3 shows the survival probabilities calculated by the direct method. In the data set, five patients (Nos 1, 2, 6, 7 and 11) completed a follow-up time of 1 year and two died before 1 year (Nos 4 and 8). Thus, a total of seven patients were included in the denominator for 1-year survival calculation and the 1-year survival probability was 5/7 (0.71). Similarly, one patient completed 2 years of follow-up (No. 2) and four died before 2 years (Nos 1, 4, 7 and 8), hence the 2-year survival probability was 1/4 (0.20).

The actuarial method for calculating survival rates using all the available follow-up information on the 11 patients under study is illustrated in Table 3. An interval of 12 months was used. LFU and withdrawn alive (due to closure of the study) patients are treated in the same way and are considered as censored observations. The 1-year and 2-year survival rates calculated by this method are 76 and 33%, respectively.

Table 3 also shows the details of the calculation of survival rates employing the Kaplan–Meier method. One patient died at 1 month and no observations are censored before 1 month, hence the conditional probability of surviving for at least 1 month was 10/11 (91%). Similarly, one death occurred at the 10th month and five patients were censored

Table 3. Survival rates using different methods: melanoma data

Interval (months)	At risk (*n*)	Censored (*n*)	Death	Effective *n* at risk	Conditional probability of death	Conditional probability of survival	Survival rate
Direct method*							
12	7	—	2	—	—	—	0.71
24	5	—	4	—	—	—	0.20
>24	5	—	4	—	—	—	0.20
Actuarial method							
1–12	11	5	2	8.5	0.24	0.76	0.76
12–24	4	1	2	3.5	0.57	0.43	0.33
>24	1	1	—	0.5	—	—	0.33
Kaplan–Meier method							
1	11	—	1	—	0.09	0.91	0.91
10	6	—	1	—	0.17	0.83	0.76
14	4	—	1	—	0.25	0.75	0.57
18	2	—	1	—	0.50	0.50	0.29

* Data for patients 3, 5, 9 and 11 could not be used in the computations of 12-month survival due to <12 months follow-up. Similarly, for computation of the 24-month survival rate data relating to patients 1, 2, 4, 7 and 8 alone are used.

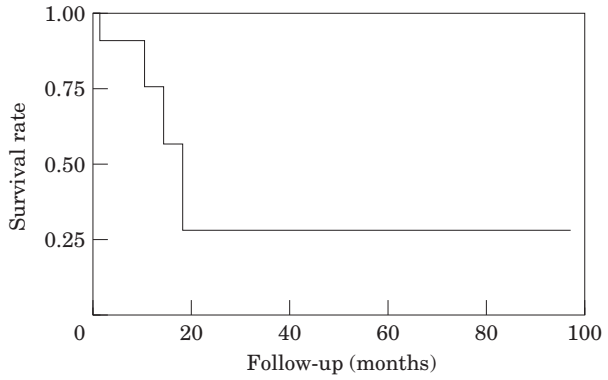

Fig. 1. Kaplan–Meier survival curve for melanoma data.

before completing 10 months of follow-up, hence the conditional probability of survival at the 10th month was 83% and the probability of surviving at least 10 months, obtained by multiplying the preceding conditional probabilities of survival, was 76%. This rate remained the same until the 18th month, as no deaths had occurred in-between. One death occurred in the 18th month and hence the survival rate for at least 18 months was 28.4%. As no deaths occurred after the 18th month, the survival rate remained the same until the completion of the follow-up period. Hence the 2-year survival rate using the K–M method was also 28.4%. The Kaplan–Meier survival curve (step-ladder) for the 11 patients is shown in Fig. 1.

Table 4 shows the confidence intervals for the survival rates estimated using Peto's and Rothman's formulae with Greenwood's variance. Confidence intervals calculated by Peto's formula are substantially wider than those calculated by Rothman's formula. Also, Peto's intervals for survival probabilities lie outside 0 and 1.

Table 4. Peto's and Rothman's confidence intervals for survival rates

Follow-up (months)	K–M survival rate	Peto's method	Rothman's method
1	0.91	0.73–1.09	0.51–0.99
10	0.76	0.42–1.10	0.31–0.94
14	0.57	0.08–1.06	0.15–0.85
18	0.28	−0.34–0.90	0.01–0.69

Table 5. Mean and median survival time according to prognostic factors

Factor	Survival time (months)	Standard error	95% CI
Sex			
Male			
Mean	57	21	16–98
Median	Undefined	—	—
Female			
Mean	12	7	0–25
Median	18	0	—
Site			
Oral cavity			
Mean	57	21	16–98
Median	Undefined	—	—
Pharynx			
Mean	12	7	0–24
Median	18	0	—
Node			
Absent			
Mean	8	3	2–14
Median	Undefined	—	—
Present			
Mean	39	19	1–77
Median	18	3	12–24

Overall mean survival time: 37 months (95% CI: 1–72).
Median survival time: 18 months (95% CI: 12–24).

The overall mean and median survival times were 37 and 18 months, respectively, for the above data. The median survival time could not be calculated for some levels of prognostic factors, such as male sex, site at diagnosis = oral cavity and no node involvement, as the survival rates at the end of the study exceeded 50% for these factors (Table 5).

To compare the survival rates according to the various prognostic factors, the Z-test, log-rank test and Breslow's

Table 6. Comparison of survival rates, using the Z-test, log-rank and Breslow tests, according to prognostic factors

Factor	Censored (*n*)	Deaths (*n*)	K–M 2-year survival rate (%)	Z-test *P*-value	Log-rank *P*-value	Breslow's *P*-value
Sex						
Male	6	2	53.3	0.0316†	0.357	0.223
Female	1	2	0			
Site						
Oral cavity	6	2	53.3	0.0316†	0.357	0.223
Pharynx	1	2	0			
Node						
Absent	2	1	0*	0.2077	0.334	0.194
Present	5	3	30.0			

* Of three node-negative patients, one died while the other two patients were included in 1996 and hence have limited follow-up.

† The Z-test shows a significant difference in survival between men and women. However, this became insignificant when tested using the log-rank and Breslow's tests.

Table 7. Results of Cox's proportional hazards model

Factor	Hazard ratio	95% CI
Sex		
Male	1	—
Female	1.68	0.04–667.4
Site		
Oral cavity	1	—
Pharynx	1.68	0.04–667.4
Node		
Absent	1	—
Present	0.25	0.01–4.50

test were used. The 2-year survival rates were compared using the Z-test. These rates were significantly different at the 5% level of significance for two levels of prognostic factors, sex and nodal status. When overall survival curves were compared using both the log-rank and Breslow's tests, none of the variables were significant. The log-rank test provided higher *P*-values than Breslow's test for all the prognostic factors studied (Table 6).

Survival estimates could not be calculated for sex when stratified according to the nodal status as there were no observations in the stratum for men with no node involvement. This limitation was overcome by using Cox's proportional hazards method. All three factors could be included in the model simultaneously. The hazards ratio for each prognostic factor, after adjusting for the two others, could be estimated. None of the factors showed significance (Table 7) and the wide confidence intervals are due to the small number of observations.

Discussion

We chose patients with mucosal malignant melanoma of the upper aero-digestive tract in order to illustrate all the methods of survival analysis using a small sample data set. Primary malignant melanoma is a rare disease in Asia, and mucosal melanoma is even rarer. The number of patients reported at the Regional Cancer Centre, Trivandrum, in the 15-year study period was, hence, very small, therefore, these data were determined to be most suitable to identify the caveats and pitfalls of the various methods under discussion.

When estimating survival using the direct method, rates cannot be computed for censored observations for whom the period of follow-up is less than the specified time of survival estimation. On the other hand, the actuarial method has the advantage of using all follow-up information accumulated up to the closing date of the study. The 2-year survival rate obtained by the actuarial method for melanoma patients does not support the rate obtained by the direct method. The conflict between the two survival rates is because of the exclusion of six patients due to follow-up of less than 2 years in the direct method. Thus, the actuarial method is more reliable than the direct method as it is based on all observations.

The K–M method is particularly suited to small sample sizes. However, both the actuarial and the K–M methods give similar results when applied to a large group of data sets. The fixed intervals of the actuarial method are perfectly suited to the classic presentation of survival probability, for a given number of years, after diagnosis. However, with the greater use of computers seen in recent years, it is now common practice to base an analysis on precise survival times, rather than grouped data, making the application of the Kaplan–Meier method much easier than in the past. Hence, it would be more appropriate to use the K–M method irrespective of sample size, as its use is generally more widespread.[22]

Even though both the actuarial and K–M methods utilize all censored observations, the difference in survival rates observed by the two methods is mainly due to differences in the estimation of the number of observations at risk, i.e. the number at risk in the actuarial method is obtained by subtracting half the number of subjects censored from the number of observations at the beginning of interval, whereas in the K–M method the true number at risk at a given time is calculated by subtracting all the censored observations prior to the specified time point. Hence survival rates based on the K–M method are better and more reliable than the actuarial method, even though in both methods all 11 patients were included for rate calculation.

The approach suggested by Peto *et al.*[6] to estimate the confidence interval is not a good approximation for small sample sizes or for very large or very small survival probabilities, even though it is an easily obtained estimate of the magnitude of variability. In such circumstances, the confidence interval for survival probability can even go outside the range of 0–1, as illustrated in the present study. The formula for standard error derived by Greenwood[5] depends on the estimate of the variance. This formula can lead to an underestimation of the variance for long time intervals when the sample size is not sufficiently large. Hence it is valid for the large samples only. Anderson *et al.*, in 1982,[23] have shown that Rothman's method with Greenwood's variance provides the most satisfactory results on average. In the present illustration, the confidence intervals using Peto's method were substantially wider than those found using Rothman's method. Also, Peto's confidence interval for survival probability had gone beyond the range of 0–1, whereas Rothman's intervals were within the probability limits.

Median survival time is a better estimate than mean survival time. This is because one or two exceptionally long survivors will cause the mean survival time to be larger than the median survival time. Also, calculations of the mean survival time based on censored observations cannot be relied upon.[14] The survival estimation can never equal zero if the largest observation is censored. Censored data can also complicate the calculation of median survival time. If the survival rate at the end of follow-up exceeds 50%, the median survival time cannot be estimated.

In the present illustration, the overall mean survival time was more than double the median survival time. This is clearly due to the largest censored observation having a follow-up of 96 months. However, median survival time is also undefined for certain levels of prognostic factors for which the survival rates at the end of the study exceeded 50% (Table 5).

The general application of a comparison of survival rates between two groups at a particular time point to the entire survival curve can lead to incorrect conclusions. Survival probabilities in the two groups might be different at a particular time point, but the two survival curves may not reveal statistical significance when all the time points are considered. The difference between two or more survival curves should be tested, making optimum use of all available follow-up information on the survival of patients in the different groups. In the present illustration, 2-year survival rates for different levels of variables such as sex and nodal status were significantly different when the Z-test was employed. However, they were not found to be significant when the entire survival curves were compared using the log-rank and Breslow's tests.

Obviously, not every test described for the comparison of survival curves needs to be used for each data set under study, nor should a test be selected solely for its convenience. In fact, the choice of a test should, as always, be directed by a hypothesis made prior to the examination of the survival curves. When the ratio of the mortality rates is not constant over time, the difference between survival curves will be more correctly determined by the Gehan–Breslow test than the log-rank test. However, as Gehan–Breslow's test can behave unpredictably in certain circumstances, caution is recommended in its use. In particular, the result of this test should be ignored if it does not show significance when the log-rank test does, especially when there is a large number of censored observations.[22]

The effectiveness of stratification as a means of adjusting for other prognostic factors is limited because it is necessary to retain a moderate number of subjects in each stratum. With an increase in the number of prognostic factors, the strata soon become too small to be meaningful unless the study is very large, therefore the use of a larger number of strata is generally unwise. In the present illustration, stratified analysis could not be performed for nodal status with sex as there were no observations in the stratum for men with no involved nodes.

Other useful procedures exist in survival analysis to take into account a large number of prognostic factors. In describing all the data using one model, the comparison being made can be summarized. Cox's regression model is a useful procedure when the number of prognostic factors is large. The suitability of the exponential and Weibull models was not tested in the present illustration as the failure time distribution of mucosal melanoma was unknown. The more general applicability of Cox's model is because it does not assume any specific mathematical distribution. The hazard ratios for the prognostic factors sex, site at diagnosis and nodal status, after adjusting each factor for the others, could be estimated in the present illustration with a sample size of 11 patients.

The analysis of the present data set clearly illustrates that, despite the small sample size and multiple factors, survival analysis can be carried out with a high degree of conviction. All that is required is to choose the appropriate method for any given data set. We suggest that, in publishing results from survival analysis, measurements of location and variability such as median survival time and confidence interval, respectively, should be made mandatory. The use of mean survival time should be uniformly discouraged.

References

1. Berkson J, Gage RP. Calculation of survival rates for cancer. *Proc Staff Meet Mayo Clin* 1950; **25**: 270–86.
2. Cutler SJ, Ederer F. Maximum utilization of the life table method in analysing survival. *J Chron Dis* 1958; **8**: 699–712.
3. Chiang CL. *Introduction to Stochastic Process in Biostatistics*. New York: John Wiley, 1968.
4. Kaplan EL, Meier P. Non-parametric estimation from incomplete observation. *J Am Stat Assoc* 1958; **53**: 457–81.
5. Greenwood M. *A Report on the Natural Duration of Cancer Reports in Public Health and Medical Subjects*. Ministry of Health, London: HMSO, 1926; **33**: 1–26.
6. Peto R, Pike MC, Armitage P, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient II. Analysis and examples. *Br J Cancer* 1977; **35**: 1–39.
7. Rothman K. Estimates of confidence limits for the cumulative probability of survival in life table analysis. *J Chron Dis* 1978; **31**: 557–60.
8. Breslow N. A generalised Kruskal–Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika* 1970; **57**: 579–94.

9. Mantel N. Evaluation of survival rates and two rank order statistics arising in its consideration. *Cancer Chemother Rep* 1966; **50**: 163–70.
10. Cox DR. Regression models and life-tables (with discussion). *J Roy Stat Soc* 1972; **B-34**: 187–220.
11. Pandey M, Mathew A, Abraham EK, Ahamed MI. Malignant mucosal melanoma of the upper aerodigestive tract. *Int J Oral Maxillofac Surg* 1998; **28**: 45–9.
12. AJCC (American Joint Committee on Cancer). *Reporting of Cancer Survival and End-results, Task Force on Statistics.* Chicago, IL, 1982.
13. Mould RF. Calculation of survival rates by the life table and other methods. *Clin Radiol* 1976; **27**: 33–8.
14. Gross AJ, Clark VA. *Survival Distributions: Reliability Applications in the Biomedical Sciences*. New York: John Wiley, 1975.
15. Mantel N, Heinzel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Nat Can Inst* 1959; **22**: 719–48.
16. Gehan EA. A generalised Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965; **52**: 203–23.
17. Peto R. Rank tests of maximal power against Lehman, Type alternatives. *Biometrika* 1972; **59**: 472–4.
18. Peto R, Peto J. Asymptomatically efficient rank invariant test procedures with discussion. *J Roy Stat Soc A* 1972; **135**: 185–207.
19. Peto R, Pike MC. Conversion of approximation $(O\text{-}E)^2/E$ in the log-rank test for survival data or tumour incidence data. *Biometrics* 1973; **29**: 579–84.
20. Breslow N. Statistical methods for censored survival data. *Environ Health Perspec* 1979; **32**: 181–92.
21. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley, 1980.
22. Esteve J, Benhamou E, Raymond L. *Statistical Methods in Cancer Research Volume IV, Descriptive Epidemiology.* Translated from French by Sinclair M. Lyon, France: IARC Scient Pub 1994; **128**.
23. Anderson JR, Bernstein L, Pike MC. Approximate confidence intervals for probabilities of survival and quantiles in life table analysis. *Biometrics* 1982; **38**: 407–16.