British Journal of Cancer (2003) 89, 232–238
© 2003 Cancer Research UK All rights reserved 0007–0920/03 $25.00

www.bjcancer.com

Tutorial Paper

Survival Analysis Part I: Basic concepts and first analyses

TG Clark*,[1], MJ Bradburn[1], SB Love[1] and DG Altman[1]

[1]Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, University of Oxford, Old Road, Oxford OX3 7LF, UK

British Journal of Cancer (2003) **89,** 232–238. doi:10.1038/sj.bjc.6601118 www.bjcancer.com
© 2003 Cancer Research UK

Keywords: survival analysis; statistical methods; Kaplan-Meier

INTRODUCTION

In many cancer studies, the main outcome under assessment is the time to an event of interest. The generic name for the time is *survival time*, although it may be applied to the time 'survived' from complete remission to relapse or progression as equally as to the time from diagnosis to death. If the event occurred in all individuals, many methods of analysis would be applicable. However, it is usual that at the end of follow-up some of the individuals have not had the event of interest, and thus their true time to event is unknown. Further, survival data are rarely Normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. It is these features of the data that make the special methods called *survival analysis* necessary.

This paper is the first of a series of four articles that aim to introduce and explain the basic concepts of survival analysis. Most survival analyses in cancer journals use some or all of Kaplan–Meier (KM) plots, logrank tests, and Cox (proportional hazards) regression. We will discuss the background to, and interpretation of, each of these methods but also other approaches to analysis that deserve to be used more often. In this first article, we will present the basic concepts of survival analysis, including how to produce and interpret survival curves, and how to quantify and test survival differences between two or more groups of patients. Future papers in the series cover multivariate analysis and the last paper introduces some more advanced concepts in a brief question and answer format. More detailed accounts of these methods can be found in books written specifically about survival analysis, for example, Collett (1994), Parmar and Machin (1995) and Kleinbaum (1996). In addition, individual references for the methods are presented throughout the series. Several introductory texts also describe the basis of survival analysis, for example, Altman (2003) and Piantadosi (1997).

TYPES OF 'EVENT' IN CANCER STUDIES

In many medical studies, time to death is the event of interest. However, in cancer, another important measure is the time between response to treatment and recurrence or relapse-free survival time (also called disease-free survival time). It is important to state what the event is and when the period of observation starts and finishes. For example, we may be interested in relapse in the time period between a confirmed response and the first relapse of cancer.

CENSORING MAKES SURVIVAL ANALYSIS DIFFERENT

The specific difficulties relating to survival analysis arise largely from the fact that only some individuals have experienced the event and, subsequently, survival times will be unknown for a subset of the study group. This phenomenon is called censoring and it may arise in the following ways: (a) a patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time of the close of the study; (b) a patient is lost to follow-up during the study period; (c) a patient experiences a different event that makes further follow-up impossible. Such censored survival times underestimate the true (but unknown) time to event. Visualising the survival process of an individual as a time-line, their event (assuming it were to occur) is beyond the end of the follow-up period. This situation is often called *right censoring*. Censoring can also occur if we observe the presence of a state or condition but do not know where it began. For example, consider a study investigating the time to recurrence of a cancer following surgical removal of the primary tumour. If the patients were examined 3 months after surgery to determine recurrence, then those who had a recurrence would have a survival time that was *left censored* because the actual time of recurrence occurred less than 3 months after surgery. Event time data may also be *interval censored,* meaning that individuals come in and out of observation. If we consider the previous example and patients are also examined at 6 months, then those who are disease free at 3 months and lost to follow-up between 3 and 6 months are considered interval censored. Most survival data include right censored observations, but methods for interval and left censored data are available (Hosmer and Lemeshow, 1999). In the remainder of this paper, we will consider right censored data only.

In general, the feature of censoring means that special methods of analysis are needed, and standard graphical methods of data exploration and presentation, notably scatter diagrams, cannot be used.

ILLUSTRATIVE STUDIES

Ovarian cancer data

This data set relates to 825 patients diagnosed with primary epithelial ovarian carcinoma between January 1990 and December 1999 at the Western General Hospital in Edinburgh. Follow-up data were available up until the end of December 2000, by which time 550 (75.9%) had died (Clark *et al*, 2001). Figure 1 shows data from 10 patients diagnosed in the early 1990s and illustrates how patient profiles in calendar time are converted to time to event

*Correspondence: Mr TG Clark; E-mail: taane.clark@cancer.org.uk

Received 6 December 2002; accepted 30 April 2003

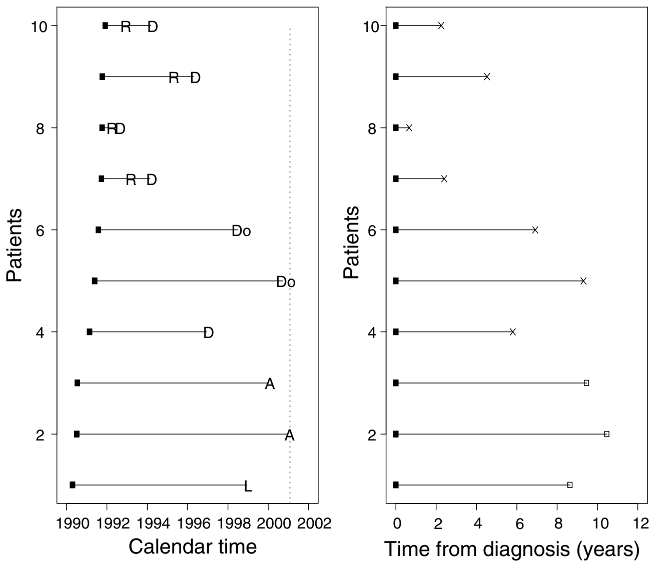


Figure 1 Converting calendar time in the ovarian cancer study to a survival analysis format. Dashed vertical line is the date of the last follow-up, R = relapse, D = death from ovarian cancer, Do = death from other cause, A = attended last clinic visit (alive), L = loss to follow-up, X = death, □ = censored.

(death) data. Figure 1 (left) shows that four patients had a nonfatal relapse, one was lost to follow-up, and seven patients died (five from ovarian cancer). In the other plot, the data are presented in the format for a survival analysis where all-cause mortality is the event of interest. Each patient's 'survival' time has been plotted as the time from diagnosis. It is important to note that because overall mortality is the event of interest, nonfatal relapses are ignored, and those who have not died are considered (right) censored. Figure 1 (right) is specific to the outcome or event of interest. Here, death from any cause, often called overall survival, was the outcome of interest. If we were interested solely in ovarian cancer deaths, then patients 5 and 6 – those who died from nonovarian causes – would be censored. In general, it is good practice to choose an end-point that cannot be misclassified. All-cause mortality is a more robust end-point than a specific cause of death. If we were interested in time to relapse, those who did not have a relapse (fatal or nonfatal) would be censored at either the date of death or the date of last follow-up.

Lung cancer clinical trial data

These data originate from a phase III clinical trial of 164 patients with surgically resected (non-small cell) lung cancer, randomised between 1979 and 1985 to receive radiotherapy either with or without adjuvant combination platinum-based chemotherapy (Lung Cancer Study Group, 1988; Piantadosi, 1997). For the purposes of this series, we will focus on the time to first relapse (including death from lung cancer). Table 1 gives the time of the earliest 15 and latest five relapses for each treatment group, where it can be seen that some patients were alive and relapse-free at the end of the study. The relapse proportions in the radiotherapy and combination arms were 81.4% (70 out of 86) and 69.2% (54 out of 78), respectively. However, these figures are potentially misleading as they ignore the duration spent in remission before these events occurred.

SURVIVAL AND HAZARD

Survival data are generally described and modelled in terms of two related probabilities, namely *survival* and *hazard*. The survival probability (which is also called the survivor function) $S(t)$ is the

Table 1 A sample of times (days) to relapse among patients randomised to receive radiotherapy with or without adjuvant chemotherapy

Radiotherapy ($n = 86$)	18, 23[a], 25, 27, 28, 30, 36, 45, 55, 56, 57, 57, 57, 59, 62, …, 2252[a], 2286[a], 2305[a], 2318[a], 2940[a]
Radiotherapy+CAP ($n = 78$)	9, 22, 35, 53, 76, 81, 94, 97, 103, 114, 115, 126, 147, 154, …, 2220[a], 2375, 2566, 2875[b], 3067[b]

CAP = cytoxan, doxorubicin and platinum-based chemotherapy. [a]Lost to follow-up and considered censored. [b]Relapse-free at time of analysis and considered censored.

probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified future time t. It is fundamental to a survival analysis because survival probabilities for different values of t provide crucial summary information from time to event data. These values describe directly the survival experience of a study cohort.

The hazard is usually denoted by $h(t)$ or $\lambda(t)$ and is the probability that an individual who is under observation at a time t has an event at that time. Put another way, it represents the instantaneous event rate for an individual who has already survived to time t. Note that, in contrast to the survivor function, which focuses on not having an event, the hazard function focuses on the event occurring. It is of interest because it provides insight into the conditional failure rates and provides a vehicle for specifying a survival model. In summary, the hazard relates to the incident (current) event rate, while survival reflects the cumulative non-occurrence.

KAPLAN–MEIER SURVIVAL ESTIMATE

The survival probability can be estimated nonparametrically from observed survival times, both censored and uncensored, using the KM (or product-limit) method (Kaplan and Meier, 1958). Suppose that k patients have events in the period of follow-up at distinct times $t_1 < t_2 < t_3 < t_4 < t_5 < \cdots < t_k$. As events are assumed to occur independently of one another, the probabilities of surviving from one interval to the next may be multiplied together to give the cumulative survival probability. More formally, the probability of being alive at time t_j, $S(t_j)$, is calculated from $S(t_{j-1})$ the probability of being alive at t_{j-1}, n_j the number of patients alive just before t_j, and d_j the number of events at t_j, by

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

where $t_0 = 0$ and $S(0) = 1$. The value of $S(t)$ is constant between times of events, and therefore the estimated probability is a step function that changes value only at the time of each event. This estimator allows each patient to contribute information to the calculations for as long as they are known to be event-free. Were every individual to experience the event (i.e. no censoring), this estimator would simply reduce to the ratio of the number of individuals events free at time t divided by the number of people who entered the study.

Confidence intervals for the survival probability can also be calculated. The KM *survival curve*, a plot of the KM survival probability against time, provides a useful summary of the data that can be used to estimate measures such as median survival time. The large skew encountered in the distribution of most survival data is the reason that the mean is not often used.

Survival analysis of the lung cancer trial

Table 2 shows the essential features of the KM survival probability. The estimator at any point in time is obtained by multiplying a sequence of conditional survival probabilities, with the estimate

Table 2 Calculation of the relapse-free survival probability for patients in the lung cancer trial

Radiotherapy ($n = 86$)		Radiotherapy+CAP ($n = 78$)	
Survival times (days)	**Kaplan–Meier survivor function $S(t)$**	**Survival times (days)**	**Kaplan–Meier survivor function $S(t)$**
18	1 × (1-1/86) = 0.988	9	1 × (1-1/78) = 0.987
23[a]	S(18) × (1-0/85) = 0.988	22	S(18) × (1-1/77) = 0.974
25	S(23) × (1-1/84) = 0.977	35	S(22) × (1-1/76) = 0.962
27	S(25) × (1-1/83) = 0.965	53	S(35) × (1-1/75) = 0.949
28	S(27) × (1-1/82) = 0.953	76	S(53) × (1-1/74) = 0.936
30	S(28) × (1-1/81) = 0.941	81	S(76) × (1-1/73) = 0.923
36	S(30) × (1-1/80) = 0.930	94	S(81) × (1-1/72) = 0.910
45	S(36) × (1-1/79) = 0.918	97	S(94) × (1-1/71) = 0.897
55	S(45) × (1-1/78) = 0.906	103	S(97) × (1-1/70) = 0.885
56	S(55) × (1-1/77) = 0.894	114	S(103) × (1-1/69) = 0.872
57	S(56) × (1-3/76) = 0.859	115	S(114) × (1-1/68) = 0.859
57	S(56) × (1-3/76) = 0.859	121[a]	S(115) × (1-0/67) = 0.859
57	S(56) × (1-3/76) = 0.859	126	S(121) × (1-1/66) = 0.846
59	S(57) × (1-1/73) = 0.847	147	S(126) × (1-1/65) = 0.833
62	S(59) × (1-1/72) = 0.835	154	S(147) × (1-1/64) = 0.820
	⋮		⋮
2252[a]	S(2209) × (1-0/5) = 0.115	2220[a]	S(2218) × (1-0/5) = 0.273
2286[a]	S(2286) × (1-0/4) = 0.115	2375	S(2220) × (1-0/4) = 0.205
2305[a]	S(2305) × (1-0/3) = 0.115	2566	S(2375) × (1-0/3) = 0.137
2318[a]	S(2318) × (1-0/2) = 0.115	2875[b]	S(2566) × (1-0/2) = 0.137
2940[a]	S(2940) × (1-0/1) = 0.115	3067[b]	S(2875) × (1-0/1) = 0.137

$S(0) = 1$, (CAP = cytoxan, doxorubicin and platinum-based chemotherapy.) [a]Lost to follow-up and considered censored. [b]Relapse-free at time of analysis and considered censored.

being unchanged between subsequent event times. For example, the probability of a member of the radiotherapy alone treatment group surviving (relapse-free) 45 days is the probability of surviving the first 36 days multiplied by the probability of then surviving the interval between 36 and 45 days. The latter is a *conditional* probability as the patient needs to have survived the first period of time in order to remain in the study for the second. The KM estimator utilises this fact by dividing the time axis up according to event times and estimating the event probability in each division, from which the overall estimate of the survivorship is drawn.

Figure 2 shows the survival probabilities for the two treatment groups in the conventional KM graphical display. The median survival times for each group are shown and represent the time at which $S(t)$ is 0.5. The combination group has a median survival time of 402 days (1.10 years), as opposed to 232 days (0.64 years) in the radiotherapy alone arm, providing some evidence of a chemotherapy treatment benefit. Other survival time percentiles may be read directly from the plot or (more accurately) from a full version of Table 2. There appears to be a survival advantage in the combination therapy group, but whether this difference is statistically significant requires a formal statistical test, a subject that is discussed later.

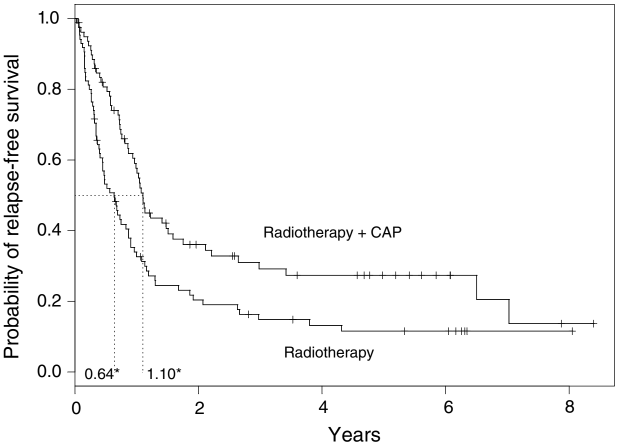


Figure 2 Relapse-free survival curves for the lung cancer trial. * Median relapse-free survival time for each arm, + censoring times, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

Survival function of the ovarian data

The KM survival curve of the ovarian cancer data is shown in Figure 3A. The steep decline in the early years indicates poor prognosis from the disease. This is also indicated by changes in the cumulative number of events and number at risk. Specifically, of the 825 women diagnosed with ovarian cancer, about a third had died within the first year, accounting for 43% of the total deaths as recorded by the last date of follow-up. The number lost to follow-up can be deduced from the total number in the cohort and the cumulative number of events and number at risk.

The 95% confidence limits of the survivor function are shown. In practice, there are usually patients who are lost to follow-up or alive at the end of follow-up, and confidence limits are often wide at the tail of the curve, making meaningful interpretations difficult. Thus, it may be sensible to curtail plots before the end of follow-up on the x-axis (Pocock *et al*, 2002). Curtailing of the y-axis, a common practice for diseases or events of low incidence, should not be performed. Instead, the incidence of death curve, or $1-S(t)$, (Figure 3B) may be presented (Pocock *et al*, 2002). The cumulative incidence at a time point is simply one minus the survival probability. For example, Figure 3A shows how the 5-year survival of 0.29 (29%) is calculated, and could also be read from Figure 3B as a cumulative incidence of 71% for the first 5 years.

HAZARD AND CUMULATIVE HAZARD

There is a clearly defined relationship between $S(t)$ and $h(t)$, which is given by the calculus formula:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}[\log S(t)].$$

The formula is unimportant for routine survival analyses as it is incorporated into most statistical computer packages. The point here is simply that if either $S(t)$ or $h(t)$ is known, the other is automatically determined. Consequently, either can be the basis of statistical analysis.

Unfortunately, unlike $S(t)$ there is no simple way to estimate $h(t)$. Instead, a quantity called the *cumulative hazard* $H(t)$ is commonly used. This is defined as the integral of the hazard, or the area under the hazard function between times 0 and t, and differs from the log-survivor curve only by sign, that is $H(t) = -\log[S(t)]$. The interpretation of $H(t)$ is difficult, but perhaps the easiest way to think of $H(t)$ is as the cumulative force of mortality, or the number of events that would be expected for each individual by time t if the event were a repeatable process. $H(t)$ is used an intermediary measure for estimating $h(t)$ and as a diagnostic tool in assessing model validity. A simple nonparametric method for estimating $H(t)$ is the Nelson-Aalen estimator (Hosmer, 1999), from which it is possible to derive an estimate of $h(t)$ by applying a kernel smoother to the increments (Ramlau-Hansen, 1983). Cox (1979) suggests another method to estimate the hazard based on order statistics but similar in spirit to the previous method.

Another approach to estimating the hazard is to assume that the survival times follow a specific mathematical distribution. Figure 4 shows the relationship between four parametrically specified hazards and the corresponding survival probabilities. It illustrates a constant hazard rate over time (which is analogous to an exponential distribution of survival times), strictly increasing/decreasing hazard rates based on a Weibull model, and a combination of decreasing and increasing hazard rates using a log-Normal model. These curves are illustrative examples and other shapes are possible. The specification of hazards using fully parametric distributions is an important and under-utilised modelling technique that will be discussed in subsequent papers.

Hazard function in the ovarian data

Figure 3C shows the cumulative hazard for the ovarian cancer data. The hazard is shown in Figure 3D. As the hazard function is generally very erratic, it is customary to fit a smooth curve to enable the underlying shape to be seen. Figure 3D shows that the (instantaneous) risk of death appears to be high in the first year after diagnosis and decreases afterwards. This observation corresponds to the steeply descending survival probability (Figure 3A) and marked increase in cumulative incidence (Figure 3B) in the first year. The y-axis is difficult to interpret for the hazard and the cumulative hazard, but the decreasing shape of the hazard may be consistent with a decreasing Weibull's model (see Figure 4).

NONPARAMETRIC TESTS COMPARING SURVIVAL

Survival in two or more groups of patients can be compared using a nonparametric test. The logrank test (Peto *et al*, 1977) is the most widely used method of comparing two or more survival curves. The groups may be treatment arms or prognostic groups (e.g. FIGO stage). The method calculates at each event time, for each group, the number of events one would expect since the previous event if there were no difference between the groups. These values are then summed over all event times to give the total expected number of events in each group, say E_i for group i. The logrank test compares observed number of events, say O_i for treatment group i, to the expected number by calculating the test statistic

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}.$$

This value is compared to a χ^2 distribution with $(g-1)$ degrees of freedom, where g is the number of groups. In this manner, a

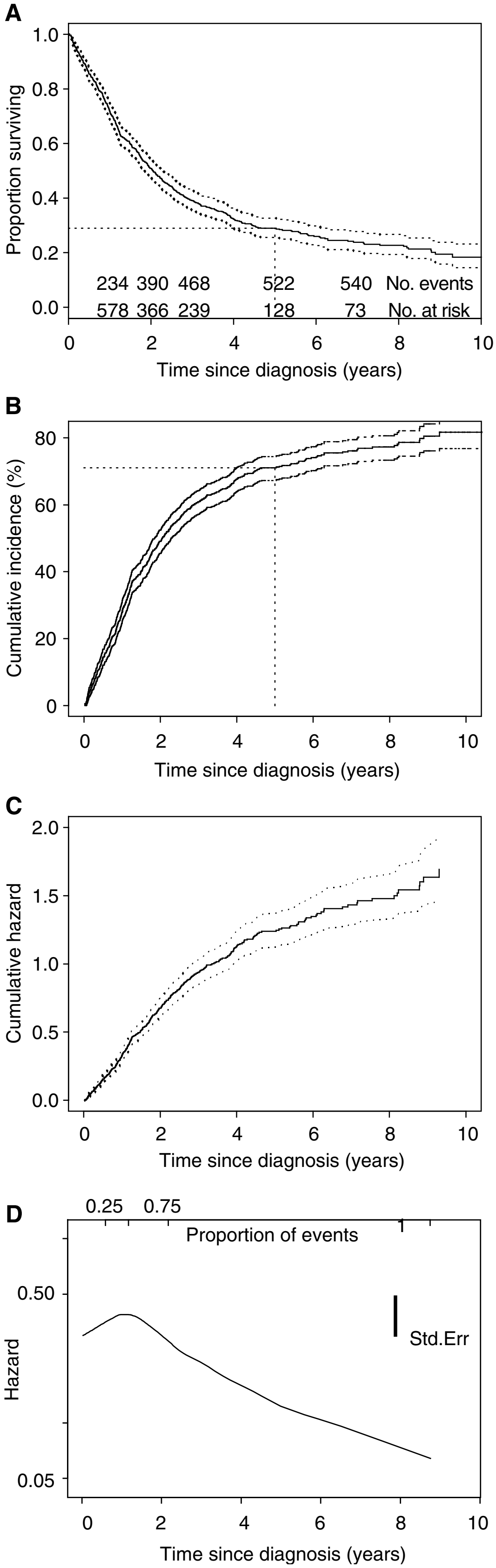


Figure 3 Survival and cumulative hazard curves with 95% CIs for the ovarian cancer study. Std.Err = standard error. (**A**) Kaplan–Meier survivor function, (**B**) cumulative incidence curve, (**C**) cumulative hazard function, (**D**) hazard function (smoothed).

P-value may be computed to calculate the statistical significance of the differences between the complete survival curves.

If the groups are naturally ordered, a more appropriate test is to consider the possibility that there is a trend in survival across them, for example, age groups or stages of cancer. Calculating O_i and E_i for each group on the basis that survival may increase or decrease across the groups results in a more powerful test. For the new O_i and E_i, the test statistic for trend is compared with the χ^2 distribution with one degree of freedom (Collett, 1994).

When only two groups are compared, the logrank test is testing the null hypothesis that the ratio of the hazard rates in the two groups is equal to 1. The hazard ratio (HR) is a measure of the relative survival experience in the two groups and may be estimated by

$$\mathrm{HR} = \frac{O_1/E_1}{O_2/E_2}$$

where O_i/E_i is the estimated relative (excess) hazard in group *i*. A confidence interval (CI) for the HR can be calculated (Collett, 1994). The HR has a similar interpretation of the strength of effect as a risk ratio. An HR of 1 indicates no difference in survival. In practice, it is better to estimate HRs using a regression modelling technique, such as Cox regression, as described in the next article.

Other nonparametric tests may be used to compare groups in terms of survival (Collett, 1994). The logrank test is so widely used that the reason for any other method should be stated in the protocol of the study. Alternatives include methods to compare median survival times, but comparing confidence intervals for each group is not recommended (Altman and Bland, 2003). The logrank method is considered more robust (Hosmer and Lemeshow, 1999), but the lack of an accompanying effect size to compliment the *P*-value it provides is a limitation.

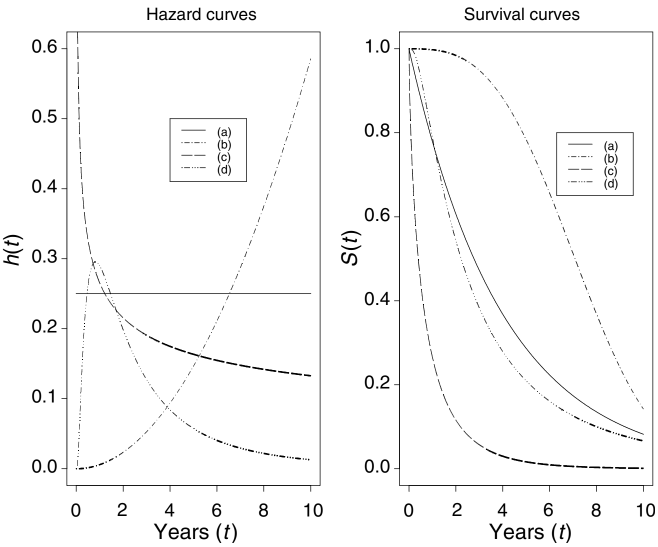


Figure 4 Relationships between (parametric) hazard and survival curves: (a) constant hazard (e.g. healthy persons), (b) increasing Weibull (e.g. leukaemia patients), (c) decreasing Weibull (e.g. patients recovering from surgery), (d) increasing and then decreasing log-normal (e.g. tuberculosis patients).

Survival differences in the lung cancer trial

We have already seen that median survival is greater in the combination treatment arm. Table 3 provides information about (relapse-free) survival differences between the trial arms. A test of differences between median survival times in the groups is indicative of a difference in survival ($P<0.01$). The number of relapses observed among patients treated with radiotherapy + CAP (cytoxan, doxorubin and platinum-based chemotherapy) and radiotherapy alone were 54 and 70, respectively. Using the logrank method, the expected number of relapses for each group were 70.6 and 53.4, respectively. Thus, the logrank test yields a χ^2 value of 9.1 on 1 degree of freedom ($P<0.002$). The HR of 0.58 indicates that there is 42% less risk of relapse at any point in time among patients surviving in the combination treatment group compared with those treated with radiotherapy alone. Overall, there is an indication that the combination treatment is more efficacious than radiotherapy treatment, and may be preventing or delaying relapse.

Survival differences in the ovarian study

In the ovarian study, we wished to compare the survival between patients with different FIGO stages – an ordinal variable. Figure 5 shows overall survival by FIGO stage. A logrank test of trend is statistically significant ($P<0.0001$), and reinforces the visual impression of prognostic separation and a trend towards better survival when the disease is less advanced.

SOME KEY REQUIREMENTS FOR THE ANALYSIS OF SURVIVAL DATA

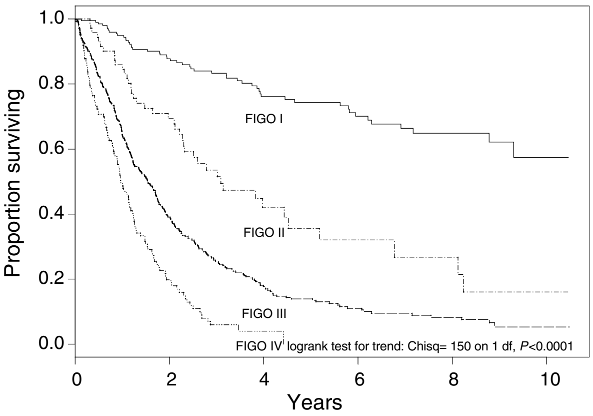
Uninformative censoring

Standard methods used to analyse survival data with censored observations are valid only if the censoring is ‘noninformative’. In practical terms, this means that censoring carries no prognostic information about subsequent survival experience; in other words, those who are censored because of loss to follow-up at a given point in time should be as likely to have a subsequent event as those individuals who remain in the study. Informative censoring may occur when patients withdraw from a clinical trial because of drug toxicity or worsening clinical condition. Standard methods for survival analysis are not valid when there is informative censoring. However, when the number of patients lost to follow-up is small, very little bias is likely to result from applying methods based on noninformative censoring.

Table 3 Differences in (relapse-free) survival in the lung cancer trial

	Radiotherapy ($n=86$)	**Radiotherapy+CAP ($n=78$)**
Number of relapses (O_i)	70	54
Median survival time(years) (95% CI)	0.64 (0.45–0.87)	1.10 (0.96–1.59)
Expected number of relapses (E_i)	53.4	70.6
Hazard ratio (95% CI)	0.58 (0.41–0.83)	
Logrank test	$\chi^2=9.1$, 1 df, $P<0.002$	

df = degree of freedom: CAP = cytoxan, doxorubicin and platinum-based chemotherapy.



Number at risk at time (years)	0	2	4	6	8
FIGO I	200	151	93	66	44
FIGO II	72	45	17	10	6
FIGO III	414	142	48	23	14
FIGO IV	123	22	3	1	1

Figure 5 FIGO stage and prognosis in the ovarian study. Chisq $= \chi^2$.

Length of follow-up

In general, the design of a study will influence how it is analysed. Time to event studies must have sufficient follow-up to capture enough events and thereby ensure there is sufficient power to perform appropriate statistical tests. The proposed length of follow-up for a prospective study will be based primarily on the severity of the disease or prognosis of the participants. For example, for a lung cancer trial a 5-year follow-up would be more than adequate, but this follow-up duration will only give a short- to-medium-term indication of survivorship among breast cancer patients.

An indicator of length of follow-up is the median follow-up time. While this could in theory be given as the median follow-up time of all patients, it is better calculated from follow-up among the individuals with censored data. However, both these methods tend to underestimate follow-up, and a more robust measure is based on the reverse KM estimator (Schemper and Smith, 1996), that is the KM method with the event indicator reversed so that the outcome of interest becomes being censored. In the ovarian cohort example, the median follow-up time of all the patients is 1.7 years, although is influenced by the survival times which were early deaths. The median survival of the censored patients was 3.2 years, but the reverse KM estimate of the median follow-up is 5.3 years (95% CI: 4.7–6.0 years).

Completeness of follow-up

Each patient who does not have an event can be included in a survival analysis for the period up to the time at which they are censored, but completeness of follow-up is still important. Unequal follow-up between different groups, such as treatment arms, may bias the analysis. A simple count of participants lost to follow-up is one indicator of data incompleteness, but it does not inform us about time lost and another measure has been proposed (Clark *et al*, 2002). In general, disparities in follow-up caused by differential drop-out between arms of a trial or different subgroups in a cohort study need to investigated.

Cohort effect on survival

In survival analysis, there is an assumption of homogeneity of treatment and other factors during the follow-up period. However, in a long-term observational study of patients of cancer, the case mix may change over the period of recruitment, or there may be an innovation in ancillary treatment. The KM method assumes that the survival probabilities are the same for subjects recruited early and late in the study. On average, subjects with longer survival times would have been diagnosed before those with shorter times, and changes in treatments, earlier diagnosis or some other change over time may lead to spurious results. The assumption may be tested, provided we have enough data to estimate survival probabilities in different subsets of the data and, if necessary, adjusted for by further analyses (see next section).

Between-centre differences

In a multicentre study, it is important that there is a consistency between the study methods in each centre. For example, diagnostic instruments, such as staging classification, and treatments should be identical. Heterogeneity in case mix among centres can be adjusted for in an analysis (see next section).

NEED FOR SURVIVAL ANALYSIS ADJUSTING FOR COVARIATES

When comparing treatments in terms of survival, it is often sensible to adjust for patient-related factors, known as covariates or confounders, which could potentially affect the survival time of a patient. For example, suppose that despite the treatment being randomised in the lung cancer trial, older patients were assigned more often to the radiotherapy alone group. This group would have a worse baseline prognosis and so the simple analysis may have underestimated its efficacy compared to the combination treatment, referred to as confounding between treatment and age. Also, we sometimes want to determine the prognostic ability of various factors on overall survival, as in the ovarian study. Figure 5 shows overall survival by FIGO stage, and there is a significant decrease in overall survival with more advanced disease.

Multiple prognostic factors can be adjusted for using multivariate modelling. For example, if those women with early stage disease were younger than those with advanced disease, then the FIGO I and II groups might be surviving longer because of lower age and not because of the effect of FIGO stage. In this case, the FIGO effect is confounded by the effect of age, and a multivariate analysis is required to adjust for the differences in the age distribution. The appropriate analysis is a form of multiple regression, and is the subject of the next paper in this series.

SUMMARY

Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is *time until an event occurs*. Because of censoring–the nonobservation of the event of interest after a period of follow-up–a proportion of the survival times of interest will often be unknown. It is assumed that those patients who are censored have the same survival prospects as those who continue to be followed, that is, the censoring is uninformative. Survival data are generally described and modelled in terms of two related functions, the survivor function and the hazard function. The survivor function represents the probability that an individual survives from the time of origin to some time beyond time *t*. It directly describes the survival experience of a study cohort, and is usually estimated by the KM method. The logrank test may be used to test for differences between survival curves for groups, such as treatment arms. The hazard function gives the instantaneous potential of having an event at a time, given survival up to that time. It is used primarily as a diagnostic tool or for specifying a mathematical model for survival analysis.

In comparing treatments or prognostic groups in terms of survival, it is often necessary to adjust for patient-related factors that could potentially affect the survival time of a patient. Failure to adjust for confounders may result in spurious effects. Multivariate survival analysis, a form of multiple regression, provides a way of doing this adjustment, and is the subject the next paper in this series.

ACKNOWLEDGEMENTS

We thank John Smyth for providing the ovarian cancer data, and Victoria Cornelius and Peter Sasieni for invaluable comments on an earlier manuscript. Cancer Research UK supported all the authors. Taane Clark holds a National Health Service (UK) Research Training Fellowship.

REFERENCES

Altman DG (2003) *Practical statistics for medical research*. London: Chapman & Hall

Altman DG, Bland JM (2003) Statistics notes: interaction revisited: the difference between two estimates. *BMJ* **326:** 219

Clark TG, Altman DG, De Stavola BL (2002) Quantifying the completeness of follow-up. *Lancet* **359:** 1309 – 1310

Clark TG, Stewart ME, Altman DG, Gabra H, Smyth J (2001) A prognostic model for ovarian cancer. *Br J Cancer* **85:** 944 – 952

Collett D (1994) *Modelling Survival Data in Medical Research*. London: Chapman & Hall

Cox DR (1979) A note on the graphical analysis of survival data. *Biometrika* **66:** 188 – 190

Hosmer DW, Lemeshow S (1999) *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* **53:** 457 – 481

Klembaum DG (1996) *Survival analysis*: A self learning text. New York: Springer

Lung Cancer Study Group (1988) The benefit of adjuvant treatment for resected locally advanced non-small cell lung cancer. *J Clin Oncol* **6:** 9 – 17

Parmer M, Machin D (1995) *Survival analysis*. UK: John Wiley and Sons Ltd

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* **35:** 1 – 39

Piantadosi S (1997) *Clinical Trials: A Methodologic Perspective*. New York: Wiley

Pocock S, Clayton TC, Altman DG (2002) Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet* **359:** 1686 – 1689

Ramlau-Hansen H (1983) Smoothing counting process intensities by means of kernel functions. *Ann Statist* **11:** 453 – 466

Schemper M, Smith TL (1996) A note on quantifying follow-up in studies of failure time. *Control Clin Trials* **17:** 343 – 346

British Journal of Cancer (2003) 89, 431 – 436
© 2003 Cancer Research UK All rights reserved 0007 – 0920/03 $25.00
www.bjcancer.com

Tutorial Paper

Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods

MJ Bradburn*^,1^, TG Clark[1], SB Love[1] and DG Altman[1]

[1] *Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 431 – 436. doi:10.1038/sj.bjc.6601119 www.bjcancer.com
© 2003 Cancer Research UK

Keywords: survival analysis; Cox model; AFT model; model selection

INTRODUCTION

Survival analysis involves the consideration of the time between a fixed starting point (e.g. diagnosis of cancer) and a terminating event (e.g. death). The key feature that distinguishes such data from other types is that the event will not necessarily have occurred in all individuals by the time the study ends, and for these patients, their full survival times are unknown. For instance, in studies that measure the length of survival after diagnosis of cancer, it is common for a proportion of individuals to remain alive and disease-free at the end of the follow-up period, and for these patients, we know only a lower limit on their actual time to event. Thus, special methods are required for these type of data. The explanation and demonstration of some of the methods proposed to analyse such data are the basis of this series.

In the first paper of this series (Clark *et al*, 2003), we described initial methods for analysing and summarising survival data including the definition of hazard and survival functions, and testing for a difference between two groups. We continue here by considering various statistical models and, in particular, how to estimate the effect of one or more factors that may predict survival.

THE NEED FOR MULTIVARIATE STATISTICAL MODELLING

The previous paper demonstrated the construction of (Kaplan–Meier) survival curves for different patient groups, and introduced the logrank test to investigate differences between them. Both these methods are examples of *univariate* analysis; they describe the survival with respect to the factor under investigation, but necessarily ignore the impact of any others. It is more common, at least in clinical investigations, to have a situation where several (known) quantities or *covariates*, potentially affect patient prognosis. For example, suppose two groups of patients are compared: those with and those without a specific genotype. If one of the groups also contains older individuals, any difference in survival may be attributable to genotype or age or indeed both. Hence, when investigating survival in relation to any one factor, it is often desirable to adjust for the impact of others. Moreover, while the logrank test provides a *P*-value for the differences between the groups, it offers no estimate of the actual effect size; in other words, it offers a statistical, but not a clinical, assessment of the factor's impact. The use of a statistical model improves on these methods by allowing survival to be assessed with respect to several factors simultaneously, and in addition, offers estimates of the strength of effect for each constituent factor. Therefore, statistical models are important and frequently used tools which, when constructed appropriately, offer valuable insight into the survival process.

Several statistical methods have been proposed for modelling survival analysis data. We will describe the most important models and illustrate their application using example datasets described in the previous paper (Clark *et al*, 2003). As before, we will assume throughout that all survival times are independent of each other and that censoring occurs solely as right-censoring and is uninformative. The focus is on covariates that are measured at the time of entry to the study, that may be continuous (e.g. the patient age or tumour size), binary (e.g. gender), unordered categorical (e.g. histology) or ordered categorical or ordinal (e.g. performance status or FIGO stage). In the next paper in this series, we will discuss the statistical assumptions made when using statistical models, and provide advice on choosing the appropriate model and covariates therein. We will also consider how to model covariates that change values over time (called 'time-dependent' or 'updated' covariates).

The methods we present here may be divided into two broad categories: proportional hazard approaches (including the semi-parametric Cox model and fully parametric approaches) and accelerated failure time models. These methods have different properties and interpretations, but all may be used to summarise survival data.

THE COX ('SEMI-PARAMETRIC') PROPORTIONAL HAZARDS MODEL

The Cox (proportional hazards or PH) model (Cox, 1972) is the most commonly used multivariate approach for analysing survival time data in medical research. It is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and a set of covariates. A fuller explanation of the hazard function was given in the previous article (Clark *et al*, 2003). Put briefly, the hazard is the instantaneous event probability at a given time, or the probability

*Correspondence: Mr M Bradburn; E-mail: mike.bradburn@cancer.org.uk

Received 6 December 2002; accepted 30 April 2003

that an individual under observation experiences the event in a period centred around that point in time.

Mathematically, the Cox model is written as

$$h(t) = h_0(t) \times \exp\{b_1x_1 + b_2x_2 + \cdots + b_px_p\}$$

where the hazard function $h(t)$ is dependent on (or determined by) a set of p covariates (x_1, x_2, …, x_p), whose impact is measured by the size of the respective *coefficients* (b_1, b_2, …, b_p). The term h_0 is called the baseline hazard, and is the value of the hazard if all the x_i are equal to zero (the quantity exp(0) equals 1). The 't' in $h(t)$ reminds us that the hazard may (and probably will) vary over time. An appealing feature of the Cox model is that the baseline hazard function is estimated nonparametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution.

The Cox model is essentially a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an 'intercept' term that varies with time. The covariates then act multiplicatively on the hazard at any point in time, and this provides us with the key assumption of the PH model: the hazard of the event in any group is a constant multiple of the hazard in any other. This assumption implies that the hazard curves for the groups should be proportional and cannot cross (see Figure 1 for examples of each). Proportionality implies that the quantities exp(b_i) are called *hazard ratios*. A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases. Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival. This *proportionality assumption* is often appropriate for survival time data but it is important to verify that it holds. We discuss methods for assessing proportionality in the next paper in this series.

The Cox PH model fitted to the ovarian cancer data

This large database, as described in the previous paper of this series (Clark *et al*, 2003), was used to derive a prognostic index for overall survival among ovarian cancer patients in Clark *et al* (2001). Their analysis included 10 variables, but for simplicity we will consider five, all of which were measured at diagnosis: FIGO stage (an ordinal covariate taking values of 1, 2 3 or 4), histology (one of seven subtypes), grade (1, 2 or 3), ascites (yes/no) and patient age.

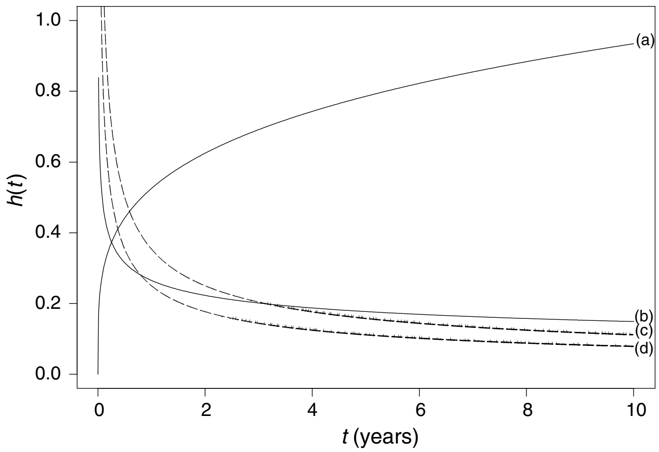


Figure 1 Example of (non-) proportional hazards (groups (c) and (d) only have proportional hazards) using the Weibull distribution. For the Weibull survival model, the hazard function $h(t) = \lambda s(\lambda t)^{s-1}$ for λ, $s > 0$: (a) increasing hazard ($\lambda = 0.5$, $s = 1.25$); (b) decreasing hazard ($\lambda = 0.25$, $s = 0.75$); (c) decreasing hazard ($\lambda = 0.5$, $s = 0.5$); (d) decreasing hazard ($\lambda = 0.25$, $s = 0.5$).

Table 1 shows the effect sizes (given as hazard ratios), 95% confidence intervals (CI), regression coefficients and statistical significance for each of these in relation to overall survival. Each factor is assessed through separate univariate Cox regressions (left-hand columns). However, the aim of the database is to describe how the factors jointly impact on survival, and so all five factors were incorporated into the multivariate model (right-hand columns). It may be seen that higher FIGO stage, higher grade, presence of ascites and increased age impaired survival to varying (and statistically significant) degrees. The histology was also of importance: the figures describe the survival of patients with each histology type in comparison with the serous type. In principle, any type with a reasonable number of patients could be chosen as the baseline of comparison. On multivariate analysis Mucinous and serous were the tumour types with the best prognosis, whereas undifferentiated and mixed mesodermal were the worst. It is possible to present P-values for the comparisons between each type and serous, but we have given an overall likelihood ratio test for the differences between the categories as a whole. The FIGO stage could be modelled as a categorical variable in the same manner as grade and histology, but assuming it is a continuous variable with a linear trend across the four categories performed sufficiently well.

PARAMETRIC PH MODELS

Parametric PH models are a class of models similar in concept and interpretation to the Cox (PH) model. The key difference between the two is that the hazard is assumed to follow a specific statistical distribution when a fully parametric PH model is fitted to the data, whereas the Cox model enforces no such constraint. Other than this, the two model types are equivalent. Hazard ratios have the same interpretation, whether derived from a Cox or a fully parametric regression model, and the proportionality of hazards is still assumed.

A number of different parametric PH models may be derived by choosing different hazard functions. As shown previously, there is a direct link between the survival and hazard, and the choice of hazard distribution determines that of the survival. In fact, the models commonly applied, such as the *Exponential*, *Weibull* or *Gompertz* models, take their names from the distribution that the survival times are assumed to follow, but the most distinguishing features between them are in the hazard function. Examples of survival and hazard functions derived from some of these parametric models were presented in the previous paper (Clark *et al*, 2003). Figure 1 shows increasing and decreasing Weibull hazard functions, as well as two groups with the latter that are proportional to each other.

Parametric models fitted to the ovarian cancer data

The estimated hazard function of the ovarian cancer data as displayed in the previous paper (Clark *et al*, 2003) may be consistent with that derived from a Weibull PH model with decreasing hazard. Fitting this to the ovarian cancer database gives similar results as the Cox model (see Table 2), and may be interpreted in the same manner. Methods to check for the appropriateness of the Weibull distribution will be discussed in the next paper of this series.

COMPARISON OF THE TWO PH APPROACHES

The main drawback of parametric models is the need to specify the distribution that most appropriately mirrors that of the actual

Table 1 Hazard ratios from the Cox PH model for the ovarian dataset

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.809	2.24	(2.03–2.48)	<0.001	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.727	0.48	(0.38–0.61)		−0.422	0.66	(0.50–0.85)	
Endometroid	−1.162	0.31	(0.22–0.45)		0.198	1.22	(0.80–1.85)	
Clear cell	−0.343	0.71	(0.52–0.97)		0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.119	1.13	(0.74–1.72)		0.501	1.65	(0.91–2.99)	
Undifferentiated	0.390	1.48	(0.81–2.70)		0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.614	1.85	(1.28–2.66)		0.789	2.20	(1.45–3.35)	
Grade				<0.001				<0.001
1	(0.000)	(1.00)			(0.000)	(1.00)		
2	1.116	3.05	(1.90–4.91)		0.885	2.42	(1.40–4.19)	
3	1.650	5.20	(3.31–8.18)		0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.798	0.45	(0.37–0.55)	<0.001	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.153	1.17	(1.12–1.21)	<0.001	0.133	1.14	(1.09–1.19)	<0.001

HR = hazard ratio, CI = confidence interval.

Table 2 Hazard ratios from the Weibull PH model for the ovarian dataset

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.862	2.37	(2.14–2.62)	<0.001	0.768	2.16	(1.89–2.46)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.804	0.45	(0.35–0.57)		−0.496	0.61	(0.47–0.79)	
Endometroid	−1.276	0.28	(0.20–0.40)		0.120	1.13	(0.75–1.70)	
Clear cell	−0.419	0.66	(0.48–0.90)		0.346	1.41	(0.99–2.02)	
Adenocarcinoma	0.113	1.12	(0.73–1.71)		0.499	1.65	(0.91–2.97)	
Undifferentiated	0.397	1.49	(0.82–2.71)		0.765	2.15	(1.06–4.37)	
Mixed Mesodermal	0.638	1.89	(1.31–2.73)		0.804	2.23	(1.47–3.40)	
Grade								<0.001
1	(0.000)	(1.00)		<0.001	(0.000)	(1.00)		
2	1.154	3.17	(1.97–5.10)		0.928	2.53	(1.47–4.36)	
3	1.727	5.62	(3.58–8.84)		0.895	2.45	(1.43–4.20)	
Absence of ascites	−0.840	0.43	(0.36–0.52)	<0.001	−0.404	0.67	(0.54–0.83)	<0.001
Age (per 5-year increase)	0.165	1.18	(1.14–1.22)	<0.001	0.138	1.15	(1.10–1.20)	<0.001

HR = hazard ratio, CI = confidence interval.

survival times. This is an important requirement that needs to be verified and an appropriate distribution may be difficult to identify. Where a suitable distribution can be found, however, the parametric model is more informative than the Cox model. It is straightforward to derive the hazard function and to obtain predicted survival times when using a parametric model, which is not the case in the Cox framework (the use of such quantities is discussed in the next section). Additionally, the appropriate use of these models offers the advantage of being slightly more *efficient*; they yield more precise estimates (i.e. smaller standard errors).

The results from the Cox or parametric PH models may be compared directly, as the model types are merely different approaches to assessing the same quantity. For either method to be valid: (a) the covariate effect needs to be at least approximately constant throughout the duration of the study, and (b) the proportionality assumption must hold. These important issues will be addressed in the subsequent paper in this series.

INTERPRETING THE PH MODEL: BEYOND THE HAZARD RATIO

In addition to the ratio of two hazards, it is possible to obtain other information from a PH regression model. One simple (and possibly underused) quantity that may be derived from a survival model is the predicted survival proportion at any given point in time for a particular risk group. The survival proportion for a given risk group at any time, $S(t)$, is equal to

$$S(t) = S_0(t)^{\exp(\gamma)}$$

where $S_0(t)$ is the baseline survival (the survival proportion when all covariates are equal to zero) and γ is equal to $b_1x_1 + b_2x_2 + \cdots + b_px_p$. Once the value of the baseline survival at a given time is derived, then the predicted survival probabilities for patients with any specified covariate values x_i are easily obtained. This information could then be displayed via tabular or

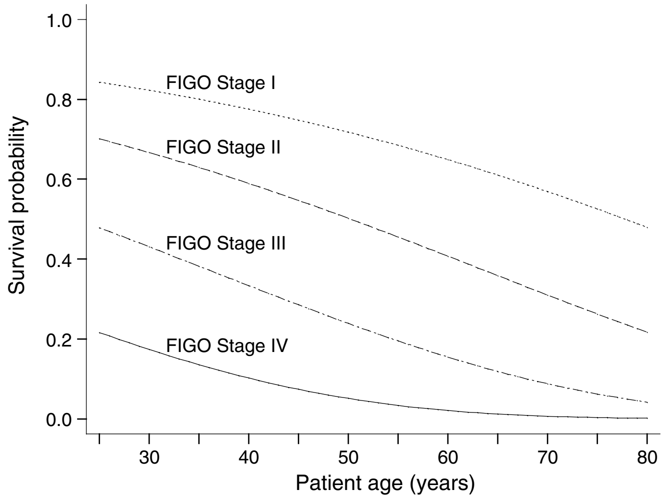


Figure 2 Predicted 5-year survival of ovarian cancer patients by age and FIGO stage.

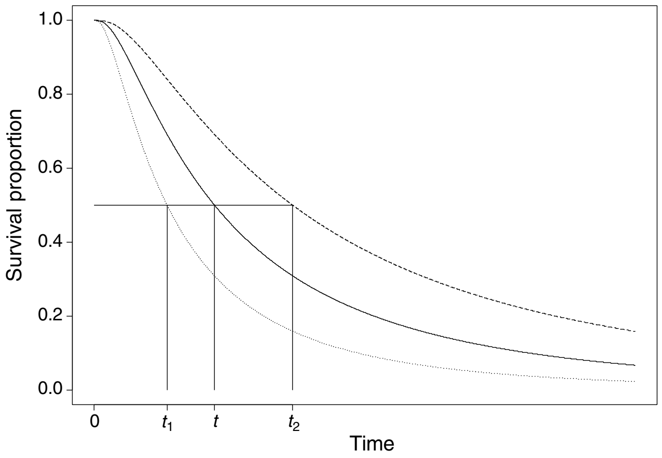


Figure 3 Illustration of the AFT model: (——), $S_0(t)$ the baseline survival function; (·······), $S(t_1) = S_0(\varphi t)$ for $\varphi < 1$; (------), $S(t_2) = S(\varphi t)$ for $\varphi > 1$.

graphical displays. Figure 2 illustrates this by giving predicted 5-year survival according to patient age and FIGO stage. Further examples are demonstrated by Christensen (1987) based on the Cox model, but can also be used when fitting fully parametric models. In a previous analysis that involved some of the patients in the present data, Clark *et al* (2001) produced a nomogram to summarise the impact of these and other covariates, and thus allows the reader to predict the median survival and the 2- and 5-year survival probabilities for patients with given prognostic information.

The advantage of fitting a parametric survival model is that predictions of the event survival, event hazard, mean and median survival times are readily available. For FIGO stages I–IV, the median survival times are estimated to be 7.8, 4.0, 2.0 and 1.0 years, respectively.

ACCELERATED FAILURE TIME MODELS

The accelerated failure time (AFT) model is a different type of model that may be used for the analysis of survival time data. For a group of patients with covariates (x_1, x_2, … x_p), the model is written mathematically as

$$S(t) = S_0(\varphi t)$$

where $S_0(t)$ is the baseline survivor function and φ is an 'acceleration factor' that depends on the covariates according to the formula

$$\varphi = \exp\{(b_1x_1 + b_2x_2 + \cdots + b_px_p)\}.$$

The principle here is that the effect of a covariate is to stretch or shrink the survival curve along the time axis by a constant relative amount φ. Figure 3 demonstrates this for the case of a single covariate (x_1) with two levels, for example, $x_1 = 0$ for a placebo group and $x_1 = 1$ for a new treatment group. The survival probabilities, $S(t)$, for the placebo and new treatment groups are $S_0(t)$ and $S_0(\varphi t)$, respectively. The proportion of patients who are event-free in the placebo group at any time point t_1 is the same as the proportion of those who are event-free in the new treatment group at a time $t_2 = \varphi t_1$. Figure 3 shows the cases where $\varphi > 1$ and $\varphi < 1$, which represent situations where the length of survival is increased and decreased in the new treatment group compared with the placebo, respectively.

The AFT model is commonly rewritten as being log-linear with respect to time, giving

$$\log(T) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon$$

where ε is a measure of (residual) variability in the survival times. Thus, the survival times can be seen to be multiplied by a constant effect under this model specification, and the exponentiated coefficients, $\exp(b_i)$, are referred to as *time ratios*. A time ratio above 1 for the covariate implies that this 'slows down', or prolongs the time to the event, while a time ratio below 1 indicates that an earlier event is more likely.

When the survival times follow a Weibull distribution, it can be shown that the AFT and PH models are the same. However, the AFT family of models differs crucially from the PH model types in terms of their interpretation of effect sizes as time ratios as opposed to hazard ratios.

The survival times are usually assumed to follow a specific distributional form in the AFT framework. Distributions such as the *Log-Normal, Log-Logistic, Generalised Gamma* and *Weibull* may be used to represent such survival data. Alternative methods include the method of Buckley and James (1979), which is discussed by Stare *et al* (2000), and semiparametric AFT models, in which the baseline survivor function is estimated nonparametrically (see Wei, 1992, for an overview), but have not yet been widely implemented in statistical software.

As with the PH approach, other quantities such as projected survival probabilities may be derived. Also in keeping with PH models is the fact that AFT models make assumptions; the appropriate choice of statistical distribution needs to be made, and also the covariate effects are assumed to be constant and multiplicative on the timescale, that is, that the covariate impacts on survival by a constant factor.

Parametric AFT models fitted to the lung cancer trial data

We use the non-small cell lung cancer dataset to illustrate the AFT model, focusing on the relapse-free survival (i.e. , the time from diagnosis to the reappearance of cancer, with patients censored at time of death if no recurrence had appeared). Again, we present both the univariate and multivariate effect sizes in Table 3. The specific comparison of interest was the effect of adjuvant (platinum-based) chemotherapy and radiotherapy compared with radiotherapy alone. The unadjusted treatment effect may be

Table 3 Time ratios from the generalised gamma AFT model for the lung cancer trial

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**TR exp(b_i)**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**TR exp(b_i)**	**95% CI**	***P*-value**
Treatment (RT+CAP *vs* RT alone)	0.648	1.91	(1.21–3.01)	0.005	0.718	2.05	(1.29–3.23)	0.002
Cell type (Sq *vs* non-Sq)	0.506	1.66	(1.01–2.71)	0.04	0.511	1.67	(1.04–2.68)	0.03
Performance status (8–10 *vs* 5–7)	0.767	2.15	(1.11–4.19)	0.02	0.729	2.07	(1.00–4.29)	0.05
Tumour status				0.59				0.60
1	(0.000)	(1.00)	—		(0.000)	(1.00)	—	
2	−0.189	0.83	(0.40–1.70)		−0.353	0.70	(0.35–1.41)	
3	−0.388	0.68	(0.31–1.48)		−0.378	0.69	(0.31–1.53)	
Nodal involvement				0.87				0.97
None	(0.000)	(1.00)	—		(0.000)	(1.00)	—	
Limited	0.122	1.13	(0.46–2.79)		−0.059	0.94	(0.36–2.48)	
Extensive	0.206	1.23	(0.55–2.73)		0.029	1.03	(0.42–2.54)	
Age at diagnosis (/years)	−0.013	0.99	(0.96–1.01)	0.34	−0.011	0.99	(0.96–1.01)	0.41
Gender (male *vs* female)	0.032	1.03	(0.62–1.71)	0.90	−0.007	0.99	(0.59–1.67)	0.98
Weight loss ($\geqslant$10 *vs* <10%)	−0.477	0.62	(0.29–1.33)	0.22	−0.337	0.71	(0.34–1.51)	0.38
Race (white *vs* non-white)	0.440	1.55	(0.81–2.98)	0.19	0.202	1.22	(0.61–2.46)	0.57

TR = time ratio, CI = confidence interval, RT = radiotherapy, CAP = cytoxan, doxorubicin and platinum-based chemotherapy, Sq = squamous.

summarised by a time ratio of 1.91 (95% CI: 1.21–3.01; $P = 0.005$), which, having allowed for other covariates increased slightly to 2.05. Therefore, we can conclude that the time to recurrence was significantly prolonged (approximately doubled) among patents given adjuvant chemotherapy in comparison with those who were not.

Again, we can derive model-based predictions: overall, patients allocated to receive adjuvant chemotherapy had a predicted median survival time of approximately 16 months, as opposed to 8 months among those treated with radiotherapy alone. Other factors are also significant and would influence these times, but these are of less importance in the context of the comparative trial. We will return to this example in the next paper of this series.

WHICH MODEL SHOULD WE USE: PH OR AFT?

From a statistical viewpoint, an obvious way to choose between the two model types is to fit a type that is in keeping with the data. If the AFT model clearly fits the data better than the PH model, or *vice versa*, this model may be adopted as being the more appropriate. However, in some cases, either type of model may appear to fit the data adequately. In such instances, the choice of model may be influenced by other factors. For instance, if other studies of a similar nature had all used the Cox regression and reported the results as hazard ratios, one may be tempted to follow suit to aid comparability. Against this, the parametric approach offers more in the way of predictions, and the AFT formulation allows the derivation of a time ratio, which is arguably more interpretable than a ratio of two hazards. As yet, however, AFT models are relatively unfamiliar and seen rarely in medical research papers (see Kay and Kinnersley, 2002).

OTHER APPROACHES

Stratified survival analysis

A more straightforward way to incorporate covariates into a survival analysis is to use a stratified survival analysis. For example, suppose the covariate of primary interest is treatment, but we wish to control for the clinical stage of the tumour when making the comparison. Here, the survival in each treatment group can be compared within each stage of disease (the 'strata') by the logrank or some other method, and the differences within each stratum are then combined to give an overall comparison of treatments that has been adjusted for the stage.

The strength of this method is in its simplicity: as the logrank test is nonparametric, few distributional assumptions are made, and its interpretation is straightforward. Its main limitation is that it is only applicable when the covariate is categorical (or with continuous variables that have been arbitrarily categorised). Further, this method does not perform well with several covariates, as the number of individuals in each stratum quickly becomes too small to allow reasonable comparisons. In addition, it does not quantify the strength of effect of each variable, or even offer a *P*-value for factors other than the one of primary interest. This method is not generally regarded as a formal statistical model, but is of use where a very small number of covariates are to be considered, if only as an exploratory method of analysis.

Aalen's additive model

Another approach to modelling the relationship between survival and covariates is to assume that the covariates act additively on the hazard. Aalen's additive hazard model (Aalen, 1989) is one method that has been suggested for this, but its properties are rather unlike any other model described in this paper. The covariates are assumed to impact additively upon a (unknown) baseline hazard, but the effects are not constrained to be constant. The impact is therefore allowed to vary freely over time according to the underlying equation

$$h(t) = h_0(t) + b_1(t)x_1 + b_2(t)x_2 + \cdots + b_p(t)x_p$$

where $h(t)$ is the hazard, $h_0(t)$ is the baseline hazard and the $b_i(t)$ are coefficients, which may change in magnitude and even sign with time. Compare this with the Cox regression, where $h_0(t)$ is also estimated nonparametrically, but the b_i quantify the *multiplicative* effect of covariate i on the hazard and are assumed constant at all times.

As it is not straightforward to estimate $h_0(t)$ nonparametrically, the cumulative baseline hazard is used and the regression coefficients that are actually estimated from the data are also the cumulative (additional) hazard

$$B_i(t) = \int_0^t b_i(u)\mathrm{d}u$$

The usual method of representing these effects is to graph them against time. The further $B_i(t)$ is from zero at time t, the greater the effect the covariate has had on the hazard over the course of the study up to t. The values of $b_i(t)$, the absolute increase in hazard at time t, are not actually observed, but their relative size may be inferred from the slope of the line. These plots are sometimes called Aalen plots, and they are also used to provide an informal assessment of the adequacy of the proportional hazards assumption in the Cox model, although Aalen considered its primary role as an alternative model in its own right (Aalen, 1993).

The flexibility of this approach is tempered by the lack of an easy interpretation. The $B_i(t)$ coefficients are not easy to understand, and as they change repeatedly over time, can offer no single quantifiable effect size. Formal tests of statistically significant covariate effects may be carried out, but Aalen plots are essentially the only manner with which to interpret the effect sizes. These reasons, together with the relative lack of statistical software, are probably the deciding factors in the relatively minimal use of Aalen's model.

Classification trees and artificial neural networks

Two relatively recent developments are classification trees and artificial neural networks. These methods differ substantially in their complexity and interpretation to the methods presented here and to each other. Both approaches are described in more detail in a later paper of this series.

DISCUSSION

The principal strength of statistical models is their ability to assess several covariates simultaneously. The strengths of the stratified logrank test and other such methods are their obvious simplicity and the fact that they make fewer parametric assumptions of the data. Although these reasons are usually insufficient to suggest that the stratified method be used more widely, this second feature is a relevant one, because it needs to be kept in mind that all the models introduced here make certain distributional assumptions of the survival times that will not always be met.

We have focused on the Cox model, the class of parametric PH models and AFT models as tools with which to analyse survival time data. Other models exist (see, e.g., Collett (1994) for a more practical demonstration of some alternatives and Bagdonavičius and Nikulin (2001) for the theoretical background), but many are similar to, if not extensions of, the approaches we have discussed. The use of the Cox model offers greater flexibility than parametric alternatives and, in particular, does not require the direct estimation of the baseline hazard function (i.e. it avoids the need to specify the distribution of the survival times). However, the assumption of proportional hazards is a crucial one that needs to be fulfilled for the results to be meaningful, and will not always be satisfied. Further, while the Cox PH model may be valid, other parametric models will produce more precise estimates where the distribution is specified correctly.

A further concern is that the choice of covariates to include is also far from simple. In the third paper of this series, we will consider ways to choose between the various model types, to identify and assess the importance of covariates, and to verify that the final model is adequate.

ACKNOWLEDGEMENTS

We wish to thank John Smyth for providing the ovarian cancer data, and Victoria Cornelius and Peter Sasieni for invaluable comments on an earlier manuscript. Cancer Research UK supported all authors. Taane Clark holds a National Health Service (UK) Research Training Fellowship.

REFERENCES

Aalen OO (1989) A linear regression model for the analysis of life times. *Stat Med* **8:** 907–925

Aalen OO (1993) Further results on the non-parametric linear regression model in survival analysis. *Stat Med* **12:** 1569–1588

Bagdonavičius V, Nikulin M (2001) *Accelerated Life Models: Modeling and Statistical Analysis*. London: Chapman & Hall/CRC

Buckley K, James I (1979) Linear regression with censored data. *Biometrics* **66:** 429–436

Christensen E (1987) Multivariate survival analysis using Cox's regression model. *Hepatology* **7:** 1346–1358

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis. Part I: basic concepts and first analyses. *Br J Cancer* **89:** 232–238

Clark TG, Stewart ME, Altman DG, Gabra H, Smyth J (2001) A prognostic model for ovarian cancer. *Br J Cancer* **85:** 944–952

Collett D (1994) *Modelling Survival Data in Medical Research*. London: Chapman and Hall/CRC

Cox DR (1972) Regression models and life tables (with discussion). *J R Statist Soc B* **34:** 187–220

Kay R, Kinnersley N (2002) On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug Inf J* **36:** 571–579

Stare J, Heinzl H, Harrell F (2000) On the use of Buckley and James least squares regression for survival data. New approaches in applied statistics: Metodološki zvezki 16 (http://mrvar.fdv.uni-lj.si/pub/mz/mz16/stare.pdf)

Wei LJ (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* **11:** 1871–1879

British Journal of Cancer (2003) 89, 605–611
© 2003 Cancer Research UK All rights reserved 0007–0920/03 $25.00
www.bjcancer.com

Tutorial Paper

Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit

MJ Bradburn*,[1], TG Clark[1], SB Love[1] and DG Altman[1]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 605–611. doi:10.1038/sj.bjc.6601120 www.bjcancer.com
© 2003 Cancer Research UK

Keywords: survival analysis; Cox model; AFT model; model checking; choice of coavriates; goodness of fit

INTRODUCTION

In this series of papers, we have described a selection of statistical methods used for the initial analysis of survival time data (Clark *et al*, 2003), and introduced a selection of more advanced methods to deal with the situation where several factors impact on the survival process (Bradburn *et al*, 2003). The latter paper focused on proportional hazards (PH) and accelerated failure time (AFT) models, and we continue the series by demonstrating the application of these models in more detail. Whereas the focus of the previous paper was to outline the purpose and interpretation of statistical models for survival analysis, we concentrate here on approaches with which to undertake the actual modelling process. In other words, the aim of this paper is to promote the correct use of the models that have been suggested for the analysis of survival data.

When used inappropriately, statistical models may give rise to misleading conclusions. Checking that a given model is an appropriate representation of the data is therefore an important step. Unfortunately, this is a complicated exercise, and one that has formed the subject of entire books. Here, we aim to present an overview of some of the major issues involved, and to provide general guidance when developing and applying a statistical model. We start by presenting approaches that can be used to ensure that the correct factors have been chosen. Following this, we describe some approaches that will help decide whether the statistical model adequately reflects the survivor patterns observed. Lastly, we describe methods to establish the validity of any assumptions the modelling process makes. We will illustrate each using the two example datasets (a lung cancer trial and an ovarian cancer dataset) that were introduced in the previous papers (Bradburn *et al*, 2003; Clark *et al*, 2003).

CHOICE OF COVARIATES

The covariates that we consider here are fixed, that is, known at baseline or entry to the study. The handling of covariates that change values over time (e.g. white blood cell count as measured at different time points) will be described in the subsequent paper in this series.

Sample size considerations

It is implicitly assumed that the subjects in a study are representative of a wider population to enable the study aims to be addressed. Another important requirement is to have data from an adequate number of subjects. Any estimate based on a small number of individuals will be less reliable than one based on a larger number, and when multivariate models are fitted to small datasets, the estimated impact of the covariates is too imprecise to give reliable answers. The use of variable selection procedures as described below is especially problematic with such data, and often leads to overoptimistic results. Finally, smaller data sets may not have sufficient power to detect a covariate that has a significant impact on survival.

The power (and indeed in some cases validity) of a survival analysis is related to the number of events rather than the number of participants. Simulation work has suggested that at least 10 events need to be observed for each covariate considered, and anything less will lead to problems, for example, the regression coefficients become biased (Peduzzi *et al*, 1995). In the ovarian study, there were 550 deaths and 11 covariates for the five prognostic factors, implying 50 events per covariate. In the liver cancer trial with 114 events, a full model of 11 covariates has approximately 10 events per covariate.

For prospective studies, several books (e.g. Machin *et al*, 1998) and software packages (e.g. nQuery, power and precision) are available to assist the calculation of adequate sample sizes, and many general purpose statistical packages also perform such calculations.

The aim of the study influences the choice of covariates

Before embarking on any statistical modelling, it is helpful to be clear as to why the multivariate model is to be fitted. The models we have presented have the considerable advantage of being able to handle several factors simultaneously, but the choice of which to incorporate lies with the analyst. This choice depends on the study aims. We suggest three possible scenarios as to why a study may use a multivariate model, and deal with each in turn.

(a) A single factor is under investigation for its association with survival, but several other factors exist
The rationale of such a study is to perform a specific test of one factor. This scenario may arise in a randomised controlled trial, such as the lung cancer example, where the aim is to decide

*Correspondence: Mr M Bradburn; E-mail: mike.bradburn@cancer.org.uk

Received 6 December 2002; accepted 30 April 2003

whether a new treatment prolongs survival, but also to adjust for prognostic factors that may or may not be equally matched between treatment groups. Another situation occurs where an association between a marker and patient survival is being assessed. In either case, any terms that are of potential importance could be incorporated whether significant or not, depending on the adequacy of the sample size. All of the covariates (other than the one of primary interest) are essentially 'nuisance' factors that are considered only to ensure they have been taken due account for in assessing the importance of the (prespecified) factor under investigation. Less important covariates may be removed.

(b) A collection of factors of known relevance are under investigation for their ability to predict survival
This arises when one wishes to assess the individual importance of a series of factors, and/or to attempt to build a model that helps predict patient survival. In such cases, the simplest strategy is to attempt to model all covariates, obtain effect sizes and gauge how well the model predicts survival. It may be desirable to remove factors from the model for simplicity, provided this does not compromise the predictive ability of the model. Statistical significance alone is an insufficient measure of assessing the extent to which a covariate can predict survival. Methods that may be used for this evaluation are given in the final paper of this series.

(c) Where a collection of factors are under investigation for their potential association with survival, possibly with additional known factors
Such studies are more 'exploratory' in nature, and the aim is to identify quantities of potential importance for further investigation. Here it is often desired to reduce the number of covariates in the model by excluding those that are not statistically significant and thus concentrate only on 'potentially interesting' ones for future research. Care must be exercised when several covariates are investigated, as the false-positive rate (or the chance of finding a spurious effect) increases with each additional test.

This selection of scenarios is far from exhaustive, and in practice a study may combine all of the above types. The ovarian study is a combination of (b) and (c).

Approaches to adding or removing covariates

Common choices for model building focus on 'semiautomated' methods such as stepwise selection, but other approaches exist. Models that are based purely on statistical significance may not be clinically meaningful. Henderson and Velleman (1981) state this simply: 'The data analyst knows more than the computer', and appropriate use of this knowledge should be incorporated into the analysis. We recommend that the choice of covariates should be verified by a degree of hands-on modelling, where terms are added or removed in a logical order rather than solely according to statistical significance.

We illustrate some straightforward approaches to the choice of covariates in the two example datasets used in this paper. In the final paper of this series, we will outline the rationale behind semiautomated methods (together with their limitations) and give further advice on hands-on modelling.

Selecting covariates for the lung cancer trial

As stated before, the lung cancer trial as presented in the earlier paper is an example of scenario (a). The table of coefficients for the full AFT multivariate model was presented in the previous paper (Bradburn *et al*, 2003). A simpler model would be to consider just the performance status, cell type and treatment covariates. Removing the remaining covariates reduces the model likelihood, but not to a significant degree ($\chi^2 = 3.34$ on 8 degrees of freedom; $P = 0.91$). The new time ratios, confidence intervals and *P*-values are presented in Table 1. They are virtually unchanged from the previous analysis, and thus the earlier conclusions remain the same.

Selecting covariates for the ovarian cancer database

As stated previously, the analysis of the ovarian cancer database (as described in the previous paper) could be considered as a mixture of scenarios (b) and (c). However, as the database is large and the aim is to derive a prognostic model, we will focus on (b). We consider five covariates here, all of which were measured at diagnosis: FIGO stage (an ordinal covariate taking values 1, 2, 3 or 4), histology (with seven possible subtypes), grade, ascites (yes/no) and patient age.

In this analysis, all the covariates were included yielding the model presented in Table 2. Advanced FIGO stage, higher grade, presence of ascites and increased age all impaired survival to varying degrees. The mucinous and serous histology types had a better prognosis, and undifferentiated and mixed mesodermal a lesser one. No grade–histology interactions were included in the final model, either due to insufficient numbers of patients to allow meaningful modelling (e.g. clear cell, mixed mesodermal, adenocarcinoma or undifferentiated), or for statistical insignificance (the remainder). In fact no second-order interaction or nonlinearity was detected.

If this model were to be used for the purpose of predicting future survival patterns, it is appropriate to ensure that the effect sizes are robust. One approach is to use bootstrap sampling, which involves randomly resampling the data and fitting the model to these modified datasets (Clark and Altman, 2002). These produce a series of effect sizes that should be similar to those derived from the original data if the model is sufficiently stable, and indeed do so here.

ASSESSING THE ADEQUACY OF A MODEL

Regardless of which type of model is fitted and how the variables are selected to be in the model, it is important to evaluate how well the model represents the data. A survival model is adequate if it

Table 1 Generalised gamma AFT model applied to the lung cancer data

Covariate	Coefficient (b_i)	TR [exp(b_i)]	95% CI	*P*-value
Treatment (RT+CAP *vs* RT alone)	0.640	1.90	(1.23–2.93)	0.004
Cell type (squamous *vs* nonsquamous)	0.536	1.71	(1.08–2.71)	0.02
Performance status (8–10 *vs* 5–7)	0.765	2.15	(1.09–4.24)	0.03

TR = time ratio; CI = confidence Interval; RT = radiotherapy; CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

Table 2 Cox model applied to the ovarian data

Covariate	Coefficient (b_i)	HR [exp(b_i)]	95% CI	*P*-value
FIGO stage	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001
Serous	(0.000)	(1.00)		
Mucinous	−0.422	0.66	(0.50–0.85)	
Endometroid	0.198	1.22	(0.80–1.85)	
Clear cell	0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.501	1.65	(0.91–2.99)	
Undifferentiated	0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.789	2.20	(1.45–3.35)	
Grade				<0.001
1	(0.000)	(1.00)		
2	0.885	2.42	(1.40–4.19)	
3	0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.133	1.14	(1.09–1.19)	<0.001

HR = hazard ratio; CI = confidence interval.

Table 3 Suggested plots for residual-based diagnostics

Y-axis	*X*-axis	Potential implication	Suggested remedy
Martingale residual	Any omitted covariate	Covariate excluded wrongly	Refit model with covariate included
Martingale residual	Any included covariate	Covariate modelled incorrectly (e.g. nonlinear effect)	Fit nonlinear term (e.g. a squared term)
Martingale residual	Date of enrolment in study	Evidence of temporal effect	Incorporate time of entry as a covariate
Deviance residual	Survival time, log(survival time) or ranks of survival time	Model fails to predict consistently for all survival times	Fit time-dependent PH model or consider using a different (i.e. non-PH) model
Deviance residual	Subject identifier	Individual is an outlier	(1) Check if the data are correct (2) Refit model with individual removed. If effect sizes alter substantially, consider removing individual altogether
Scaled Schoenfeld residual	Survival time, log(survival time) or ranks of survival time	Non-PH	Fit time-dependent PH model or consider using a different (i.e. non-PH) model

PH = proportional hazards. All of the above *X*–*Y* plots should give rise to a plot evenly scattered along a horizontal line that displays no trend. The possible implications where this does not occur and suggested remedies are presented.

represents the survival patterns in the data to an acceptable degree. This aspect of a model is known as *goodness of fit*. For example, if a given group of patients have a poor (or good) prognosis, then the model should predict this group to have that outcome. In practice, the issues in choosing the most appropriate type of model and the most appropriate covariates are heavily related, and the adequacy of a model may be assessed in several ways. In this section, we discuss methods to verify fit that are common across all survival models, before describing approaches specific to different model types. We will use the ovarian database example to demonstrate these checks.

Residuals from survival models

Residuals are a useful method for checking the fit of a statistical model. Essentially, they are the difference between an observed and a model-predicted quantity, with large or systematic differences between the two indicative of a poor model. Several residuals have been proposed, but unfortunately most are rather difficult to understand in the context of survival analyses due to censoring (Collett, 1994). In general, the residuals are skewed and need to have smoothing functions (e.g. Kernel smoother) applied to aid interpretation. Nevertheless, the graphical displays suggested in Table 3 (with appropriate smoothing as required) should all give rise to an evenly scattered horizontal band and display no obvious trend (e.g. no slope). If a trend in these plots is apparent, it should be investigated, perhaps using the method suggested in Table 3. Overall model adequacy may be assessed by use of Cox-Snell residuals (Collett, 1994).

Residual plots for the ovarian cancer data set

Figures 1A illustrates a plot of the Martingale residuals against the patient's age, with a Kernel smoother marked as the dashed line. Figure 1B shows the Martingale residuals plotted against FIGO stage, with the median within each stage represented by the solid bar. Both FIGO and age were modelled as linear effects. If FIGO or age had been wrongly excluded or modelled incorrectly (i.e. nonlinear), the figures should display a trend other than a strictly horizontal line. The age residual plot shows no evidence of a trend. Although there appears to be evidence of a trend in the FIGO plot, the inclusion of this covariate as a categorical covariate fails to

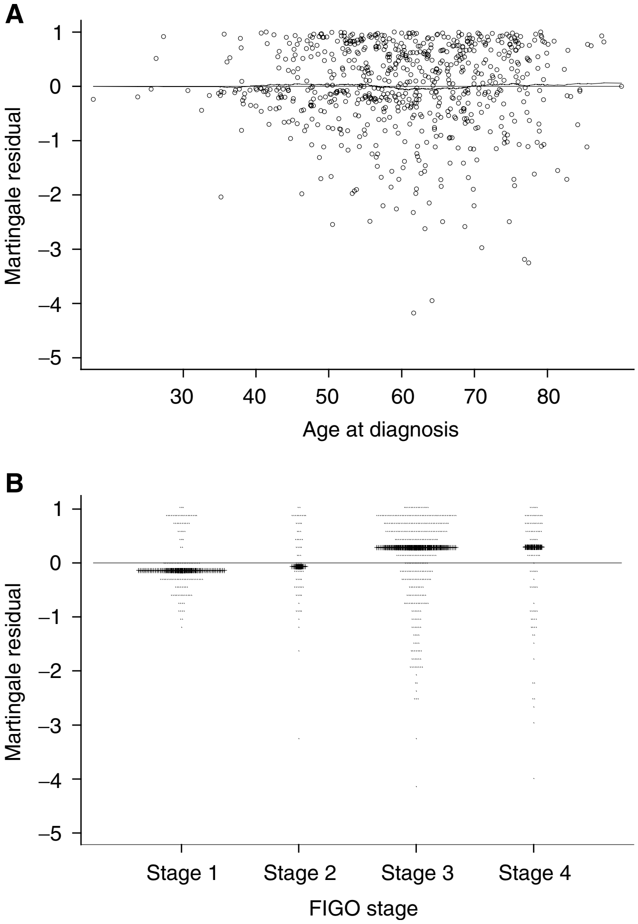


Figure 1 Martingale residuals plotted against (**A**) patient age and (**B**) FIGO stage; median for each stage is denoted by a horizontal line.

improve the fit to a significant degree. Thus, we may be reasonably confident that the model is adequate for both covariates.

Identifying the correct parametric model

When fitting a fully parametric model, the survival times are assumed to follow a statistical distribution. Several different distributions have been proposed, and the identification of a suitable one is a crucial step. The most obvious distinguishing feature between parametric models is in the shape of the hazard they assume the data follow. The Weibull and Gompertz distributions are appropriate when the hazard is always increasing or decreasing; the Log-Logistic may be used where the hazard either rises to a peak and then decreases or always decreases; the Log-Normal and Generalised Gamma models are preferable when the hazard rises to a peak before decreasing. In the Exponential model, the hazard is assumed to be constant over time. The actual shapes of these distributions (e.g. the point in time at which a hazard ‘peaks’ or the gradient at which it increases/decreases) depend on *ancillary* parameters that are also estimated from the data. For example, when using the Weibull distribution, the hazard function, $h(t)$, is $\lambda s(\lambda t)^{s-1}$. In this case, the shape (s) and scale (λ) parameters are the ancillary parameters to be estimated (see Figure 1 in the previous paper Bradburn *et al*, 2003).

If the shape of the disease hazard is known to be different from that of a particular distribution, then the data should not be analysed with this parametric model. For example, consider the hazard for overall survival after cancer diagnosis. The hazard is rarely constant, thus ruling out an Exponential distribution. In some cases, the hazard rises sharply (due to treatment deaths) before tailing off, which would also rule out the Weibull. An informal assessment of a parametric model’s appropriateness may be made via plotting the (smoothed) empirical hazard or cumulative hazard against those estimated by the model, or by log(−log(survival)) plots which are discussed later. Akaike’s Information Criterion (AIC) (Akaike, 1974), a statistic that trades off a model’s likelihood against its complexity, may also be used when comparing the viability of different parametric models. The AIC of a model may be defined as

$$\mathrm{AIC} = -2\mathrm{LL} + 2(c + a)$$

where LL is the logarithm of the model likelihood (*log-likelihood*), c is the number of *covariates* and s the number of ancillary parameters (e.g. 2 in the case of the Weibull; λ and s). A *lower* value of the AIC suggests a better model. Note, however, that the likelihood computed in a Cox model is a partial likelihood, and so it is not possible to compare Cox PH models to fully parametric ones in this manner.

In the PH framework, it may be clear that none of the parametric models suggested here or elsewhere adequately capture the distributional form of the data. In such cases, the more flexible Cox model is the obvious choice. Commonly used parametric models in the AFT framework are arguably more flexible than those available in the PH framework, and so fitting a parametric AFT model is another option.

Overall goodness-of-fit tests

A simple test for the model adequacy is to compare the overall (Kaplan–Meier) survival curve to the model-based predicted survival and, ideally, for any group of patients the two should be close, if not identical. Hosmer and Lemeshow (1999) suggest using a more formal measure of fit based on comparing observed and expected events in different *risk groups* as defined by the model. Specifically, the predicted risk or prognostic index (PI) from a model consisting of covariates x_1, x_2, …, x_p with estimated coefficients b_1, b_2, …, b_p, respectively, is

$$\mathrm{PI} = b_1x_1 + b_2x_2 + \cdots + b_px_p$$

PI is calculated for each patient. Risk groups are constructed by categorising the (ranked) PIs, for example, three risk groups can be created using the highest, middle and lowest tertiles of PI. A score test is then applied to the differences between the observed and expected events in the risk groups. A simple approximation to this calculation may be obtained by adding the risk groups as a series of covariates to the survival model itself. A significant improvement in the model likelihood suggests that the original covariates form an insufficient model for the data.

Assessing overall goodness of fit on the ovarian cancer data

The predicted survival curves for the ovarian cancer model are potentially misleading. Several factors are associated with length of survival, and some are also related to (or correlated with) each other (e.g. histology and stage). Predicted survival curves for each histological group may be estimated by fixing all other covariates at their mean values. However, this approach will give an estimate of survival that is different to those observed in the data because correlations are ignored. The test of Hosmer and Lemeshow (1999) is more useful here. The patients are split into ten risk groups, with the proportion of deaths in each ranging from 10% in the best prognosis group to 94% in the worst. The approximate score test, derived from adding nine covariates to the model, produced no evidence of a poor fit (likelihood ratio test $\chi^2 = 7.84$ on 9 degrees of freedom, $P = 0.55$).

ASSESSING WHETHER PH IS APPROPRIATE

The PH assumption, that is, the hazards are proportional (and not overlapping) at all points in time, should be verified. An obvious approach is to plot the hazard in each group, but this is of limited use. The empirical hazard function is generally not well estimated, and instead the cumulative hazard is generally preferred to assess the PH assumption. If a PH model is valid, a plot of the logarithm of the cumulative hazard function in each group against the logarithm of time should give rise to lines that are parallel. Continuous variables need to be categorised into groups. The plot described is also known as the log(−log(survival)) plot, as the cumulative hazard is equal to the negative logarithm of the survival proportion. This approach requires a subjective assessment. Unfortunately, convergent or divergent lines may be due to either a lack of proportionality or to the omission of an important covariate. In practice, it is not known which, but this phenomenon suggests an inadequate model. On the other hand, parallel lines suggest that models assuming PHs may be suitable. In the case of fitting a Weibull or an Exponential parametric model, the lines should be parallel and straight.

Several formal statistical tests have been proposed for assessment of proportionality of hazards. A simulation study by Ng'andu (1997) described and compared several tests in the Cox PH framework, and concluded that the (weighted) scaled Schoenfeld residuals test (Grambsch and Therneau, 1994), the linear correlation test (Harrell, 1986) and the time-dependent covariate test (Cox, 1972) were the most powerful diagnostic tools for proportionality. The first two of these test for an association between residuals and time (evidence of which indicates a bad fit), and the third tests whether the effect (coefficient) of a covariate changes with time (i.e. nonconstant hazard ratio). This latter method is appealing as it not only detects nonproportionality, but allows it to be modelled validly. An alternative is to fit a *stratified* model, wherein a covariate that displays nonproportionality is modelled without the constraint of proportionality. Such a covariate must obviously be categorical (or be categorised), but more importantly has no estimated effect size provided when forming the strata of a stratified model, and thus is suitable only for covariates that are not of primary interest. Abandoning the PH approach in favour of some other model is clearly another option.

Assessing the appropriateness of PH for the ovarian cancer data

The Kaplan–Meier survival curves and log(−log(survival)) *vs* log(time) plots are shown for FIGO stage and histology in Figure 2A–D. The log(-log(survival)) plot for FIGO stage gave rise to reasonably parallel lines and therefore suggests proportionality. However, in the case of the histology, this appears to be violated. In particular, the prognosis for the endometroid group sits in the middle of all the groups in the first year but improves thereafter. A similar feature was apparent for the presence of ascites, where the initial detrimental effect becomes less important with time (data not shown). The (weighted) scaled Schoenfeld residuals test suggested significant overall nonproportionality ($P=0.05$), as did the time-dependent covariate tests for these terms. Therefore, despite other aspects of this model appearing adequate, the assumption of proportionality appears to be violated.

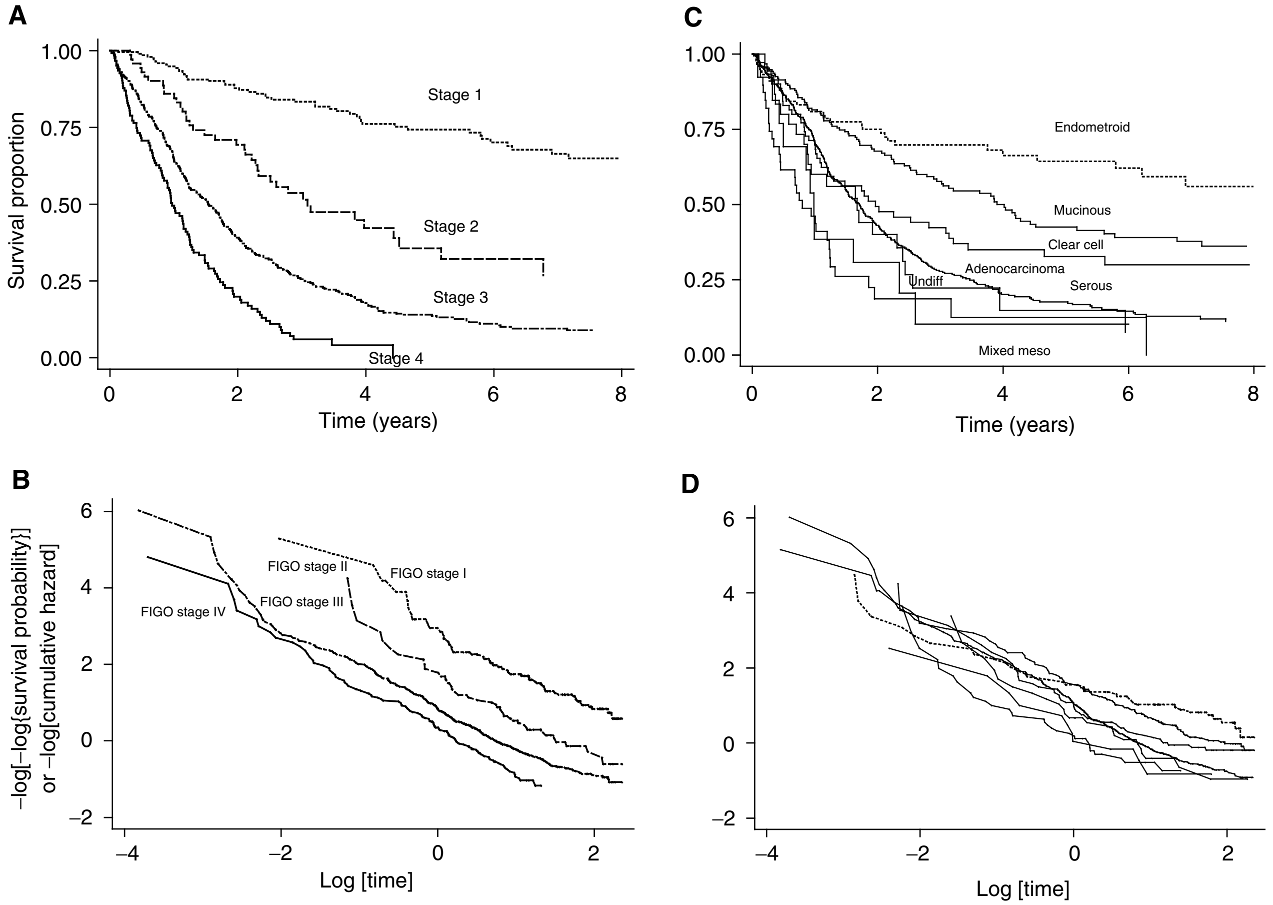


Figure 2 (**A**) Survival according to FIGO stage. (**B**) Log(−log(survival)) for FIGO stage. (**c**) Survival according to histology. (**D**) Log(−log(survival)) for histology. The endometroid group is shown by the dotted line.

Table 4 The Cox model applied to the ovarian data, with a time dependency added to ascites and endometroid terms

Covariate	Coefficient (b_i)	HR [exp(b_i)]	95% CI	*P*-value
FIGO	0.734	2.09	(1.83–2.38)	<0.001
Histology				<0.001
Serous	(0.000)	(1.00)		
Mucinous	−0.432	0.65	(0.50–0.85)	
Clear cell	0.344	1.41	(0.99–2.01)	
Adenocarcinoma	0.494	1.64	(0.91–2.96)	
Undifferentiated	0.769	2.16	(1.06–4.40)	
Mixed mesodermal	0.825	2.28	(1.50–3.47)	
Endometroid	0.312	1.37	(0.90–2.07)	
Endometroid × log(time)	−0.500	0.61	(0.45–0.82)	0.001
Grade				<0.001
1	(0.000)	(1.00)		
2	0.826	2.28	(1.32–3.95)	
3	0.843	2.32	(1.35–4.00)	
Absence of ascites	−0.466	0.63	(0.50–0.80)	<0.001
Ascites × log(time)	0.233	1.26	(1.01–1.58)	0.04
Age (per 5-year increase)	0.134	1.14	(1.09–1.20)	<0.001

HR = hazard ratio; CI = confidence interval.

Table 5 Akaike Information Criterion (AIC) of five different distributions fitted to the full model

Model	Log likelihood (LL)	No. of covariates (*c*)	No. of ancillary parameters (*a*)	AIC
Exponential	−259.28	11	1	542.55
Weibull	−253.21	11	2	532.41
Log-Normal	−238.22	11	2	502.44
Log-Logistic	−236.33	11	2	498.65
Generalised Gamma	−235.79	11	3	499.58

AIC = −2LL+2(*c*+*a*).

Nevertheless, we can still use a Cox PH model with time-dependent covariates implemented, which is a model that includes interaction terms between the covariates and (log) time, and thus allows the effect of the relevant covariates to change with time. Table 4 shows the amended model that now allows the effects to vary with time. The time-dependent terms suggest that the absence of ascites and endometroid histology have effects that diminish (the hazard ratios tend towards 1) with time. For example, the absence of ascites is judged to have a hazard ratio of $\exp(-0.466 + 0.233 \times \log(2)) = 0.74$ at 2 years but $\exp(-0.466 + 0.233 \times \log(5)) = 0.91$ at 5 years.

ASSESSING WHETHER AN AFT MODEL IS ADEQUATE

In the AFT model, the survival proportion in one group at any time t is equal to the survival proportion in the second at time φt, where φ is constant. Therefore, a Quantile–Quantile (Q–Q) plot of the times of survival percentiles should lie on a straight line of slope φ that passes through (0, 0). As with the log(−log(survival)) plot in PH models, this is a useful but limited approach as departures from linearity could be due to the AFT model being inappropriate or that one or more important covariates have been omitted. The methods of stratification or modelling with time-dependent covariates suggested in the PH section may be applied here as well.

The lung cancer trial data

We assessed the adequacy of the Generalised Gamma and four other parametric models (each with all covariates included) and present their AIC values in Table 5. The Generalised Gamma model has a higher log-likelihood than the other models and a lower AIC, indicating that this distribution may be the most accurate. To check for excluded covariates, the Martingale residuals were plotted against potential model terms as before. None of these plots suggested that a covariate was incorrectly omitted. Figure 3

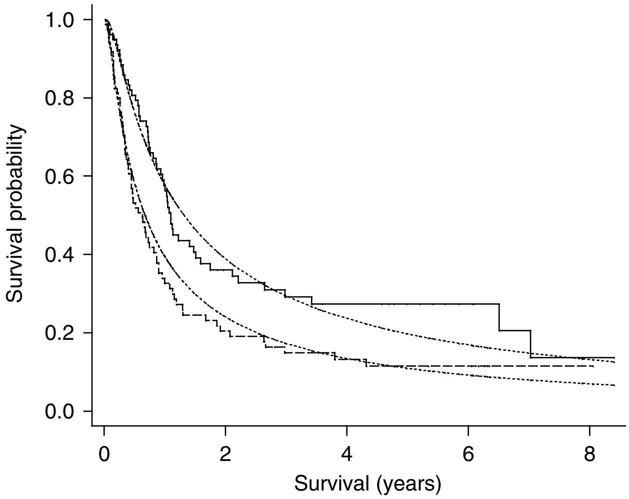


Figure 3 Kaplan–Meier survival probabilities for patients treated by RT + CAP (solid line) and RT alone (dashed line). The respective predicted survival proportions of a generalised gamma multivariate model are given by the faint dotted lines for grouped mean covariates. RT = radiotherapy, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

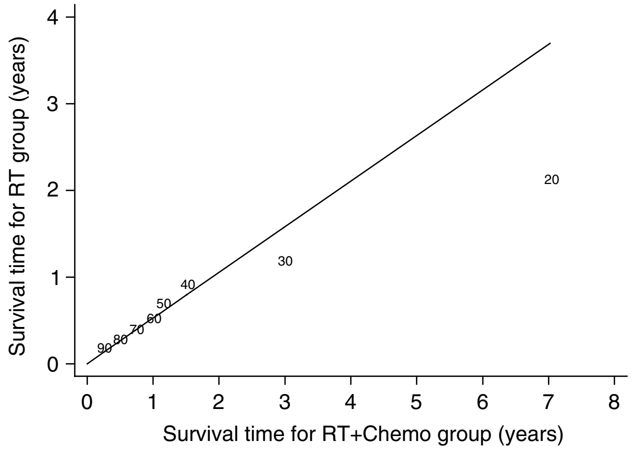


Figure 4 Q–Q plot (percentiles of survival distribution) for patients RT + CAP against those with RT only. The plot symbols are the survival percentiles* and the slope corresponds to the value of the time ratio (= 1/1.90). * The 10th percentile is omitted: 4.9 and 11.4 years for RT and RT + CAP respectively, RT = radiotherapy, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

gives the predicted observed survival curves together with the predicted survival under a Generalised Gamma model; for each treatment group, the Karnofsky performance status and cell type were fixed at their mean values. The medium-term survival is not as well fitted by the model but is tolerably close. The long-term survival is also less well estimated, but because few patients survive this length of time the estimated survival is imprecise and so this does not cause grounds for concern. The survival times for the 10th, 20th, …, 90th survival percentiles for each treatment group are plotted as a Q–Q plot in Figure 4 and, again, apart from the later times seem to fit adequately.

DISCUSSION

This paper has sought to demonstrate the models introduced in the previous paper in this series (Bradburn *et al*, 2003), to offer practical advice on how to select a method that represents the data fairly, and how to present and interpret it. Good modelling of survival data is not a straightforward exercise, and it is not possible to suggest an ‘off the peg’ solution. Before starting the process of deciding which (if any) of the models suggested is most suitable for an individual dataset, the important question of why the model should be fitted needs to be considered. The answer should inform the modelling process. Although it is possible to choose a model from those suggested that is optimal from a purely statistical point of view (e.g. goodness-of-fit measures), nonstatistical considerations should to be taken into account. The choice of model and of covariates therein should, in general, be suggested from experience and based on the specific question under investigation. However, good nonstatistical reasons informing model choice should not override good statistical reasons for not choosing that model. The diagnostics (e.g. residuals) for the different models may be difficult to interpret, but they will give an indication of whether modelling assumptions hold and, ultimately, should be considered when model building.

In some cases, all of the models mentioned above may not be wholly appropriate either for modelling the data or answering the relevant question. Consider an example where the time between treatment and possible multiple cancer relapse is to be investigated. The methods introduced assume one survival time (culminating in one type of event), but we may be dealing with patients who have one or more relapses of different type or levels. In the final paper of this series, we introduce models that extend the types of models described here to incorporate recurrent events. We also present approaches to modelling continuous covariates in a nonlinear fashion, validating models and discuss alternatives when fundamental censoring assumptions do not hold.

ACKNOWLEDGEMENTS

We wish to thank John Smyth for providing the ovarian cancer data, and Victoria Cornelius and Peter Sasieni for invaluable comments on an earlier manuscript. Cancer Research UK supported all authors. Taane Clark holds a National Health Service (UK) Research Training Fellowship.

REFERENCES

Akaike H (1974) A new look at the statistical model identification. *IEEE Transaction and Automatic Control* **AC-19:** 716–723

Bradburn MJ, Clark TG, Love S, Altman DG (2003) Survival analysis. Part II: multivariate data analysis – an introduction to concepts and methods. *Br J Cancer* **89**(3)**:** 431–436

Clark TG, Altman DG (2002) Developing a prognostic model in the presence of missing data: an ovarian cancer case-study. *J Clin Epidemiol* **56:** 28–37

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis. Part I: basic concepts and first analyses. *Br J Cancer* **89**(2)**:** 232–238

Collett D (1994) *Modelling Survival Data in Medical Research*. London: Chapman & Hall

Cox DR (1972) Regression models and life tables (with discussion). *J R Stat Soc B* **34:** 187–220

Grambsch PM, Therneau TM (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81:** 515–526

Harrell FE (1986) The PHGLM procedure. *SAS Supplemental Library Users Guide*, Version 5 edition. Cary, NC: SAS Institute

Henderson HV, Velleman PF (1981) Building multiple regression models interactively. *Biometrics* **37:** 391–411

Hosmer DW, Lemeshow S (1999) *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley

Machin D, Campbell MJ, Fayers PM, Pinol APY (1998) *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Science

Ng’andu NH (1997) An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox’s model. *Stat Med* **16:** 611–626

Peduzzi P, Concato J, Feinstein AR, Holford TR (1995) Importance of events per independent variable in proportional hazards regression analysis II: accuracy and precision of regression estimates. *J Clin Epidemiol* **48:** 1503–1510

British Journal of Cancer (2003) 89, 781 – 786
© 2003 Cancer Research UK All rights reserved 0007 – 0920/03 $25.00
www.bjcancer.com

Tutorial Paper

Survival Analysis Part IV: Further concepts and methods in survival analysis

TG Clark*,[1], MJ Bradburn[1], SB Love[1] and DG Altman[1]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Oxford OX3 7LF, UK*

British Journal of Cancer (2003) **89,** 781–786. doi:10.1038/sj.bjc.6601117 www.bjcancer.com
© 2003 Cancer Research UK

Keywords: survival analysis; missing data; validation; repeated events

INTRODUCTION

In the previous papers in this series (Bradburn *et al*, 2003a, b; Clark *et al*, 2003), we discussed methods for analysing survival time data, both univariate and multivariate. We have dealt with only a portion of the methods available for analysing survival time data, and in many cases, useful alternatives to (or extensions of) these methods exist. We have also left unanswered other questions regarding the design and analysis of studies that measure survival time and, in particular, dealing with situations where some standard modelling assumptions do not hold. We conclude this series by tackling these issues. These ideas are described in a question and answer format, and introductory references are provided for the reader to investigate further.

IN A SURVIVAL ANALYSIS, CONTINUOUS VARIABLES ARE SOMETIMES CATEGORISED. SHOULD WE DO THIS (AND IF SO, HOW)?

In medical research, it is common to see continuous measures grouped into categories to simplify a covariate's relationship with survival and its interpretation. There is no statistical reason for grouping and it can lead to as many problems as it seeks to avoid. The categorisation of a continuous covariate by definition discards data and can be seen as introducing measurement error. It also leads to biased estimates and a reduced ability to detect real relationships (Schmoor and Schumacher, 1997; Altman, 1998). Nevertheless, there are sometimes good reasons to categorise a continuous covariate in the analysis of survival (and indeed any) data. When doing so, it is wise to note the following points:

1. Use cut-points that have been predetermined rather than testing multiple values. A common choice of boundaries is fixed centiles such as quartiles. It is preferable though to use established cut-points that have clinical meaning, and therefore provide consistent groupings between studies. Examples include dividing oestrogen receptor level at 10 fmol, and age into 5- or 10-year intervals.
2. Do not choose cut-points based on minimising *P*-values, as this method gives biased results (Altman *et al*, 1994; Altman, 1998).
3. If possible, use more than two categories to reduce the loss of information and allow some assessment of the linearity of any trend.
4. Ensure that each group contains an adequate number of individuals (and events).

Grouping is sometimes used because there are concerns with mismodelling the relationship when there is a nonlinear relationship between the variable and log hazard. The simplest approach is to evaluate the effect of adding a quadratic term to the model, but better approaches to use are smoothing splines (Therneau and Grambsch, 2000) or fractional polynomials (Royston *et al*, 1999). Figure 1 shows the result of modelling a new covariate, (log) CA125, in the previously used ovarian cancer data, by the method of smoothing splines (with 11 degrees of freedom). There is evidence of nonlinearity ($P = 0.002$) and the plot suggests that CA125 might be modelled as a cubic effect. It is clear that modelling the data using a binary or linear variable would be inappropriate here (see Figure 1). Knorr *et al* (1992) discussed these issues in the context of prognostic studies in cancer.

IN OUR CLINICAL TRIAL, WE COLLECTED MEASUREMENTS AT PREARRANGED VISITS. CAN WE INCLUDE MULTIPLE MEASUREMENTS FOR THE SAME COVARIATE IN OUR SURVIVAL ANALYSIS?

If variables measured after entry into the study are to be included in the survival model, special methods are required. Such methods are called *time-dependent* (or *updated*) *covariate methods*, as the variables they incorporate may change value over time. For example, if a longitudinal study seeks to assess the effects of smoking on cancer, a variable for each patient may be defined, being equal to 0 (nonsmoker) or 1 (smoker) at any time. If a nonsmoker begins smoking after entering the study, then this covariate is updated (from '0' to '1') at the time that smoking begins. This covariate contributes more information than using smoking status at time of entry alone. It is important to note that post-entry measurements cannot be validly incorporated into a survival model without using these methods.

Recall that for the proportional hazards model, the formula relating a covariate x_1 to the hazard $h(t)$ at time t is

$$h(t) = h_0(t)\exp[b_1\, x_1]$$

where $h_0(t)$ is the baseline hazard. If repeated measurements of a covariate x_1 are available, the formula changes to

$$h(t) = h_0(t)\exp[b_1\, x_1(t)]$$

*Correspondence: Mr TG Clark; E-mail: taane.clark@cancer.org.uk

Received 6 December 2002; accepted 30 April 2003

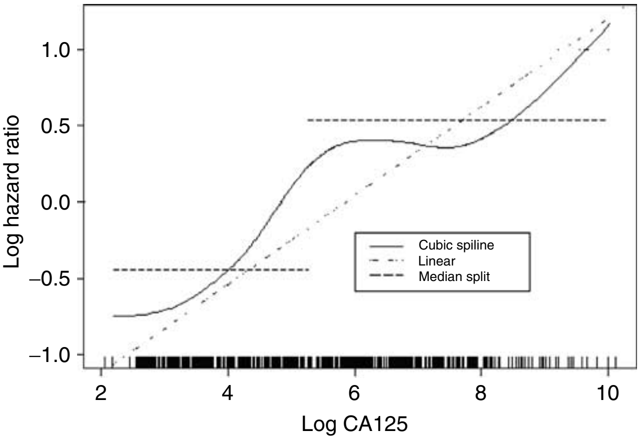


Figure 1 Modelling log CA125 using spline functions: | corresponds to measurements.

where $x_1(t)$ is the value of x_1 at time t. (It is also possible to use, but harder to interpret, an accelerated failure time model here.) The covariate x_1 may be continuous or categorical, and may change freely or at fixed time intervals. The coefficient b_1 represents the additional relative hazard for each unit increase in x_1 at any given time. This model is different from models with *time-dependent coefficients* (Bradburn *et al*, 2003b), in which the *effect* of a covariate changes rather than the value of the covariate itself, that is, $h(t) = h_0(t)\exp[b_1(t)\ x_1]$.

The time dependent method can be applied in many standard statistical software packages. However, the approach described requires a large amount of data and is therefore rarely seen. One also has to be confident that the collection process is not *itself* dependent on clinical progress, perhaps by using scheduled assessments. Further details of the method, and some precautions, are noted in Altman and De Stavola (1994).

MOST SURVIVAL ANALYSIS METHODS ASSUME THE CENSORING IS NONINFORMATIVE. WHAT IF THE CENSORING IS INFORMATIVE?

Informative censoring occurs when individuals are lost to follow-up for reasons that may relate to their (unknown) outcome. For example, in a randomised trial in which the main outcome is time to cancer recurrence, a patient who is lost to follow-up may be more likely to have experienced drug toxicity or ill health and thus may also be more susceptible to (earlier) relapse. Informative censoring introduces bias into the standard methods discussed previously. Unfortunately, it is difficult both to identify informative censoring and to assess its impact. It is helpful though to know what proportion of censored individuals were lost to follow-up before the end of the study (Clark *et al*, 2002).

A simple, *ad hoc* approach to the problem is to perform sensitivity analyses, to assess the impact of assigning different survival times to those patients whose observed (censored) survival times may have been affected in this manner. For example, if a patient suspected to be in ill health exits the study at 4 weeks, a first analysis may be performed with this patient censored at 4 weeks and a second where the patient is assumed to have relapsed at 4 weeks (i.e. a ‘best case – worst case’ scenario). This approach works best when there are few such patients, but in that situation, the possible bias will be very small. Another possibility is to decide *a priori* that all such patients will be treated in a particular way. The issue has been of particular concern in randomised trials of nicotine replacement therapy, in which losses to follow-up are considerable. In a systematic review of randomised trials, patients who were lost to follow-up were regarded as being continuing smokers (Silagy *et al*, 2002).

More formal approaches have been proposed (e.g. Robins, 1995a, b; Scharfstein *et al*, 2001). In general, they assume that a relationship exists (and can be modelled) between censoring times and baseline covariates and perhaps also post-treatment patient data. It is difficult to evaluate the assumptions of these complex methods, and implementation in statistical software is limited.

If follow-up stops because the patient has experienced a different defined event, the problem may be viewed as a competing risk scenario (see below), or handled via a mixture model (or ‘cure’ model), where the differing event types are explicitly modelled. The latter method makes particular sense if the two events are quite dissimilar, such as patient recovery and patient death.

In practice, if there is little informative censoring, the bias introduced to standard methods is minimal, and in general using these along with simply reporting loss to follow-up (perhaps with a basic sensitivity analysis) will suffice. Good patient follow-up and avoidance of unnecessary drop-out is by far the best solution, and when and why drop-out occurs should always be reported (Moher *et al*, 2001).

SOME COVARIATE DATA ARE MISSING IN OUR ANALYSIS. WHAT SHOULD WE DO?

Missing data are a common problem when developing survival models in cancer. Individuals without complete covariate data are usually omitted, but the resulting analysis has reduced power and may be an unrepresentative subset of patients. Often many covariates have missing data, and the absence of a small percentage of data points for each variable can lead to a greatly depleted sample. Unless only a few values are missing, some investigation of the missing data and methods that accommodate it should be considered. In the ovarian cancer data set presented previously, a small number of important factors containing little or no missing data were used. The database contains several other factors in which missing data were frequently encountered, and a more definitive analysis (Clark *et al*, 2001) was able to incorporate these factors, while retaining all patients by applying multiple imputation methods (Van Buuren *et al*, 1999). Multiple imputation is a framework in which missing data are imputed or replaced with a set of plausible values. Several data sets are then constructed, each being analysed separately, and their results are combined while allowing for the uncertainty introduced in the imputation. Other approaches exist (e.g. Lipsitz and Ibrahim, 1998), but imputation approaches have more software available (Horton and Lipsitz, 2001). Further details, discussion and references are given in another analysis of the ovarian data found in Clark and Altman (2002).

We recommend that authors of research papers are explicit about the amount of missing data for each variable and indicate how many patients did not have complete data. Imputation techniques are powerful tools and are increasingly available in software, but are not a panacea. Inherent in the method is the assumption that a model relating data absence to other measured covariates (and possibly survival too) exists and can be specified. This has much in common with the situation where informative censoring is suspected, and similarly, their practical experience is limited at the present time. Researchers should be aware of the assumptions, most of which are untestable, and use sensitivity analysis to assess the robustness of results. Ultimately, these problems are best avoided by minimising missing data.

HOW SHOULD WE CHOOSE WHICH VARIABLES TO INCLUDE IN OUR SURVIVAL MODEL?

In some cases, the factors to be included in the model will be predetermined. In many others, there will be several possible covariates from which only a handful are to be chosen. This is often because there are a large number of covariates of which some are unimportant, but the identification and elimination of these is not always easy. As a starting point, it is good practice to include known prognostic factors and any that are specifically required by the study aims (e.g. the treatment identifier in the analysis of a clinical trial). It is then the burden of new factors to add significant additional predictive ability (Simon and Altman, 1994).

If there are a large number of factors of interest and there is relatively little information about their prognostic influence, automated selection techniques such as stepwise methods can be used. There are variations on these that start either with all covariates (backward elimination) or none (forward selection), adding or removing covariates according to statistical significance at some predecided level. A disadvantage of both is that they only evaluate a small number of the set of possible models. Instead, each possible model could be fitted, with the best being picked on the basis of a goodness-of-fit measure such as Mallow's C (Hosmer and Lemeshow, 1999). However, this may be time-consuming with many covariates, multiple testing is problematic, and is seldom used due to its noninclusion in many software packages.

Unfortunately, all these methods are problematic. The 'best' model is derived solely on statistical grounds (and indeed may lack any clinical meaning), the regression coefficients produced are biased (too large) and standard errors and *P*-values are too small, especially for smaller sample sizes and when few events occur. Backward elimination is possibly the best of the above methods for identifying the important variables, and it allows one to examine the full model, which is the only fit providing accurate standard errors and *P*-values (Harrell, 2001). An alternative, the lasso method (Harrell, 2001) attempts to force some regression coefficient estimates to be exactly zero, thus achieving variable selection while shrinking the remaining coefficients toward zero to reflect the overfitting and overestimation caused by data-based model selection.

If one cannot completely prespecify a model, it may be best to apply backward elimination or lasso to a full model of prespecified covariates of interest, and use bootstrap methods to compare the stability and predictive accuracy of the full model with that of a reduced one (see next question for further details).

WE HAVE DEVELOPED A PROGNOSTIC MODEL FOR OVERALL SURVIVAL. HOW CAN WE MEASURE ITS PREDICTIVE ABILITY? HOW CAN THE MODEL BE VALIDATED?

In survival analysis, statistical models are employed to identify or propose combinations of risk factors that might predict patient survival. It follows that to be of use, the model must be able to: (1) make unbiased predictions, that is, give predicted probabilities that match closely those observed, and (2) distinguish higher and lower risk patients. These are the two components of predictive ability, and are called *calibration* and *discrimination,* respectively. Importantly, models rarely perform as well on either basis when used to predict survival in patients other than those used to derive the model. A model that closely mirrors the survival patterns of the present data is said to have *internal validity*, but to be of wider use should do so for other groups of patients as well (be *externally valid*). Before a model is applied routinely in clinical practice, it should have been shown to meet both criteria.

Measures of discrimination include the *c*-index and Nagelkerke's $R^2(R^2_{\mathrm{N}})$ (Harrell, 2001). The *c*-index, a generalisation of the area under the receiver operating characteristic (ROC) curve, is the probability of concordance between observed and predicted survival based on pairs of individuals, with $c=0.5$ for random predictions and $c=1$ for a perfectly discriminating model. Similarly, $R^2_{\mathrm{N}}=0$ indicates no predictive ability and $R^2_{\mathrm{N}}=1$ indicates perfect predictions. Calibration may be quantified using an estimate of slope shrinkage (Harrell, 2001). Each quantity may be evaluated for the data used in the modelling by randomly splitting the patients into two samples, one to derive the model and the other to validate it. The proportion of data to include in each sample is, however, arbitrary and although estimates of predictive accuracy from this approach are unbiased, they also tend to be imprecise. Bootstrapping, a method that involves analysing subsamples from a data set, or 'leave-one out' cross-validation may be more beneficial. For these analyses, an alternative is to estimate shrinkage factors and apply these to regression coefficients to counter overoptimism. These techniques allow evaluation on multiple data sets. Once the internal validity of a model has been established, it can be tested for its generalisability by applying the model to other patients, and using the above methods to assess the adequacy of the predictions.

A good summary of important issues can be found in Justice *et al* (1999) and Wyatt and Altman (1995), and more details on the statistical methods are given in Altman and Royston (2000). In summary, internal validation is necessary before a model is proposed, and external validation is highly recommended before it is to used in clinical practice.

CAN WE PERFORM AN ANALYSIS WHERE THERE ARE UNMEASURED FACTORS THAT MAY AFFECT SURVIVAL TIME?

In practice, one cannot be sure that all important prognostic variables have been measured. In general, omitting variables will simply reduce the predictive ability of a model, so that patients with similar measured covariates will exhibit large variability in their survival. When a strongly prognostic variable is omitted, however, the model may be biased. In particular, the estimated treatment effect in a randomised trial may be biased if an important prognostic variable is not adjusted for, even when that variable is balanced between the treatment groups (Schmoor and Schumacher, 1997; Chastang *et al*, 1988). It is inappropriate to proceed at all if vital information such as clinical stage in breast cancer patients is unavailable.

Another form of missing covariate is when some individuals have a shared exposure that is unmeasured. For example, members of the same family will have shared dietary and other environmental exposures, so that their outcomes cannot be considered to be independent. A similar situation arises in cluster randomised trials and multicentre trials in general (Yamaguchi *et al*, 2002). Such data can also be considered as being 'multilevel', with variation both between and within groups. Random effects (or 'frailty') models can be used to allow covariate effects to vary across groups (O'Quigley and Stare, 2002). Such models are widely used in other contexts, in particular, in meta-analysis. Frailty can also be considered to apply to individuals, relating to the idea of unmeasured variables as a possible explanation for observed heterogeneity of outcome. Use of such models depends on precise knowledge of the frailty distribution, which is generally not available (Keiding *et al*, 1997).

Lack of fit of a Cox model may be better explained by other modelling approaches (O'Quigley and Stare, 2002), such as the accelerated failure time model (Keiding *et al*, 1997).

SEVERAL PAPERS IN OUR RESEARCH AREA HAVE APPLIED (ARTIFICIAL) NEURAL NETWORKS AND REGRESSION TREES AS AN ALTERNATIVE TO THE COX MODEL. WHAT ARE THESE METHODS?

Artificial neural networks

Artificial neural networks (ANNs) are a relatively new method for assessing the extent to which a series of covariates explain patient outcomes. The key feature of the ANN methodology is to assume that there are some latent, or 'hidden', intermediary variables in the input (covariate) and output (survival probability) processes. The most common model is the three-layer model shown in Figure 2. Under this model, the covariates (input) do not act directly on the response variable (output), but channel their influence into a series of latent (hidden) variables. It is the relative importance of these unobservable variables which determines the survival. For a more detailed introduction to these methods, see Cross *et al* (1995).

This methodology is appealing in that it can incorporate complex relationships between covariates and survival more easily than standard approaches such as Cox regression, which may be too simplistic. However, there have been several major criticisms of the method: (a) the high chance of overfitting the data, (b) the lack of easy interpretation of the model and of the impact of individual covariates, (c) the perceived 'black box' methodology involved, and (d) the difficulty in handling censored survival times. The last issue arises because it is usually the status of the individuals (i.e. alive or dead) at a given point (or points) in time that is taken to be the response. Biganzoli *et al* (1998) and others have modelled the hazard functions directly, in a promising attempt to extend this method. Reviews comparing the examples where both ANN and regression methods had been used to derive prognostic models have found that overall ANNs are little better than classical statistical modelling approaches (Sargent, 2001), and misuses of ANNs in oncology are common (Schwarzer *et al*, 2000). We therefore advise caution in their use, and the involvement of an experienced statistician.

Classification and regression trees

The classification and regression tree (C&RT) approach is based on dividing the cohort into groups of similar response patterns, using covariates (Lausen *et al*, 1994). The partitioning algorithm starts with the covariate that best discriminates the survival outcome between two subgroups. For continuous or multicategory variables, the method thus needs to determine the threshold that best dichotomises the variable. This process is repeated for each subgroup in turn using all the available covariates. The same covariate may be used more than once, and the process stops eventually with either no covariate adequately dividing the subgroups further or when the subgroups have reached a specified minimum size. Figure 3 shows an unpublished C&RT analysis in a Dukes' B colonic cancer study, in which four categorical variables (perforation, peritoneal involvement, venous and margin) were assessed for their prognostic value in overall survival. Using a logrank test at each step, it was found that peritoneal involvement (levels 1, 2 *vs* 3, 4) discriminated best between good and bad survival, and level 1 venous subdivided patients with high levels of peritoneal involvement. The stopping rule employed was the first occurrence of either (a) the maximum logrank statistic is not statistically significant at the 1% level or (b) when any subgroup contains less than 25 patients. The latter condition ceased the partitioning algorithm in the example, yielding the three groups of patients described in Figure 3.

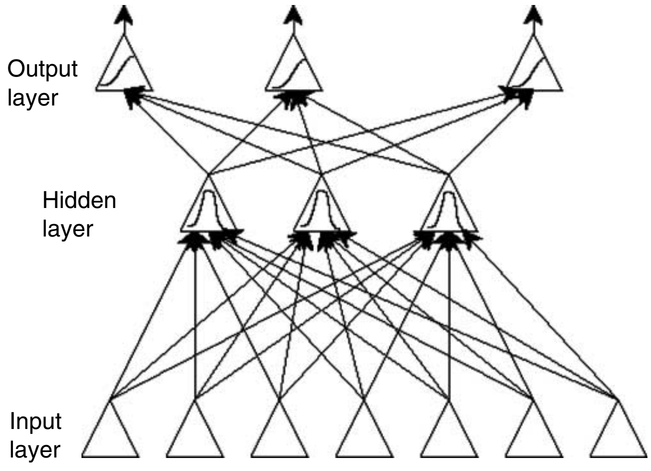
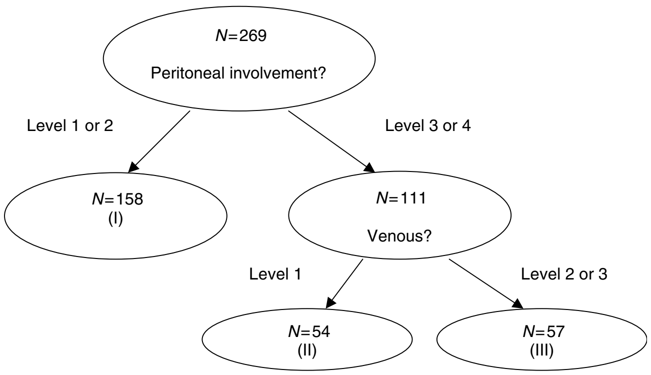


Figure 2 An example of an ANN.



	5-year survival (95% CI)	*N*	Deaths
(I) Peritoneal 1 or 2	87.1 (80.2, 91.7)	158	21
(II) Peritoneal 3 or 4 and venous 1	73.7 (58.8, 83.9)	54	13
(III) Peritoneal 3 or 4 and venous 2 or 3	45.7 (31.0, 59.3)	57	29

Figure 3 A CART for Dukes' B colonic cancer study.

The major advantage of C&RT is its ease of interpretability – it reflects how many decisions are made. It also relies on fewer distributional assumptions (Schmoor *et al*, 1993) and is particularly useful in situations where there are interactions. The disadvantages of C&RT lie in having to decide what threshold to use for continuous covariates, and to correct for multiple testing and overfitting. The automated covariate selection is similar to forward stepwise methods in regression, and hence shares their problems (see the choice of covariate section). Finally, as C&RT seeks to classify patients into groups, it offers little in the way of estimated effect of risk factors. Nevertheless, C&RT is a useful complement to other methods, in particular as an exploratory tool that can inform future research.

CAN WE ANALYSE DIFFERENT TYPES OF EVENTS OR REPEATED EVENTS?

Traditional survival analysis methods (including all those discussed so far) assume that only one type of event of interest occurs, and at most once. More advanced methods exist to allow the investigation of several types of events (e.g. cancer death, vascular death, other), or an event that may occur repeatedly (e.g. cancer recurrence). We will describe methods for each in turn.

Where the survival duration is ended by the first of several events, it is referred to as *competing risks analyses*. Analysing the time to each event separately can be misleading, and in this context the Kaplan–Meier method, in particular, tends to overestimate the proportion of subjects experiencing each event. The cumulative incidence method, in which the overall event probability at any time is the sum of the event-specific probabilities, may be used to address this. Univariate tests and statistical models also exist, and an overview of several of the methods proposed can be found in Tai *et al* (2001). Models are generally implemented by entering each patient several times – one per event type – and for each patient, the time to any event is censored on the time at which the patient experienced another event.

Where multiple events of the same type occur, it is common practice to use the first event only, but this ignores information. Three approaches to use this extra information are demonstrated using artificial patient data in Table 1. In a *conditional* model, follow-up time is broken up into segments defined by events, with each patient being at risk for an *i*th event once the $(i-1)$th has occurred. Patient 1 in Table 1 is therefore assumed not to be at risk of a second event until the first has occurred, and so is at risk of experiencing this from time 8 until time 12. This model comes in two types: using either the time since the beginning of the study (type A) or the time since the previous event (type B). The *marginal* model, on the other hand, considers each event to be a separate process and, by definition, the time for each event starts at the beginning of follow-up for each patient. Here, all patients are considered to be at risk for all events, regardless of how many events they have previously had, and so patient 2, for example, was considered at risk of events 3 and 4 despite being lost to follow-up at the second. A third approach, called the *independent increment* model, is closest in spirit to a conditional model but takes no account of the number of previous events experienced by a patient, and for this reason the conditional and marginal models are often preferable. For each model, the data should be entered in the form of one patient record per event number as illustrated in Table 1.

Table 1 Data layout under four recurrent event models with patient 1 having three events (at times 8, 12 and 26) and patient 2 having two events (at times 10, 18)

Model	Patient i.d.	Time interval	Event[a]	Stratum[b]
Conditional A	1	(0,8]	1	1
	1	(8,12]	1	2
	1	(12,26]	1	3
	1	(26,31]	0	4
	2	(0,10]	1	1
	2	(10,18]	1	2
Conditional B[c]	1	(0,8]	1	1
	1	(0,4]	1	2
	1	(0,14]	1	3
	1	(0,5]	0	4
	2	(0,10]	1	1
	2	(0,8]	1	2
Marginal model	1	(0,8]	1	1
	1	(0,12]	1	2
	1	(0,26]	1	3
	1	(0,31]	0	4
	2	(0,10]	1	1
	2	(0,18]	1	2
	2	(0,18]	0	3
	2	(0,18]	0	4
Independent increment	1	(0,8]	1	1
	1	(8,12]	1	1
	1	(12,26]	1	1
	1	(26,31]	0	1
	2	(0,10]	1	1
	2	(10,18]	1	1

[a] 1 = had event of interest, 0 = censored. [b] Relates to the number of events and is used in the fitting of the model as the strata variable. [c] Gap time model.

All of the above models are usually applied within a Cox model framework, although accelerated failure time methods may equally be used. These models are fitted using the same basis as standard approaches, with two exceptions: (1) a cluster effect is used to adjust the standard errors because patients are repeated in the study, and (2) the analysis is stratified – with the exception of the independent increment method – with the event type (for competing risks) or number (for recurrent events) defining the strata. Interaction effects between covariates and strata may be used to assess whether covariate effects vary across competing outcomes or event number. For example, Kay (1986) presents an example of a treatment that reduces the risk of death from one cause, but increases the risk of death from another.

More thorough reviews of the above (and other related) methods can be found in Hosmer and Lemeshow (1999), and Therneau and Grambsch (2000). These modelling procedures are generally only a little more difficult than for single-event data, and software is widely available. As with any statistical model though, it is still important to assess its adequacy and fit. In each case, the choice of the best method of analysis will depend on the disease in question and the goals of the analysis. However, the aims such as those described here can often be highly relevant, and where this is the case these methods should be strongly considered.

SUMMARY

Most analyses of survival data use primarily Kaplan–Meier plots, logrank tests and Cox models. We have described the rationale and interpretation of each method in previous papers of this series, but here we have sought to highlight some of their limitations. We have also suggested alternative methods that can be applied when either the data or a given model is deficient, or when more difficult or specific problems are to be addressed. For example, analysis of recurrent events can make an important contribution to the understanding of the survival process, and so investigating repeat cancer relapses may be more informative than concentrating only on the time until the first. More fundamentally, missing data are a common issue in data collection that in some cases can seriously flaw a proposed analysis. Such considerations may be highly relevant to the analysis of a data set, but are rarely mentioned in the analysis of survival data. One possible reason for this is a perceived lack of computer software, but many of the approaches discussed here are currently incorporated into existing commercial statistical packages (e.g. SAS, S-Plus, Stata) and freeware (e.g. R). On the other hand, the desire may be to ‘keep things simple for the readership’. This view is reasonable, but is valid only where a simple analysis adequately represents the survival experience of patients in the study. Ensuring the analyses are appropriate is therefore crucial. More advanced survival methods can derive more information from the collected data; their use may admittedly convey a less straightforward message, but at the same time could allow a better understanding of the survival process.

The aim of this series has been to aid awareness, understanding and interpretation of the many and varied methods that constitute the analysis of survival data. It is paramount that analyses are performed in the knowledge of the assumptions that are made therein, and the more complex methods, in particular, are best applied by a statistician.

ACKNOWLEDGEMENTS

We thank Victoria Cornelius and Peter Sasieni for providing comments on an earlier draft. Taane Clark holds a National Health Service (UK) Research Training Fellowship. Cancer Research UK supported Douglas Altman, Mike Bradburn and Sharon Love.

REFERENCES

Altman DG (1998) Suboptimal analysis using 'optimal' cutpoints. *Br J Cancer* **78:** 556 – 557

Altman DG, De Stavola BL (1994) Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Stat Med* **13:** 301 – 341

Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* **86:** 829 – 835

Altman DG, Royston P (2000) What do we mean by validating a prognostic model? *Stat Med* **19:** 453 – 473

Biganzoli E, Boracchi P, Mariani L, Marubini E (1998) Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat Med* **17:** 1169 – 1186

Bradburn MJ, Clark TG, Love S, Altman DG (2003a) Survival analysis. Part II: multivariate data analysis – an introduction to concepts and methods. *Br J Cancer*

Bradburn MJ, Clark TG, Love S, Altman DG (2003b) Survival analysis. Part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. *Br J Cancer* (submitted)

Chastang C, Byar D, Piantadosi S (1988) A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. *Stat Med* **7:** 1243 – 1255

Clark TG, Altman DG (2002) Developing a prognostic model in the presence of missing data: an ovarian cancer case-study. *J Clin Epidemiol* **56:** 28 – 37

Clark TG, Altman DG, De Stavola BL (2002) Quantifying the completeness of follow-up. *Lancet* **359:** 1309 – 1310

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis. Part I: basic concepts and first analyses. *Br J Cancer* (submitted)

Clark TG, Stewart ME, Altman DG, Gabra H, Smyth J (2001) A prognostic model for ovarian cancer. *Br J Cancer* **85:** 944 – 952

Cross SS, Harrison RF, Kennedy RL (1995) Introduction to neural networks. *Lancet* **346:** 1075 – 1079

Harrell FE (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag

Horton NJ, Lipsitz SR (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Stat* **55:** 244 – 254

Hosmer DW, Lemeshow S (1999) *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley

Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalisability of prognostic information. *Ann Int Med* **130:** 515 – 524

Kay R (1986) Treatment effects in competing-risks analysis of prostate cancer data. *Biometrics* **42:** 203 – 211

Keiding N, Andersen PK, Klein JP (1997) The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Stat Med* **16:** 215 – 224

Knorr KL, Hilsenbeck SG, Wenger CR, Pounds G, Oldaker T, Vendely P, Pandian MR, Harrington D, Clark GM (1992) Making the most of your prognostic factors: presenting a more accurate survival model for breast cancer patients. *Breast Cancer Res Treat* **22:** 251 – 262

Lausen B, Sauerbrei W, Schumacher M (1994) Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In *Computational Statistics*, Dirschedl P, Osermann R (eds) Heidelberg/New York: Physica-Verlag

Lipsitz SR, Ibrahim JG (1998) Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* **54:** 1002 – 1013

Moher D, Schulz KF, Altman DG for the CONSORT Group (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomised trials. *Lancet* **357:** 1191 – 1194

O'Quigley J, Stare J (2002) Proportional hazards models with frailties and random effects. *Stat Med* **21:** 3219 – 3233

Robins JM (1995a) An analytic method for randomized trials with informative censoring: Part I. *Lifetime Data Anal* **1:** 241 – 254

Robins JM (1995b) An analytic method for randomized trials with informative censoring: Part II. *Lifetime Data Anal* **1:** 417 – 434

Royston P, Ambler G, Sauerbrei W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* **28:** 964 – 974

Sargent DJ (2001) Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* **91:** 1636 – 1642

Scharfstein D, Robins JM, Eddings W, Rotnitzky A (2001) Inference in randomised studies with informative censoring and discrete time-to-event endpoints. *Biometrics* **57:** 404 – 413

Schmoor C, Schumacher M (1997) Effects of covariate omission and categorization when analysing randomized trials with the Cox model. *Stat Med* **16:** 225 – 237

Schmoor C, Ulm K, Schumacher M (1993) Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Stat Med* **12:** 2351 – 2366

Schwarzer G, Vach W, Schumacher M (2000) On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* **19:** 541 – 561

Silagy C, Lancaster T, Stead L, Mant D, Fowler G (2002) Nicotine replacement therapy for smoking cessation (Cochrane Review). In: *The Cochrane Library*, Issue 4. Oxford: Update Software

Simon R, Altman DG (1994) Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* **69:** 979 – 985

Tai B, Machin D, White I, Gebski V (2001) Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. *Stat Med* **20:** 661 – 684

Therneau TM, Grambsch PM (2000) *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag

Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* **18:** 681 – 694

Wyatt JC, Altman DG (1995) Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ* **311:** 1539 – 1541

Yamaguchi T, Ohashi Y, Matsuyama Y (2002) Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Stat Methods Med Res* **11:** 221 – 236

Department of medical statistics

Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls

Stuart J Pocock, Tim C Clayton, Douglas G Altman

Survival plots of time-to-event data are a key component for reporting results of many clinical trials (and cohort studies). However, mistakes and distortions often arise in the display and interpretation of survival plots. This article aims to highlight such pitfalls and provide recommendations for future practice. Findings are illustrated by topical examples and also based on a survey of recent clinical trial publications in four major journals. Specific issues are: should plots go up or down (we recommend up), how far in time to extend the plot, showing the extent of follow-up, displaying statistical uncertainty by including SEs or CIS, and exercising caution when interpreting the shape of plots and the time-pattern of treatment difference.

In many clinical trials, the primary outcome for comparison of treatments is the time to occurrence of a disease-related event. The most widely adopted method of displaying such results is by means of Kaplan-Meier survival plots, which show the proportion of patients who experience (or do not experience) the event by time since randomisation. The event itself could be death (hence the term "survival plot" is used loosely), but is often time to a non-fatal event (eg, disease recurrence in cancer) and can sometimes be a favourable outcome such as discharge from hospital. Combined endpoints are used increasingly in clinical trials (eg, death, acute myocardial infarction, or cardiac arrest), and in such cases, the survival plot shows the time to the first event.

The statistical methods for producing survival plots and for calculating p values, estimates of treatment effects, and associated CIs are all well documented.[1–3] However, the display and interpretation of survival plots are prey to several potential distortions and deceptions that can make the right message difficult to work out, as reported in a previous survey of survival analyses in cancer trials.[4] In this article, we concentrate on treatment comparisons in clinical trials, although many of the same problems apply to survival plots in general. Our aim is to reveal some of the more common pitfalls and to give some guidelines to authors, journal editors, and readers on what constitutes desirable statistical practice.

As a practical basis for our concerns and conclusions, we identified all 35 clinical trials with survival plots that were published in four general medical journals during July to October, 1999 (19 in *The Lancet*, ten in the *New England Journal of Medicine*, four in the *British Medical Journal*, and two in the *Journal of the American Medical Association*). These trials constituted 41% of the 86 individually randomised parallel-group trials published in the four journals.

Lancet 2002; **359:** 1686–89

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK (Prof S J Pocock PhD, T Clayton MSc)**; Centre for Statistics in Medicine, Oxford, UK** (Prof D G Altman DSc)

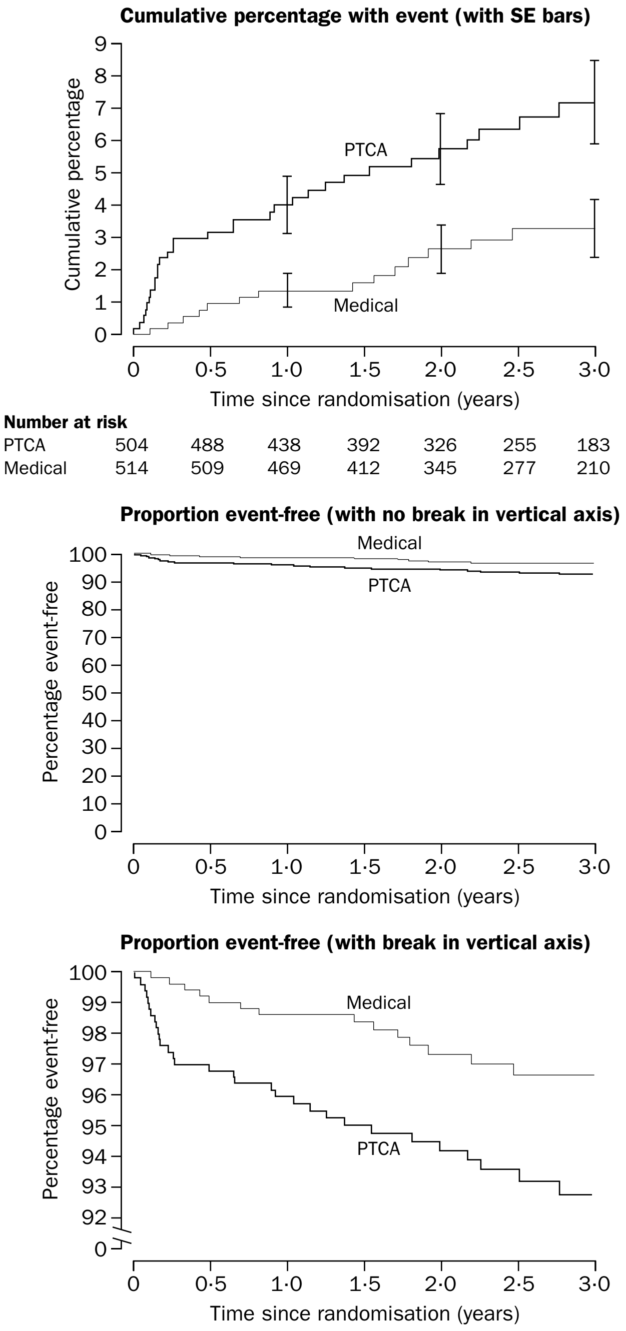
Correspondence to: Prof Stuart J Pocock (e-mail: stuart.pocock@lshtm.ac.uk)

Should plots go up or down?

A survival plot going down displays the proportion of patients free of the event (which of course declines over time), whereas a plot going up shows the cumulative proportion experiencing the event by time. In principle, both contain the same information, but the visual perceptions with regard to comparison of treatment groups can be quite different.

For instance, figure 1 shows three ways of displaying the same data on time to non-fatal myocardial infarction or death in the RITA-2 trial.[5] The first plot, going up, indicates clearly the excess of events in the group randomised to percutaneous transluminal coronary angioplasty (PTCA) compared with the group continuing on medical treatment. This plot has the same style as in the trial's publication,[5] which also gave the numbers and percentages of patients with myocardial infarction or death: 32 of 504 (6·3%) and 17 of 514 (3·3%) for the PTCA and medical treatment groups, respectively (p=0·02). The second plot, going down and using the whole vertical axis from 0 to 100%, makes the difference look much less pronounced (the corresponding proportions event-free being 93·7% and 96·7%, respectively) and mainly emphasises that most patients did not experience the event. The third plot, going down but with a break in the vertical axis seems to fill the space more informatively, but relies on the reader recognising the break in scale: if they do not, the impression is left that PTCA is harmful to a large proportion of patients. Hence having such a break in the scale is not a good style to adopt.

In practice, only one of these options can be displayed in a trial report. We recommend the first option—the plot going up—as the most reliably informative, especially if the event rate is lower than, say, 30%. To maximise the clarity of information, the highest value on the vertical axis should be a round number slightly greater than the highest value represented by the steepest curve—ie, 9% in figure 1. Some might argue that the full scale (0–100%) should always be included, but this inhibits the ability to discriminate between treatments. For instance, the ELITE 2 trial[6] included such a plot, which helped to hide the apparent survival inferiority of losartan compared with captopril. Admittedly, the treatment difference was not statistically significant, but any claim of potential equivalence was perhaps falsely magnified by the choice of survival plot going down over the full 100%



Number at risk

PTCA	504	488	438	392	326	255	183
Medical	514	509	469	412	345	277	210

Figure 1: **Time to non-fatal myocardial infarction or death in RITA-2 trial: three ways to display same data**

scale.[7] The important survival superiority of pravastatin over placebo in the LIPID trial[8] was hard to discern because of this same injudicious choice of survival plot going down over the full 100% scale, since death rates in all groups were, in fact, less than 10% after 5 years. Incidentally, the investigators claim that this choice was introduced by the journal, not the authors themselves. Such plots going down are useful only for trials in which the event rate is high, such as those in cancers with poor prognosis. For instance, for a neuroblastoma trial,[9] the same style of survival plot was perfectly clear, since the median survival was less than 2 years in a study with follow-up over 5 years for those still alive.

The applicability of trial findings should not rely on relative treatment differences alone (eg, proportional reduction in mortality), but must also include absolute treatment differences (eg, number needed to treat per life saved[10,11]). Provision of both survival plots would perhaps be ideal, one going up to reveal the detail and the relative treatment difference, and one going down to clarify the small absolute risk and hence small absolute difference in treatments. In trial reports for which space is at less of a premium and in regulatory submissions, such an approach is to be encouraged for the key outcomes, but it is unrealistic for journal publications.

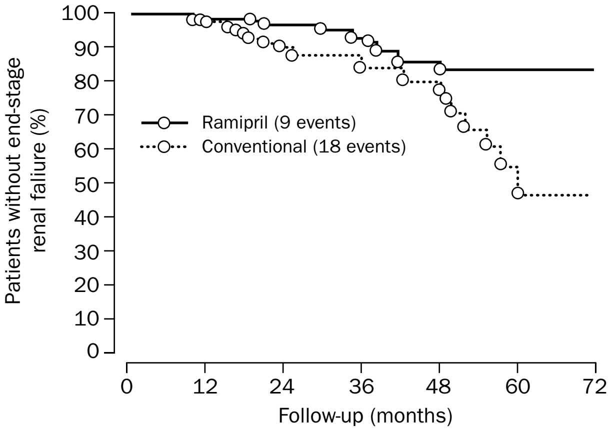


Figure 2: **Kaplan-Meier estimation of renal survival among patients on ramipril or conventional treatment**
Relative risk 2·72 (95% CI 1·22–6·08), p=0·01.

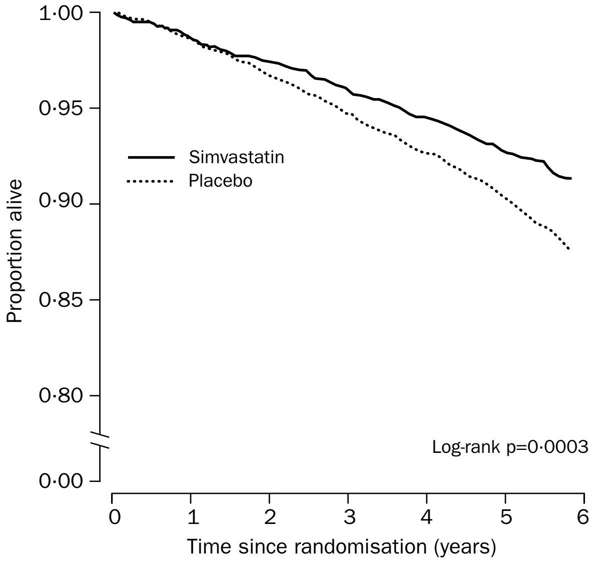
In the 35 trials we surveyed, 12 had plots going up, 15 had plots going down all the way to zero, and eight had plots going down but with a break in scale. This disparity in approach is undesirable.

How far in time to extend the plot?

Follow-up times in any one trial can vary substantially because patients are usually recruited over a long period, and some patients can be lost to follow-up. Length of follow-up is taken into account in the Kaplan-Meier life-table method[1–3] for estimating the proportion of patients who experience an event by time since randomisation. Technically, any survival plot can be extended right through to the longest follow-up time, and five trials we surveyed did just that. However, this extension is not good statistical practice, since for any such plot the eye is drawn to the right (ie, where the plot finishes), which is where there is least information and greatest uncertainty. In small trials, much of the right-hand part of the plot can depict just a few patients.

For instance, figure 2 is a reproduction of the plot of time to end-stage renal failure in a trial comparing ramipril with conventional treatment.[12] The visual impression is that treatments are similar up to 48 months, but thereafter the conventional group develops a striking excess of end-stage renal failures, reaching an estimated 50% failure, by 60 months. However, the median follow-up was 31 months and only 25% of patients assigned conventional treatment reached 48 months' follow-up. The number reaching 60 months is not stated but must be very few. Thus, for both treatment groups, there are inadequate data to estimate reliably the failure rates beyond 48 months' follow-up.

In general, we recommend that survival plots be halted once the proportion of patients free of an event, but still in follow-up, becomes unduly small. In our experience, this view is not universally held, but we hope that our recommendation is a good basis for debate.



Number at risk

Simvastatin	2221	2193	2160	2131	2097	2060	113
Placebo	2223	2193	2152	2103	2059	2011	115

Figure 3: **Kaplan-Meier curves for all-cause mortality in 4S trial**

What constitutes "unduly small" is open to debate and depends on the context. It will often be reasonable to curtail the plot when only around 10–20% are still in follow-up. For example, suppose in a trial of 500 patients, 100 had the event of interest by 2 years of follow-up, but of the remaining 400 patients, only 80 (20%) were still in follow-up beyond 2 years. In this case, restriction of the plot to 2 years' follow-up might be sensible. Such a restriction is just for the plot; all events should be retained in analysis (eg, nine *vs* 18 events in the ramipril trial should remain the basis for the statistical inference given in the legend of figure 2). In this example, the authors' dilemma is clear, since all the "action" happens beyond 48 months. However, were the later follow-up to be included in the plot, it should include a note highlighting the small number of patients on which the data were based. These problems do not arise for trials with an intended fixed length of follow-up (usually quite short), as was the case for 21 of the trials we surveyed.

Showing the extent of follow-up

So, readers need to be informed about the extent of follow-up, and stating the median follow-up time is often useful. Another helpful device is to display the numbers of patients event-free and still in follow-up in each treatment group at relevant time points, as shown in figures 1 and 3. These numbers at risk of the event convey to the reader the increasing unreliability of estimates as time gets further from randomisation; most trials we surveyed included this information. The numbers on the time axis of the published 4S trial plot[13] reproduced in figure 3 show a case for not extending the graph to 6 years. Since only a small minority of patients reached 6 years' follow-up, the apparent extra boost in treatment difference in that last year is less reliably estimated. Incidentally, plots going downwards with an axis break like figure 3 make focusing on the main finding harder. We needed a ruler and calculator to work out that mortality rates after 5 years were 7·4% on simvastatin and 9·7% on placebo.

Displaying statistical uncertainty

Most outcome results of clinical trials include measures of statistical uncertainty—eg, either SEs or CIs—for each treatment group, or a CI for the comparison of groups. However, survival plots often fail to include such measures. Hence the visual impression of any treatment differences, and how they vary over time, can look much more convincing than is really the case, especially if the clinical trial has few outcome events.

For any time since randomisation, the SE (or 95% CI) for the estimated proportion of patients with (or without) the event can be calculated.[2,3] In principle, such error bands could be displayed at all time points for each treatment group, but displaying the SE or 95% CI at a few regularly spaced time points on the plot for each treatment group is clearer. For instance, figure 1 (top panel) shows the SE bars for the estimated event rate for each treatment at 1, 2, and 3 years' follow-up. As is common in such plots, the smaller numbers of patients in follow-up at later time points is reflected in the increasing SE over time.

Although these SEs display each plot's uncertainty, they do not directly display the uncertainty of the treatment difference, which is usually of primary interest. In fact, the SE of the treatment difference in event rates is equal to the square root of the sum of the two squared SEs, but there is no conventionally accepted style (nor any easy way) of displaying this on a survival plot. One simple rule of thumb is that if the treatment difference at a particular time is less than the sum of the two plotted SEs (ie, if the plotted SEs overlap), the difference is well within the bounds of random chance. If the difference is more than twice the sum of the SEs (ie, the 95% CIs do not overlap) it is highly significant. Whether SEs or 95% CIs should be plotted is open to debate, but authors should always make clear which is being used.

One problem here is the focus on the difference between treatments at particular arbitrary time points. The overall evidence of a treatment difference is usually given by the estimated hazard ratio (sometimes called relative risk) and its 95% CI,[14] and by a log-rank test of significance,[2,3] as shown in the legend of figure 2. Thus, an alternative to plotting SEs is to present overall treatment comparisons and their uncertainty on the survival plot or its legend. Most authors do neither, leaving any comment on statistical uncertainty to the text only. In fact, only one of the 35 trial reports we surveyed included CIs at regular time points on the survival plot, five plots included the hazard ratio and its CI, and 16 plots incorporated the log-rank p value.

Summary of recommendations

- Survival plots are best presented going upwards, to maximise detail without needing a break in the scale
- Plots should only be extended through the period of follow-up achieved by a reasonable proportion of participants
- The extent of follow-up should be explained—eg, by listing at regular intervals under the time axis the number still at risk in each treatment group
- Plots should include some measure of statistical uncertainty, otherwise any visual signs of treatment differences might look more convincing than they really are. Either SEs or CIs should be displayed at regular time points, or an overall estimate of treatment difference (eg, relative risk) with its 95% CI should be given
- Authors and readers should be cautious in interpreting the shape of survival plots. The lack of follow-up and poorer estimation to the right-hand end, the lack of any prespecified hypothesis, and the lack of statistical power to explore subtleties of treatment difference other than the overall comparison should be recognised

18 plots included none of the above. We recommend that future authors include in each survival plot some indication of statistical uncertainty (panel).

Interpreting the shape of survival plots

The easiest patterns to interpret are those that show no apparent difference between treatments or when there is a steady divergence between treatments over time. However, in many instances, more complex patterns seem to exist: the treatment difference might look greater early on (figure 1), the divergence between treatments might start later on (figures 2 and 3), or the survival curves might cross. Such putative treatment–time interactions need cautious interpretation since there are rarely sufficient data to consolidate their true existence.

For instance, in the ramipril trial (figure 2), most of ramipril's benefit seems to have occurred late: nine ramipril failures versus 11 conventional treatment failures before 48 months, compared with zero failures versus seven failures, respectively, after 48 months. However, the strength of evidence for this effect is limited, since the number of failures is small, the statistical test for treatment–time interaction is of borderline significance, and such a post-hoc (data-driven) analysis is disputable. So, the overall conclusion needs to rest on events during the total follow-up rather than after any specific time point.

Even for the much larger 4S trial (figure 3), caution is required in interpreting the visual impression that the treatment effect does not occur until after 18 months' follow-up. There seem to have been 55 deaths in each group in the first 18 months, and a striking treatment difference thereafter, with 201 versus 127 deaths favouring simvastatin.[13] A test for treatment–time interaction (ie, of whether the hazard ratio is different before and after 18 months) is significant (p=0·03), but its validity can be questioned because the 18-month time-split for the data has been selected post hoc after seeing the survival plot. Thus, even in such a large trial, to expect reliable estimation of when a treatment effect first begins is unrealistic.[15] Indeed, recent evidence from the Heart Protection Study indicates that there is an observable treatment difference in survival even in early follow-up, which becomes more rapidly divergent beyond 2 years (www.hpsinfo.org).

References

1 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **53:** 457–81.

2 Collett D. Modelling survival data in medical research, section 2.1. London: Chapman and Hall, 1994.

3 Altman DG. Practical statistics for medical research, chapter 13. London: Chapman and Hall, 1991.

4 Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. *Br J Cancer* 1995; **72:** 511–18.

5 RITA-2 Trial Participants. Coronary angioplasty versus medical therapy for angina: the second Randomised Intervention Treatment of Angina (RITA-2) trial. *Lancet* 1997; **350:** 461–68.

6 Pitt B, Poole-Wilson A, Segal R, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomised trial—the Losartan Heart Failure Survival Study ELITE II. *Lancet* 2000; **355:** 1582–87.

7 Hall A. Comparison of losartan and captopril in ELITE II. *Lancet* 2000; **356:** 851.

8 Tonkin AM, Colquhoun D, Emberson J, et al. Effects of pravastatin in 3260 patients with unstable angina: results from the LIPID study. *Lancet* 2000; **356:** 1871–75.

9 Matthay KM, Villablanca JG, Seeger RC, et al. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-*cis*-retinoic acid. *N Engl J Med* 1999; **341:** 1165–73.

10 Altman DG, Anderson PK. Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ* 1999; **319:** 1492–95.

11 Lubsen J, Hoes A, Grobbee D. Implications of trial results: the potentially misleading notions of number needed to treat and average duration of life gained. *Lancet* 2000; **356:** 1757–59.

12 Ruggenenti P, Perna A, Gherardi G, et al. Renoprotective properties of ACE-inhibition in non-diabetic nephropathies with non-nephrotic proteinuria. *Lancet* 2000; **354:** 359–64.

13 Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994; **344:** 1383–89.

14 Altman DG, Machin D, Bryant TN, Gardner MJ, eds. Statistics with confidence, chapter 9. London: BMJ Publishing, 2000.

15 Boutitie F, Gueyffier F, Pocock SJ, Boissel J-P. Assessing treatment-time interaction in clinical trials with time to event data: a meta-analysis of hypertension trials. *Stat Med* 1998; **17:** 2883–903.