Archives of Disease in Childhood 1995; **73**: 374–377

STATISTICS FROM THE INSIDE

17. Survival data

M J R Healy

(1) Introduction and terminology

In certain situations a quantity of interest is the time interval between two events. The statistical theory for handling data of this kind has grown up in the cancer research community where the two events are commonly diagnosis and death. As a result the appropriate methods are often referred to under the heading of *survival data*. It is important to realise that these methods are of much wider applicability. The terminal event may be relapse rather than death, or it may be a desirable outcome such as a manifest response to treatment. None the less, the survival terminology is convenient and I shall use it here.

The statistical analysis of survival data presents two special problems which often make it inappropriate to use the commoner statistical methods. To begin with, the frequency distribution of survival times is usually very far from being Normal, or even of a form transformable to Normality by a simple transformation. Instead the distribution is often extremely skew with a relatively small proportion of survival times being much longer than the rest. Statistical models exist for data of this kind. A very simple distribution which sometimes provides a realistic model for survival times is the exponential distribution, illustrated in fig 1. The distribution has only one parameter, and it has the odd characteristic that the average length of survival for any individual who is still alive is the same no matter how long that individual has survived so far (the average time to a prize for a regular player of the National Lottery has this property). More general distributions such as the Weibull and Gompertz distributions have proved to be applicable to certain types of cancer survival data.

Suppose that the proportion of the population dying before time t is given by a function $F(t)$ (note that what is plotted in fig 1 is the proportion surviving, which is $S(t)=1-F(t)$). Then the proportion of the total which dies between times t_1 and t_2 is $F(t_2)-F(t_1)$. The survivors at time t_1 are a proportion $1-F(t_1)$ of the total and so the proportion of these survivors who die in the interval is

$$\frac{F(t_2)-F(t_1)}{1-F(t_1)}$$

If we divide this proportion by the length of the interval, (t_2-t_1), we get the death rate across the interval. Now let (t_2-t_1) be very small. Then $\{F(t_2)-F(t_1)\}/(t_2-t)$ becomes the slope of the curve of $F(t)$ against t at time t_1 which we write as $f(t_1)$, and the death rate at time t_1 is given by

$$\lambda(t_1)=\frac{f(t_1)}{1-F(t_1)}$$

This 'instantaneous death rate' is called the *hazard*. For an exponential distribution of survival times, the hazard is constant – the death rate is the same for long term as for short term survivors. In other models the hazard may increase or decrease with time. For all causes of death in the general population the hazard decreases sharply in early infancy and then rises steadily into old age.

The second peculiarity of survival time data as they arise in the context of clinical investigations lies in the fact that at the time the data are to be evaluated the terminating event may not have yet occurred for some of the subjects. In a trial of an anticancer treatment, for example, with death as the terminating event, some of the patients may still be alive at the end of the trial. For such cases we do not known the survival time exactly, but we do know that it is greater than the duration of observation so far (this, of course, may differ from one subject to another). Times of this nature are said to be *censored*, and the results of a study involving survival times will typically be a mixture of censored and exactly known times. Conventional methods of analysis cannot handle this kind of data.

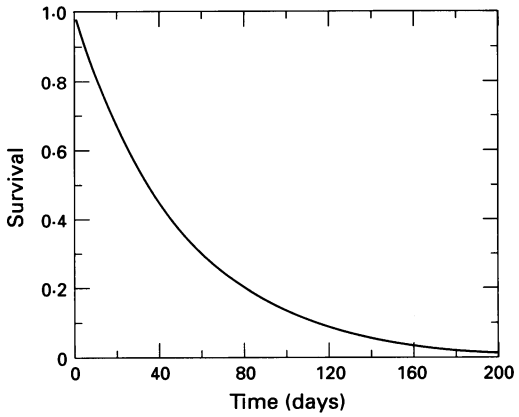

Figure 1 Exponential distribution of survival times.

23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD

Correspondence to: Professor Healy.

No reprints available.

(2) Plotting the survival curve

As usual, the first thing that needs doing when confronted with a sample of survival times is to make some kind of plot of the survival function, $S(t)=1-F(t)$. The usual way of doing this is due to Kaplan and Meier. Suppose the survival times are recorded to the nearest day. Suppose too that at the start of day j there were n_j subjects still alive (and so at risk of dying) and that on that day a number d_j (usually 0 or 1) of them died. Then an estimate of the probability of dying on day j assuming survival up to that time is simply $q_j=d_j/n_j$. The estimated probability of one of the original subjects surviving up to day k can now be found by multiplying together all the quantities $p_j=1-q_j$ for all the days from 0 up to k (we write this as $\prod_{j=0}^{k} p_j$ where the $\prod$ symbol is analogous to the more familiar Σ). Note that most of the d_j will be equal to 0 (only a minority of days will see a death occurring) and for the corresponding days p_j will equal 1. This means that the product can be taken simply over the days on which one or more deaths occur.

A miniature example is shown in tables 1 and 2. The estimated survival curve is shown in fig 2. Note that it consists of horizontal lines with a step as each of the deaths occur. The censored survival times are marked as ticks on the horizontal stretches.

The interpretation of Kaplan-Meier plots does require some experience. In particular, too much attention must not be paid to the right hand ends of the plots where the numbers surviving are small (the example above is obviously too small to carry any weight of interpretation). It is good practice to label the time axis with the numbers surviving at regular intervals.

(3) Comparing two or more groups

If in a clinical trial survival time is used as an outcome, we shall probably want to make a significance test for the difference between the survival curves for the treated and control groups (I assume for simplicity that simple randomisation had been used with no pairing or matching). One apparently simple technique is to compare the proportions surviving at some fixed time. If there is no censoring these will be binomial variates and the comparison is straightforward using standard methods. This technique has the merit of comprehensibility, both to clinician and patient. It has the drawback of wasting much of the available information, and it must not be used if some of the observations are censored. In particular, it is not sufficient when estimating survival at, say, five years simply to ignore subjects whose length of follow up is less than five years. It is obviously invalid to select the time of evaluation after rather than before inspection of the data.

Various significance tests can be devised using some kind of parametric form for the population survival curves, but it is simpler to use a non-parametric approach which makes no assumption about the shapes of the curves. A suitable test is known (for obscure theoretical reasons) as the *log rank* test.

The principle of the log rank test is quite simple. A miniature example is shown in tables 3 and 4 and the two survival curves are plotted in fig 3. The first step is to tabulate the numbers of survivors (n_j) and the number dying (d_j) for every day on which one or more deaths occur in either of the samples. In table 4 on day 7, for example, one death occurred at a time when there were nine survivors in group 1 and eight in group 2. If the null hypothesis that the two survival curves are the same is true, the 'expected' numbers of deaths on that day would be proportional to the numbers of survivors at risk, 9/17 in group 1 and 8/17 in group 2. (I put 'expected' in quotes as under some exceptional circumstances the total 'expected' number of deaths may exceed the total number of the group.) In the same way, on day 87 the numbers at risk have been reduced by deaths and censoring to two in group 1 and four in group 2. There was one death on that day, so the 'expected' numbers were 2/6 and 4/6 respectively. These expected numbers are shown in table 4 as e_1 and e_2. Now form the totals of the observed and expected deaths for each of the groups, calling them O_1, O_2 and E_1, E_2. Then the quantity

$$\frac{(O_1-E_1)^2}{E_1}+\frac{(O_2-E_2)^2}{E_2}$$

Table 1

Survival times (days)								
7	12	15+	28	33+	47+	79	103	116+

The times followed by a + sign are censored.

Table 2 Kaplan-Meier plot calculations for data in table 1

Day j	At risk n_j	Deaths d_j	d_j/n_j q_j	$1-q_j$ p_j	Πp_j
0	9	0	0	1	1·000
7	9	1	0·111	0·889	0·889
12	8	1	0·125	0·875	0·778
28	6	1	0·167	0·833	0·648
79	3	1	0·333	0·667	0·432
103	2	1	0·500	0·500	0·216

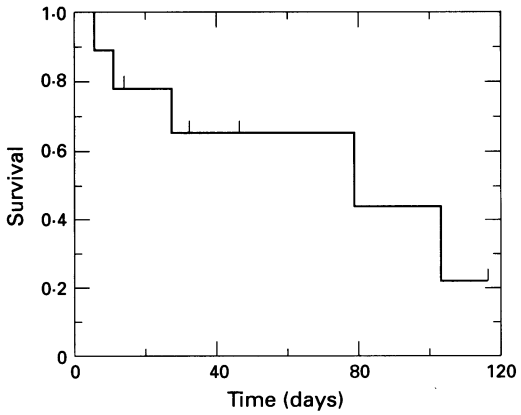

Figure 2 Kaplan-Meier plot of data from table 1.

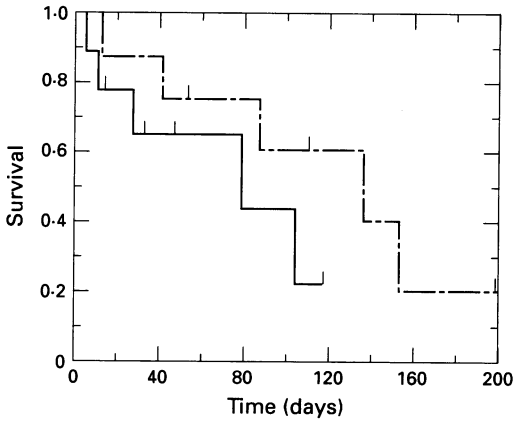

Figure 3 Kaplan-Meier plot of data from table 3.

is approximately a χ^2 with 1 degree of freedom. In the example, $\chi_1^2=1{\cdot}28$, giving $p=0{\cdot}26$ and no convincing evidence for a difference between the survival curves – not surprising with such tiny samples.

The log rank test, like all other statistical techniques, needs to be handled with care. It is designed to be especially sensitive to departures from the null hypothesis in which the two hazards (the 'instantaneous death rates') are in a constant ratio independent of time. It is very insensitive to situations in which the survival curves cross, so that the death rate in one of the groups is initially greater than that in the other but subsequently becomes less. As always, an initial plot of the data is imperative. With this proviso it is a useful technique. The analogy with the standard χ^2 test can be pursued to take in comparisons between three or more groups, including a test for trend if the groups are defined by a quantitative factor.

(4) A regression approach

The log rank test, like most non-parametric methods, provides nothing but a mere significance test. At first glance it seems difficult to know how we can estimate *the* difference between two survival curves when these can differ from each other in an infinite number of ways. A solution was provided by D R Cox who suggested making the assumption mentioned above in connection with the log rank test, that the ratio of the hazards is the same at all time points. This is known as the *proportional hazards* model – note that it does not have to make any assumptions about the actual shapes of the curves, only for the relationship between them. It is often described as a *semi-parametric* model.

Suppose then that we have a number of subjects each of whom is described by the values of one or more covariates x_1, x_2, ... and

Table 3 Survival times (days) for two groups

				Group 1				
7	12	15+	28	33+	47+	79	103	116+
				Group 2				
14	42	54+	87	110+		136	152	198+

The times followed by a + sign are censored.

Table 4 Log rank test calculations for data in table 3

Day j	*At risk*		*Observed deaths*		*'Expected deaths'*	
	n_{1j}	n_{2j}	o_{1j}	o_{2j}	e_{1j}	e_{2j}
7	9	8	1	0	0·529	0·471
12	8	8	1	0	0·500	0·500
14	7	8	0	1	0·467	0·533
28	6	7	1	0	0·462	0·538
42	4	7	0	1	0·364	0·636
79	3	5	1	0	0·375	0·625
87	2	5	0	1	0·286	0·714
103	2	4	1	0	0·333	0·667
136	0	3	0	1	0·000	1·000
152	0	2	0	1	0·000	1·000
Total			5	5	3·316	6·684

$$\chi_1^2=\frac{(5-3{\cdot}316)^2}{3{\cdot}316}+\frac{(5-6{\cdot}684)^2}{6{\cdot}684}=1{\cdot}28$$

a hazard function denoted by $\lambda(t)$. The x's may be quantitative, such as age, blood pressure or carcinoembryonic antigen level, or they may be dummy variables defining groupings such as sex or treatment. If the hazards for all these subjects are proportional, the logs of the hazards will be additive and we can write

$$\log\{\lambda(t)\}=\log\{\lambda_0(t)\}+\beta_1x_1+\beta_2x_2+\ldots+\epsilon$$

where ϵ is an error term and $\lambda_0(t)$ is a baseline hazard applying to a subject all of whose x's are equal to 0. This looks like a regression equation and can in some ways be treated as such. Fitting it involves heavy arithmetic, but suitable computer programs are readily available. The method is called *proportional hazards regression* or *Cox regression* after its originator.

The interpretation of the regression coefficients is reasonably straightforward. If $\beta_1=+0{\cdot}3$ for example, this means that the log hazard for a subject is increased by 0·3 for a unit increase in x_1. The hazard itself is increased by a factor of $e^{+0{\cdot}3}=1{\cdot}35$, in fact by 35%. The fitting procedure produces estimated standard errors for the coefficients and significance tests and confidence limits can be calculated in the usual way.

Proportional hazards regression can be applied to the data in table 3 by introducing a dummy variable which takes the value 0 for all the subjects in group 1 and 1 for all the subjects in group 2. The result is a regression coefficient of $-0{\cdot}88$ (SE 0·74) giving 94% confidence limits of $-2{\cdot}33$ to $+0{\cdot}57$ (the relevant value of t is 1·96, with infinite degrees of freedom). Taking exponentials, the relative hazard is estimated as 0·41 with 95% limits of 0·10 to 1·77. Dividing the coefficient by its standard error gives $t_\infty=1{\cdot}19$. The square of this is 1·41, in fair agreement with the χ^2 value from the log rank test. Both of these techniques assume large samples (without being very clear about how large is large) and the illustrative data that I have used in table 3 are too scanty for the results to be seriously believed.

Proportional hazards regression possesses all the complexities of ordinary multiple regression (for which see the previous two articles in this series) together with others peculiar to itself. The assumption of proportional hazards is obviously very important and adherence to it is one aspect of goodness of fit. There are methods, both arithmetic and

graphical, for investigating goodness of fit, but they are in no way straightforward. Those contemplating the use of the method are strongly advised to collaborate with a statistician who has had extensive experience of it.

(5) Further reading

There are many textbooks dealing with the analysis of survival data, but they are mostly written for statisticians. For clinical readers there are a pair of articles by a group of leading medical statisticians which deserve to be read by anyone contemplating the use of data of this type.[1 2]

1 Peto R, Pike MC, Armitage P, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br J Cancer* 1976; **34:** 585–612.
2 Peto R, Pike MC, Armitage P, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br J Cancer* 1977; **35:** 1–39.

Hirschsprung's enterocolitis

Enterocolitis of uncertain cause is a feared complication of Hirschsprung's disease. It occurred in 57 (34%) of 168 children with Hirschsprung's disease in Michigan (Essam A Elhalaby and colleagues, *Journal of Pediatric Surgery* 1995; **30:** 76–83). In 21 (12·5%) it was the presenting feature of the Hirschsprung's disease. Thirty six children (63%) developed it both before and after operation, 13 (23%) had it only preoperatively, and eight (14%) only after operation. The 57 children had 119 bouts of enterocolitis and the main clinical features were: abdominal distension (93% of episodes), explosive diarrhoea (69%), and vomiting (51%). Fever was a feature of 34% of episodes and lethargy was noted in 27%. Rectal bleeding occurred in six episodes (5%) and colonic perforation in three (2·5%). Thirty one children (54%) had diarrhoea persisting for weeks or months and in them growth failure was common. In most children there was no significant bacterial or viral infection in the stools but enteropathogenic *Escherichia coli* and retrovirus were each found separately in six (7%) episodes.

Plain abdominal *x* ray films were taken during 72 acute episodes and 78 between episodes. Dilatation of colon and/or small bowel and multiple fluid levels were all seen commonly both during and between attacks but the 'intestinal cut-off sign' (sudden cessation of intestinal gas) was present in 74% of the radiographs taken during an attack but only 14% of those taken between attacks. Intramural or intraperitoneal gas was seen on only two occasions. Barium enema appeared to add little of importance to the clinical evaluation.

Management was largely conservative with rectal tube decompression, correction of fluid balance, and antibiotics although 14 children underwent secondary surgical procedures because of recurrent bouts of enterocolitis. There were no deaths directly caused by enterocolitis but nine patients had serious complications; intestinal obstruction (n=5, with gangrenous bowel in two), colonic perforation (n=2), and cardiac arrest ((n=1). The authors attribute their 100% survival to a high degree of suspicion of enterocolitis and early treatment.

ARCHIVIST