STATISTICS IN MEDICINE, VOL. 16, 1041–1074 (1997)

TUTORIAL IN BIOSTATISTICS
SURVIVAL ANALYSIS IN OBSERVATIONAL STUDIES

KATE BULL[1] AND DAVID J. SPIEGELHALTER*[2]

[1] *Cardiothoracic Unit, Hospital for Sick Children, Great Ormond Street, London WC1N 3JH, U.K.*
[2] *MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 2SR, U.K.*

SUMMARY

Multi-centre databases are making an increasing contribution to medical understanding. While the statistical handling of randomized experimental studies is well documented in the medical literature, the analysis of observational studies requires the addressing of additional important issues relating to the timing of entry to the study and the effect of potential explanatory variables not introduced until after that time. A series of analyses is illustrated on a small data set. The influence of single and multiple explanatory variables on the outcome after a fixed time interval and on survival time until a specific event are examined. The analysis of the effect on survival of factors that only come into play during follow-up is then considered. The aim of each analysis, the choice of data used, the essentials of the methodology, the interpretation of the results and the limitations and underlying assumptions are discussed. It is emphasized that, in contrast to randomized studies, the basis for selection and timing of interventions in observational studies is not precisely specified so that attribution of a survival effect to an intervention must be tentative. A glossary of terms is provided.
 Stat. Med., Vol. 16, 1041–1074 (1997).

(No. of Figures: 7 No. of Tables: 12 No. of Refs: 31)

1. INTRODUCTION

1.1. Background

As multi-centre databases become established,[1,2] more reports relating clinical outcomes to risk factors and time are emerging. Such studies may have a variety of objectives: description of the experience of a set of patients; identification of risk factors; prediction on individuals for decision-making, and, increasingly, standardization ('risk stratification') for comparisons between centres or even between operators. Investigators may also wish to draw conclusions about the efficacy of alternative interventions or clinical strategies, although the dangers of making judgements about the benefits of treatment from the analysis of databases have been well argued.[3]

1.2. Structure of this paper

The statistical handling of randomized experimental studies has been well discussed in tutorial papers, particularly the classic articles by Peto and colleagues.[4,5] Good observational and randomized studies have much in common, but there are vital differences. Most important is that a randomized study focuses all attention on estimating the effect of an intervention, and balance between treatment groups with respect to known and unknown explanatory variables is assured,

* Correspondence to: D. J. Spiegelhalter

CCC 0277–6715/97/091041–34$17.50

apart from the play of chance, by the act of randomization. Thus, any major observed differences in outcomes may be attributed to a causal effect of the intervention. In contrast, the basis for selection and timing of interventions in observational studies is not precisely specified and attributing effect to cause must be tentative. Thus, in circumstances where randomized studies are not feasible, good observational studies not only require rigorous attention to the quality of the data but also call for more sophisticated statistical analysis.[6,7] This paper's purpose is to identify some problems in designing and analysing observational studies to increase the likelihood that valid conclusions are drawn, and to illustrate some statistical techniques that have been found helpful. The paper is particularly directed at numerate physicians and surgeons with access to a personal computer and at least one of the many statistical packages available, but also may be useful to statisticians responsible for summarizing time-related outcome data.

Two themes are developed in parallel. First, we may be interested in the occurrence of an event within a *fixed interval*, say 'death within a year of surgery'. Second, we may wish to analyse occurrence of events over a period, say 'pattern of mortality up to the age of 20'. We first deal with simple descriptions of data within these two circumstances. We then introduce a single potential risk factor and subsequently consider multiple, possibly interrelated risk factors. Using a small data set, for each analysis we define its aim, the choice of data to be used, the essentials of the methodology, a computational guide with specific attention to interpreting the output of statistical packages, and finally in a *caveat* section we consider the inferences offered in the light of the limitations of the analysis and the plausibility of its underlying assumptions. A non-technical glossary of terms is provided as an Appendix.

The main novelty in this paper concerns techniques for dealing with occurrences which arise while a study is in progress. First, in Section 4 we introduce the concept of *late entry*; for example, the incorporation of data on a patient who did not present to the hospital until late childhood into a study summarizing events for a class of patients from birth. Second, in Section 9 we consider *time-dependent variables*, in which a subject changes status in some way during follow-up, perhaps by having an operation performed. Finally, for the most determined readers, we show in Section 10 how all these concepts may be combined within a single statistical analysis. However, in discussion we emphasize the tentative nature of the conclusions to be drawn from such analyses.

Though many of the calculations for the examples can be carried out on a hand calculator, the full data set and all of these examples were analysed on a personal computer using the readily available statistical packages SPSS (SPSS Inc, 1992) and EGRET (Epidemiological GRaphics Estimation and Testing package; Statistics and Epidemiology Research Corporation 1991), though several other packages are available which will accomplish most of these analyses. We include an annotated example of the necessary SPSS commands in Appendix I.

We should make clear that this is not a comprehensive review of the appropriate statistical methodology and its limitations. For more detailed expositions on survival analysis (in increasing order of mathematical difficulty) see Healy,[8] Altman,[9] Clayton and Hills,[10] Fisher and van Belle,[11] Cox and Oakes[12] and Andersen *et al.*[13] Large prospective epidemiological studies such as the Framingham Heart Study[14] have made extensive use of analyses such as those described in this paper, while the general issues of bias in observational studies have been covered in texts on clinical epidemiology such as Sackett *et al.*[15]

2. DATA

Perhaps the most vital issue in the analysis of any clinical material is the *integrity* of the data, within which we include the quality, completeness and relevance of the information collected. No

amount of analytic sophistication will rescue a project that does not feature these properties,[16] but here we shall, perhaps naively, take them for granted.

2.1. General problems of bias

We have already stressed the problem of drawing any causal interpretation of associations found in observational studies, but there are other general problems of bias that, although they can occur in randomized trials, are particularly prevalent in observational studies. Two issues are introduced below; other potential biases associated with specific types of analysis are described later.

(a) *Bias which prejudices external generalizability.* The aim of a study will usually be to derive from an available subset of patients, statements about their patterns of survival which will be generalizable to a wider body of patients with the condition. There are many instances where there must be concern that the subset of patients in the analysis are not representative of patients as a whole (reports from institutions attracting 'difficult' caes, older cases, 'correctable' cases etc.). If a patient group whose spectrum of disease is not broadly typical is analysed and the conclusions are to be 'generalizable', factors which make them atypical must be accounted for in the analysis.[7] There is then some hope that patient-specific explanatory variables derived from the skew subset can be applied to other patients.

(b) *Ascertainment bias.* This occurs if the availability of information about a patient's status is dependent on that status. For example, patients may be discharged to the care of referring physicians. If a letter is received reporting the death of a patient, how is this information to be used? If notification is more likely to follow a death than a report that the patient is alive, use of this follow-up information will produce an unfavourable bias. To avoid this bias entirely requires complete ascertainment of status at a point in time.

2.2. Illustrative data

We shall illustrate the analyses using a subset of 30 cases extracted from a larger set of 218 patients with complex pulmonary atresia collected as the basis of an observational study on the presentation and natural history of this disease.[17] Complex pulmonary atresia is a congenital malformation with very abnormal sources of blood supply to the lungs. This particular condition is remarkable for the variability in the age at which patients present to medical attention. Patients were selected for the subset because details of their presentation and history were illustrative for our purposes, so no conclusions about the condition of complex pulmonary atresia can be inferred from these exercises.

The original data collection entailed obtaining dates from the patient record including those of birth, presentation, first operation, death and the date when the patient was last seen. Dates were entered onto a spreadsheet and a date subtraction facility allowed generation of ages at presentation, first operation and death or last contact. Features observable at presentation were defined, obtained and coded and are here exemplified by the size of the intrapericardial pulmonary arteries (`paanat`) and `sex`. The original data with the ages (in days), and appropriate time intervals already prepared, are shown in Table I. For easy reference, patients in Table I have been arranged according to age at presentation. Additional variables have been derived for use in later analyses.

The data are also summarized in Figure 1, which displays the age-interval during which each patient was followed up and the events occurring during this period. For example, we can

Table I. Sample data set

								Derived data							
Patient	agepres	agelast	ageopl	dead	sex	paanat	adfol	follow-up	opfpres	unopage	unop-fpre	pre-opded	hadop	ded-lyrpp	age-presx
1	1	1274	− 1	0	0	0	1	1273	− 1	1274	1273	0	0	0	0
2	2	123	40	1	0	1	1	121	38	40	38	0	1	1	0
3	2	119	− 1	1	1	0	1	117	− 1	119	117	1	0	1	0
4	3	120	− 1	0	1	0	0	117	− 1	120	117	0	0	2	0
5	6	10	− 1	1	0	0	1	4	− 1	10	4	1	0	1	0
6	6	5415	194	0	0	1	1	5409	188	194	188	0	1	0	0
7	7	3261	1041	0	1	0	1	3254	1034	1041	1034	0	1	0	0
8	8	1819	− 1	0	1	0	1	1811	− 1	1819	1811	0	0	0	0
9	11	696	− 1	0	0	0	1	685	− 1	696	685	0	0	0	0
10	13	6415	29	0	1	1	1	6402	16	29	16	0	1	0	0
11	29	3127	144	1	0	0	1	3098	115	144	115	0	1	0	0
12	30	423	47	1	0	0	1	393	17	47	17	0	1	0	0
13	35	5794	− 1	0	1	0	1	5759	− 1	5794	5759	0	0	0	0
14	45	292	62	1	1	1	1	247	17	62	17	0	1	1	0
15	54	68	− 1	1	1	0	1	14	− 1	68	14	1	0	1	0
16	58	1849	− 1	1	0	0	1	1791	− 1	1849	1791	1	0	0	0
17	68	343	− 1	1	1	0	1	275	− 1	343	275	1	0	1	0
18	109	3276	1294	0	0	0	1	3167	1185	1294	1185	0	1	0	0
19	119	207	207	1	0	1	1	88	88	207	88	0	1	1	0
20	121	1430	123	0	1	0	1	1309	2	123	2	0	1	0	0
21	231	308	237	1	0	1	1	77	6	237	6	0	1	1	0
22	258	347	− 1	0	0	0	0	89	− 1	347	89	0	0	2	0
23	349	3355	383	0	0	0	1	3006	34	383	34	0	1	0	0
24	369	3351	2826	1	0	1	1	2982	2457	2826	2457	0	1	0	1
25	437	547	441	0	0	0	0	110	4	441	4	0	1	2	1
26	771	3834	868	0	0	0	1	3063	97	868	97	0	1	0	1
27	1285	7209	− 1	0	1	0	1	5924	− 1	7209	5924	0	0	0	1
28	2455	3555	3532	1	1	1	1	1100	1077	3532	1077	0	1	0	1
29	5161	5354	5353	1	1	1	1	193	192	5353	192	0	1	1	1
30	5497	5639	5633	1	0	1	1	142	136	5633	136	0	1	1	1

Ages and time intervals expressed in days

agepres	age at presentation
agelast	age last seen alive (if alive) or age at death (if dead)
ageopl	age at first operation (− 1 for no operation to date)
dead	no = 0, yes = 1
sex	male = 0, female = 1
paanat	size of intrapericardial pulmonary arteries at presentation: absent or tiny = 0, normal or near normal = 1
adfol	adequate follow-up. Study closed less than 1 year since presentation = 0, study closed at least 1 year since presentation = 1
followup	duration of follow-up (agelast-agepres)
opfpres	interval from presentation to first operation (− 1 if no operation to date)
unopage	follow-up before first operation (ageopl − agepres) if operated, (lastage − agepres) if unoperated
unopfpre	interval from presentation to first operation (if operated) or to age last seen (if unoperated)
preopded	death before any operation: 0 = no, 1 = yes
hadop	had an operation 0 = no, 1 = yes
dedlyrpp	died within 1 year of presentation. 0 = no, 1 = yes, 2 = not applicable (i.e. adfol = 0)
agepresx	age at presentation grouped: 0 = less than 365 days, 1 = older than 365 days

immediately see that patient 1 presented soon after birth with normal size pulmonary arteries (paanat = 1) and is still alive without operation aged 3, while patient 10 had an operation soon after presentation and is still being followed up aged 16. In contrast, patients 29 and 30 did not present until their teens, and then both soon had an operation which they did not survive. Figure 2 shows the identical data, but with elapsed time being measured from presentation rather than birth.

3. SIMPLE DESCRIPTION OF OUTCOMES AT A FIXED TIME INTERVAL

The most straightforward studies concern 'yes/no' outcomes within a fixed time interval after some event; a common example is reporting of early post-operative mortality.

Figure 1. Summary of data showing period of observation of each patient from birth (note the change in scale of the time-axis at 2 years)

3.1. Analysis specification

3.1.1. Inclusion criteria

To be included in such studies, all patients will have had the 'event' referred to (for example, an operation) and in addition, all patients will have been followed (or, if dead, could have been followed) throughout the time interval in question. Examples to be used in our analyses include variables identifying follow-up for at least a year (adfol), and that a patient had an operation at some stage (hadop).

3.1.2. Outcomes (also known as events, responses or dependent variables)

These will include death (perhaps from a particular cause) and possibly intermediate events such as receiving definitive surgery. Examples from our data set include dead and dead within one year of presentation (dedlyrpp).

3.2. Worked example: proportion dying within one year of presentation

We illustrate, using our small data set, the proportion of patients who die within one year of presentation with complex pulmonary atresia (the interval between 0 and 1) in Figure 2. Table II shows the analysis specification using the variable names from Table I.

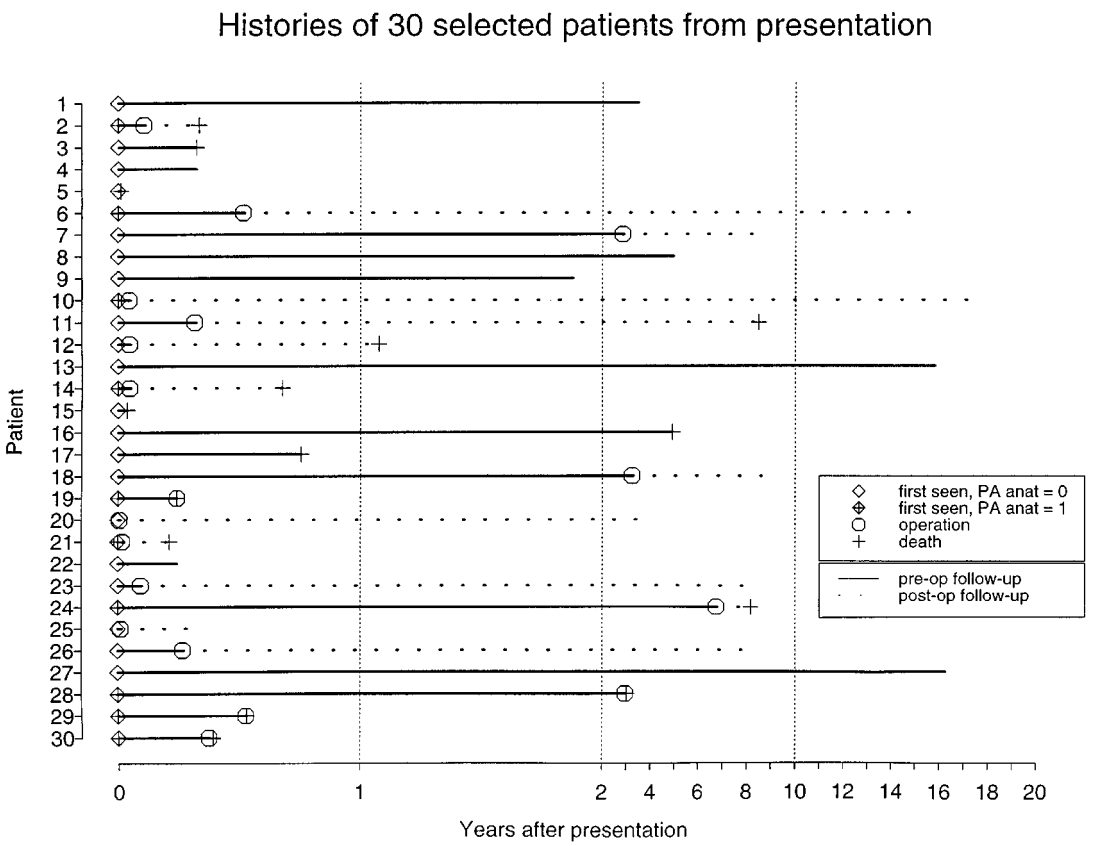

Figure 2. Summary of data showing period of observation of each patient from the time of their presentation (note the change in scale of the time-axis at 2 years)

Table II. Patients dying within one year of presentation

Question:	proportion of patients dying within one year of presentation			
Analysis specification:				
inclusion criteria	patients who have been followed up at least 1 year and patients who (if dead) could have been followed up at least 1 year			`adfol = 1 (dedlyrpp ne 2)`
outcomes	death within 1 year of presentation			`dedlyrpp`
Output:				
Dead	Alive	Total n	Proportion dying p	Odds on dying $p/(1-p)$	95% CL on proportion dying $p \pm 1{\cdot}96\sqrt{[p(1-p)/n]}$
10	17	27	$10/27 = 0{\cdot}37$	$10/17 = 0{\cdot}59$	0·19–0·55

We note that patients 4, 22 and 25 are excluded, being alive but with less than a year of follow-up since presentation (`adfol = 0`); all the deaths were more than a year before the time of analysis (`adfol = 1`) and so could have been followed up for a year had they not died.

Table II also presents simple descriptions of the results. An observed proportion $p =$ (number dying/total number) can take on values 0 to 1; it can be transformed into an odds scale $p/(1-p) =$ (the number dying/the number surviving), which can take on values from 0 to infinity. The odds may appear a somewhat unintuitive measure of risk compared to the proportion of events; however, as we shall see later, the odds provide a basis for handling multiple explanatory variables simultaneously and makes a link to the analysis of full survival data. The 95 per cent confidence limits (95 per cent CL) for the true underlying mortality rate were calculated using the standard normal approximation $p \pm 1{\cdot}96\sqrt{\{p(1-p)/n\}}$: more precise estimates[18] are appropriate for smaller number of events (in particular when no events occur) and may be obtained in some statistical packages.

3.3. Caveats and inferences

Generalizations following this simple analysis depend crucially on the *cohort* (the group of patients being followed up) being representative of the overall class of patients of interest. For example, it may be inappropriate to compare such crude mortality rates between hospitals with different referral populations without using the kind of adjustment techniques to be discussed later.

4. BASIC SURVIVAL ANALYSIS (WITHOUT EXPLANATORY VARIABLES)

4.1. Introduction

The previous analysis is limited in two ways; first, it only considers whether an event has occurred by a particular time, and second, it only includes patients who have been, or could have been, followed throughout that time. In contrast, a survival analysis uses information from the whole follow-up period and all patients can contribute information during their time under surveillance.

Generally, the aim of a survival analysis is to use the data available to provide estimates of the probability of surviving to (or being free of the event in question at) different times, this relationship being expressed as the *survival function*. A graph of the survival function provides the most appealing summary of the time-related information.

Suppose we wish to provide a summary of the pattern of survival of patients with complex pulmonary atresia from the time of presentation. If everyone had been meticulously followed from presentation, and if everyone had presented more than 20 years previously, then estimating the survival function up to age 20 would be trivial: we could simply count the proportion who still survived at each age. However, in practice the more recent patients have not been followed-up for so long and so should only contribute to the estimation of survival up to their current age. Standard analyses, demonstrated below, deal with this problem which is known as *censoring* – the loss to follow-up of patients from causes other than the event of interest.

4.2. Analysis specification

In addition to specifying the *inclusion criteria* and *outcomes*, for a survival analysis we also need to define the following terms.

4.2.1. Time origin

This specifies when the 'clock starts' and derives directly from the question posed. In a randomized study the reference for all follow-up is the point of randomization. In observational studies we might wish to ask questions about survival from 'birth' (or even conception) when analysing

natural history, from 'presentation' when studying an acute illness, or 'operation' when investigating the effects of alternative interventions. Figure 1 illustrates the data with time origin at birth and age along the horizontal axis; Figure 2 presents the same data with the time origin at presentation and with years after presentation on the horizontal axis.

4.2.2. Entry to study

Analysis of a randomized study is straightforward because the point of randomization is clearly both the time-origin of the study and the point of entry of every patient to the study. However, in an observational study, the time origin of the study and the beginning of the period of observation of the patient may not coincide (the patient may come under observation before or after the time origin of the study) and so decisions about what represents 'entry to the study' may require more thought. 'Late entry' describes situations where for some or all patients there is a delay between the time origin of the study (specified by the scientific question posed) and the entry time (limited by the data available). See Section 4.10(b).

4.2.3. Withdrawal from the study (censoring)

There are a number of reasons why follow-up of a subject may cease before the event of interest has occurred, although this will usually be simply due to the selected date for the close of the study being reached. The greatest care is required when the current status of the subject is unavailable, since we need to be able to assume that their loss to the study is unrelated to their underlying risk (an assumption known as *non-informative* censoring), since biased results would arise from systematic withdrawal of either high or low risk patients (see Section 4.10(a)).

4.2.4. Survival time

Specification of the time origin of the study, the outcome of interest and the censoring rules determines the *survival time* within the study for each patient. This is the interval between the time origin and either the occurrence of the outcome or censoring. Examples from our data set include the age when last seen (`agelast`) and the interval between presentation and last contact (`followup`).

4.2.5. Period of observation

Specificatioin of the entry time and the outcome and censoring rules determine the extremes of the *period of observation* within the study for each patient. This is the interval between the entry time and either the occurrence of the outcome or censoring. In situations with late entry, this may be shorter than the survival time.

4.3. Non-parametric survival functions

We shall first illustrate how *censoring* can be accommodated within the Kaplan–Meier (K–M) procedure,[1,2] this is known as a *non-parametric* way of estimating a survival function since it makes no assumptions about the shape of the underlying survival curve (it does not assume that it can be summarized mathematically by a limited number of parameters).

4.4. Worked example: survival from presentation (corresponding to Figure 2)

In this analysis the time origin of the study and the point of entry of every patient to the study is the same, and so survival time and period of observation are identical. This will generally be true

Table III. Estimation of survival from presentation

Question:	survival experience of all patients from presentation	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	presentation	
entry time	presentation	0
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
period of observation	presentation until death or censored	0 to followup

Output:

Patient	Event time	at risk	K–M survival estimate	95% CL on survival estimate	
5	4	30	0·97	0·79	0·99
15	14	29	0·93	0·76	0·98
21	77	28	0·90	0·72	0·97
19	88	27	0·87	0·68	0·95
3	117	24	0·83	0·64	0·93
2	121	22	0·79	0·60	0·90
30	142	21	0·76	0·55	0·88
29	193	20	0·72	0·51	0·85
14	247	19	0·68	0·47	0·82
17	275	18	0·64	0·44	0·79
12	393	17	0·60	0·40	0·76
28	1100	15	0·56	0·36	0·73
16	1791	12	0·52	0·31	0·69
24	2982	10	0·47	0·26	0·64
11	3098	7	0·40	0·20	0·60

for studies of post-operative survival (time origin at operation) or for randomized trials (time-origin at the point of randomization), so a similar analysis will be appropriate in these situations.

Table III shows the analysis specification for this example. The K–M procedure estimates the instantaneous risk of death at any particular time as the ratio of the number who died at that time to the number in the current 'risk set', which is defined to be the set of individuals currently at risk of experiencing the outcome of interest. Hence at the first death (of patient 5) 4 days after presentation, there were 30 in the risk set and hence the risk of death on the 4th day after presentation is estimated to be $1/30 = 0{\cdot}033$. Thus the chance of surviving past 4 days after presentation is estimated to be $1 - 1/30 = 0{\cdot}967$, with 95 per cent confidence limits of 0·79 to 0·99; these limits are best not obtained from an estimated standard error, but from standard formulae[12] that provide assymetric intervals. Fourteen days after presentation, patient 15 dies with a risk set now comprising 29 individuals; the chance of surviving the 14th day after presentation is therefore estimated to be $(1 - 1/29) = 0{\cdot}965$, and thus the estimated cumulative probability of surviving past 14 days becomes $0{\cdot}967 \times 0{\cdot}965 = 0{\cdot}933$ with 95 per cent confidence limits of 0·76 to 0·98, and so on. Results in Table III have been rounded to two figures to reflect the general reporting as 'percentage survival'. Figure 3 displays the estimated survival curve in the conventional 'step' manner.

In mathematical notation, suppose there are r_k subjects in the risk set at the time of the kth distinct time of death t_k, and that at that time there are f_k deaths. Then the estimated survival

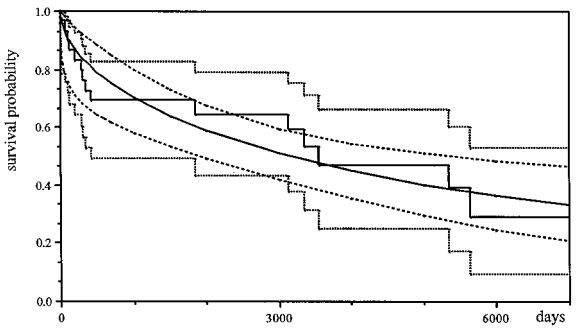

Figure 3. Kaplan–Meier and Weibull estimates of survival from the time of presentation, each with 95 per cent confidence limits

probability until the time t_K is given by

$$p_K = \left(1 - \frac{f_1}{r_1}\right)\left(1 - \frac{f_2}{r_2}\right) \cdots \left(1 - \frac{f_K}{r_K}\right).$$

4.5. Non-parametric survival functions: effect of late entry

If a summary of the survival of patients with complex pulmonary atresia from birth (see Figure 1) is required, a second problem emerges since patients do not come under observation until presentation; this is a general issue in attempting to model the natural history of disease.[19] How we handle *left-truncation* or *late entry* depends on our understanding of the reporting of events for the patients under study. If, for example, we are sure that an event occurring at any point of their life would be reported to us, whether or not the patient was under active follow-up, then we could assume surveillance started at birth and individuals would not enter the cohort late. Such a situation is only plausible in well-defined communities with efficient notification procedures, and as such is rarely an appropriate assumption. Otherwise, avoidance of bias requires that we only include information gathered from patients while they are actively under surveillance.

Because we would usually assume that if an event happened to a patient before they 'presented' we would have been unaware of it, patients should not contribute to our estimate of survival until their age at presentation. An extreme example occurs when we only have information about adult patients – we cannot use them to say anything about survival in childhood. However, just as in right-censoring, we would like to assume *non-informative* late entry,[20] meaning that individuals who present at a certain age are essentially comparable with those of the same age already being followed up; the reasonableness of this assumption is discussed in Section 4.10(a).

4.6. Worked example of survival with late entry: survival from birth (corresponding to Figure 1)

We shall summarize the whole survival experience from birth of all patients with complex pulmonary atresia based on our small selected data set (Figure 1).

Table IV sets out to compare overall survival estimated when all patients are allowed to contribute to the risk set from birth with the curve prepared only allowing patients to contribute to the risk set from presentation; the first is as if a patient's period of observation as illustrated in Figure 1 was extrapolated backwards to birth (appropriate under the optimistic assumption that all events on these patients since birth would have been reported to us). In each case the formula from Section 4.4 is used, with the appropriate size of risk set. Figure 4 plots the estimated survival

Table IV. Survival estimates: entry time 'birth' contrasted to entry time 'presentation'

Question:	whole survival experience of patients in the dataset	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	`dead`
time origin	birth	
entry time	entry: (a) at birth	`0`
	(b) at presentation	`agepres`
censoring rule	withdrawn at end of study	
survival time	birth to death or censored	`agelast`
period of observation	entry to death or censored	(a) `0` to `agelast`
		(b) `agepres` to `agelast`

Output:

Patient	Event time	(a) from birth at risk	(a) from birth K–M	(b) from presentation at risk	(b) from presentation K–M
5	10	30	0·97	8	0·88
15	68	29	0·93	16	0·82
3	119	28	0·90	17	0·77
2	123	26	0·87	16	0·72
19	207	25	0·83	15	0·68
14	292	24	0·80	16	0·63
21	308	23	0·76	15	0·59
17	343	22	0·73	14	0·55
12	423	20	0·69	14	0·51
16	1849	14	0·64	11	0·46
11	3127	13	0·59	11	0·42
42	3351	10	0·53	8	0·37
28	3555	8	0·47	6	0·31
29	5354	6	0·39	5	0·25
30	5639	4	0·29	4	0·18

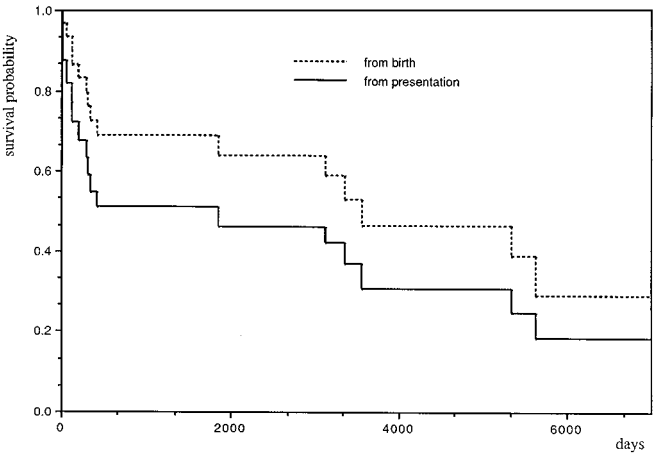

Figure 4. Kaplan–Meier estimates of survival from birth: time origin birth contrasted with time origin at presentation

functions; the survival function assuming entry to the cohort at birth is considerably more optimistic than that generated from the data using only information emerging during the time patients were then being followed up (entry to the cohort at presentation). The pattern of 'drops' reflects the fact that the only difference between the two functions is due to the size of the risk set at each time of death. For example, the first death occurred at age 10 days. We know that 30 children were alive at that age, but only 8 of them were actually being followed-up within this study. The choice of the risk set (with 8 or 30 individuals) determines the size of the denominator in the calculation, and a larger denominator will always decrease the apparent risk.

4.7. Parametric survival functions

Non-parametric techniques use the data to 'draw' the survival function directly; the methodology will conform to any pattern of survival but the survival function proceeds by downward steps which do not reflect the usual perception of an underlying continuity in nature. *Parametric* survival functions reflect both the data and some assumptions about it, including an underlying continuity. The assumptions are embodied in parameters which are themselves estimated from the data; the resulting survival function is thus a mathematical equation which describes a smooth curve. Simple parametric functions include the exponential (in which a single parameter characterizes the death rate which is assumed constant), Weibull (with two parameters allowing the death rate to either increase or decrease with time) or a variety of other forms.[12]

Figure 3 shows a fitted Weibull curve for survival from the time of presentation and its 95 per cent confidence limits (see below for details of this fitted curve). The comparable K–M function is shown with its confidence limits. We note that the confidence limits for the parametric curve are tighter than for the non-parametric curve; this additional precision has been obtained at a cost of greater assumptions which may lead to additional bias. For example, the Weibull curve does not appear to have sufficiently captured the high early mortality followed by the rapid reduction in risk for patients surviving a year from presentation. More complex parametric models are available which adapt to the different survival patterns for early and late mortality;[21] such models have more free parameters which allow greater adaptation to observed patterns and tend to produce more precise estimates than a non-parametric analysis. Such complex parametric models do have the disadvantage that the survival curve at a particular time may be substantially influenced by temporally distant events, and in particular that it may be tempting to extrapolate the curve beyond the region in which the evidence is strong.

4.8. Hazard functions

While survival curves express the *cumulative* effect of the risks faced by an individual, it is often both more convenient and interpretable to work directly in terms of *hazard* at a point in time (the risk of an event occurring per unit of time elapsed, given that the individual has survived up to that time). The *hazard function* expresses the hazard as it changes over time and contains exactly the same information as the survival function but in terms of its rate of change; where the survival curve is falling fast, hazard is high, while if the survival curve is flat the hazard is zero. This equivalence allows the hazard function to be derived by a mathematical transformation of the survival function (see following), as in Figure 5 where a smooth parametric Weibull survival function curve transforms to a smooth hazard function. An identical transformation can be applied to the survival function generated by K–M methodology, though the hazard function then appears as a series of spikes because the K–M survival function falls by discrete steps, and generally has to be smoothed to produce a reasonable plot. In the next section we show how to obtain a slightly different direct estimate of the average hazard within a time window.

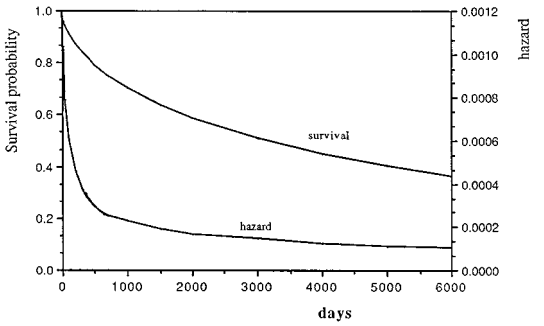

Figure 5. Weibull survival function and hazard function over the first year from presentation

4.9. Computation

Kaplan–Meier estimates are available in most packages, but rarely is adjustment for late entry allowed; exceptions include EGRET. Alternatively, late entry calculations as shown in Table IV could be carried out by hand or on a spreadsheet. Exponential and Weibull parametric models can be fitted in EGRET, but, as is common with most programs for parametric survival analysis, late entry cannot be handled.

If we denote the survival and hazard functions at time t by $S(t)$ and $h(t)$, respectively, then it can be shown that the hazard equals minus the derivative of the natural logarithm of the survival function, or $h(t) = -\mathrm{d}\log S(t)/\mathrm{d}t$. For an exponential distribution, $S(t) = \exp(-bt)$, (where the mean survival is $1/b$) and hence $h(t) = b$ reflecting the equivalent assumption of a constant hazard. For a Weibull distribution we assume $S(t) = \exp(-(bt)^k)$, and hence $h(t) = kb(bt)^{k-1}$, which means that the hazard function may be increasing ($k > 1$) or decreasing with time ($k < 1$). The fitted curve in Figure 3 has parameters $b = 0{\cdot}00021$ and $k = 0{\cdot}46$.

4.10. Inferences and Caveats

(a) *Informative right-censoring.* We have previously mentioned the necessity of assuming that being withdrawn from follow-up is unrelated to the current hazard of death ('non-informative censoring'). This will generally not be a problem if censoring occurs due to the end of the study period, or because of loss to follow-up due to an event unrelated to the underlying risk; events such as emigration or accidental death would usually be considered as non-informative, although this may not be strictly appropriate. In other circumstances censoring may well be informative. Suppose a researcher wished to estimate the 'natural history' of a disease, and censored patients at a definitive operation. This would only be non-informative if patients were selected for operation on the basis of factors unrelated to their current risk of death, which would rarely be the case.

(b) *Informative late entry.* As with censoring, we wish to assume that late arrivals into the study are at equal risk to those already under surveillance ('non-informative late entry'). This will not be a problem if those starting to be followed-up are similar to those already under surveillance, which would occur if, say, new patients were identified through a broadening of the scope of the study. Often, however, new patients will have been referred because of a worsening condition, or, conversely, because they seem a good candidate for a definitive intervention. In either case, age at presentation to a secondary referral institution is likely to be an important predictive variable, possibly being a proxy for the current severity of illness. There is a simple test to examine the independence of time of

entry and survival time,[22] although an alternative means of testing this assumption is by including age at entry into a Cox regression analysis (Section 9) and checking it has no effect upon subsequent risk.

In addition to the standard requirements for generalizability, we now need to address assumptions concerning censoring and late entry. It is apparent that each increment in complexity of analysis, while bringing with it a more realistic representation of the realities underlying the data, require associated judgements on conformity to broad assumptions. Even if we are happy about making such assumptions, it is clear that the simple descriptive analyses shown do not formally explore factors that may influence the outcome; we now need to introduce methods for making comparisons between groups of patients.

5. OUTCOMES AT A FIXED TIME INTERVAL: ONE FACTOR AT A TIME

The simple description of events within a fixed time interval can be readily extended and becomes clinically more useful when the influence of possible explanatory variables can be explored.

5.1. Analysis specification

In randomized trials all explanatory variables (also known as *risk factors*, *covariates*, *predictors*, or *independent variables*) are defined at the point of randomization and are guaranteed (apart from chance variation) to be balanced between treatment groups by the act of randomization. Typically these variables will include age, morphology, clinical status and centre. In the absence of randomization, treatment groups will not be balanced with respect to such variables and hence it is vital that all known explanatory variables are recorded so that the analysis can attempt to adjust for them. Examples from our data set include age at presentation (agepres), gender (sex) and pulmonary artery anatomy (paanat).

5.2. Worked example: deaths within one year of presentation

Here we explore the influence of two variables pulmonary artery anatomy (paanat) and age at presentation, grouped into less than or greater than one year (agepresx) as predictors of death within one year of presentation; both explanatory variables were observable at presentation. In general, variables may be discrete or continuous, although in an exploratory analysis it is generally helpful to group any continuous quantity into discrete categories (such as agepresx); these are generally called *factors*. For each category of factor explored, Table V begins by showing the proportions and percentages dying within a year of presentation.

For each factor explored in this way, a baseline category is identified relative to which comparisons are made. This baseline category is generally the lowest risk or the most common category; here paanat = 0 and agepresx = 0 are used. The odds ratio, relative to baseline, is then calculated for each non-baseline category: these odds ratios are known as *unadjusted* or *simple odds ratios* since they only take into account the association between single factors and outcome. For example, the odds ratio for paanat = 1, relative to paanat = 0, is (6/4)/(4/13) = 4·88.

5.3. Computation

All statistical packages should be able to cross-tabulate a categorical factor against an outcome measure and calculate *p*-values using simple chi-squared tests. Confidence intervals and significance levels for the odds ratio are generally obtainable from statistical packages. In the example there is an excess odds on death, associated with having normal or near-normal pulmonary artery

Table V. Univariate and multivariate analysis of outcomes after a fixed time interval

Question:	predictors of death within one year of presentation	
Analysis specification:		
inclusion criteria	patients who have been followed up at least 1 year *and* patients who (if dead) could have been followed up at least 1 year	adfol = 1 (dedlyrpp ne 2)
outcomes	death within 1 year of presentation	dedlyrpp
explanatory variables	pa anatomy	paanat
	age at presentation (grouped)	agepresx

Output:

Factor	Category	Mortality	%	Odds on death	Univariate analysis			Multivariate analysis		
					odds relative to baseline	95% CL on odds relative to baseline	*p*-value	odds relative to baseline	95% CL on odds relative to baseline	*p*-value
paanat	0	4/17	24%	4/13	1·00			1·00		
	1	6/10	60%	6/4	4·88	0·90–26·42	0·07	6·94	1·01–48·03	0·05
agepresx	0(< 1 year)	8/21	38%	8/13	1·00			1·00		
	1(⩾ 1 year)	2/6	33%	2/4	0·81	0·12–5·50	0·83	0·35	0·04–3·43	0·36
						baseline odds		0·34		

size rather than absent or hypoplastic pulmonary arteries, of 4·88 with 95 per cent confidence interval 0·90 to 26·42; this wide interval just includes 1 and hence we cannot strictly exclude the possibility that pulmonary artery size is not associated with a change in mortality within one year of presentation. This is reflected in the *p*-value of 0·07, which states that there is a 7 per cent chance of observing such an extreme odds ratio even if there were no change in risk. (In general we note that a 95 per cent confidence interval just excluding 1 is essentially the same as a chi-squared test of association rejecting at the 5 per cent level the null hypothesis of no difference from baseline risk.)

5.4. Caveats and inferences

Two concerns with explanatory variables can be identified. First, for results to be generalizable we must be sure that the variables are measured similarly in other contexts; this is easy for precise factors such as agepresx but may be more contentious when subjective judgements about morphology are involved (paanat). Second, looking at one factor at a time can be misleading and we need to consider techniques for examining multiple factors simultaneously (see Section 7.1).

6. SURVIVAL WITH ONE FIXED EXPLANATORY VARIABLE

It is clear that K–M estimated survival functions may easily be calculated for two categories of patient.

6.1. Worked example: survival in different risk groups

Here we consider the example of patients with paanat 0 and 1, using 'presentation' as the time origin. The analysis specification is shown in Table VI, and in the output, the event times correspond with the risk set in Table III broken down into pa anatomy categories. The survival functions are plotted in Figure 6.

Table VI. Example: survival estimates and direct estimates of hazard in the first year and 1–10 years after presentation according to pulmonary artery anatomy

Question:	survival from presentation according to pa anatomy	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	presentation	
entry time	presentation	
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
period of observation	presentation until death or censored	0 to followup
explanatory variables	pa anatomy	paanat

Output:

Patient	Event time	paanat = 0					paanat = 1				
		at risk	events	K–M	instantaneous hazard/day	estimated hazard/year	at risk	events	K–M	instantaneous hazard/day	estimated hazard/year
to 1 year after presentation											
5	4	20	1	0·950	1/20 = 0·050		10	0	1·00		
15	14	19	1	0·900	1/19 = 0·053		10	0	1·00		
21	77	19	0	0·900			10	1	0·900	1/10 = 0·1	
19	88	19	0	0·900			9	1	0·800	1/9 = 0·111	(0·1 + 0·111 + 0·125
3	117	16	1	0·844	1/16 = 0·063	(0·050 + 0·053	9	0	0·800		+ 0·143 + 0·167
2	121	16	0	0·844		+ 0·063 + 0·071)	8	1	0·700	1/8 = 0·125	+ 0·2)
30	142	16	0	0·844		= **0·237**	7	1	0·600	1/7 = 0·143	= **0·846**
29	193	16	0	0·844			6	1	0·500	1/6 = 0·167	
14	247	16	0	0·844		(SE 0·123)	5	1	0·400	1/5 = 0·2	(SE 0·387)
17	275	14	1	0·783	1/14 = 0·071		5	0	0·400		
1 to 10 years after presentation											
12	393	13	1	0·726	1/13 = 0·077		5	0	0·400		
28	1100	13	0	0·726		(0·077 + 0·111	4	1	0·300	1/4 = 0·25	(0·25 + 0·333)/9
16	1791	9	1	0·643	1/9 = 0·111	+ 0·2)/9	4	0	0·300		= **0·065**
24	2982	9	0	0·643		= **0·043**	3	1	0·200	1/3 = 0·333	
11	3098	5	1	0·514	1/5 = 0·2	(SE 0·029)	3	0	0·200		(SE 0·055)

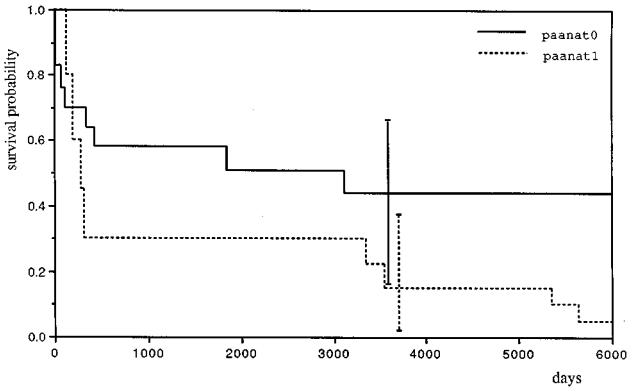

Figure 6. Kaplan–Meier survival estimates according to pulmonary artery anatomy, each with a 95 per cent confidence limit at 3650 days

6.2. Calculation of hazard and its standard error

As described in Section 4.8, general comparison of survival experience may be approached in terms of hazard. Estimates of the instantaneous hazard per day can be derived directly from the observed event rate as in Table VI and from these the average or 'smoothed' hazard over an

interval may be derived. For example, for paanat = 0, we may estimate the instantaneous risk or hazard on the 4th day after presentation, when there was one death from a risk set size 20, to be $1/20 = 0{\cdot}05$. Each subsequent event contributes to an estimate of the hazard at that time, and accumulating these provides an estimated cumulative hazard over a specific period. In this example we have estimated the annual hazard.

The appropriate formulae are provided, for example, by Cox and Oakes,[12] p. 56. Remembering the notation introduced in Section 4.4 (r_k is the size of the risk set in which f_k deaths occur), and consider a time period of length T in which the first K distinct times of death occur. Then the cumulative hazard over this period may be estimated by

$$\frac{f_1}{r_1} + \frac{f_2}{r_2} + \cdots + \frac{f_K}{r_K}$$

with estimated standard error

$$\sqrt{\left\{\frac{f_1}{r_1(r_1 - f_1)} + \frac{f_2}{r_2(r_2 - f_2)} + \cdots + \frac{f_K}{r_K(r_K - f_K)}\right\}}.$$

These formulae can be trivially generalized to any follow-up period, not necessarily starting at time zero.

The average hazard and its standard error may be obtained by dividing each of these quantities by the length of the period T, as shown in Table VI. However, this is a purely descriptive quantity; if it is believed that the hazard was constant over the entire period then an exponential survival distribution should be fitted, with hazard rate estimated by the total number of failures divided by the total follow-up time during the period.

Often the actual hazard in a group is not of primary interest, but attention focuses on the ratio of the hazards between the categories of a factor. In our example the estimated hazard ratio is $0{\cdot}850/0{\cdot}237 = 3{\cdot}59$ during the first year after presentation and $0{\cdot}064/0{\cdot}043 = 1{\cdot}49$ from years 1 to 10 post-presentation. An important question is whether it might be reasonable to assume that the hazard ratio does not depend on the patient's time from presentation, since this would mean we could unambiguously talk of the hazard ratio associated with a particular pulmonary artery anatomy. The assumption that the hazard ratio does not depend on the elapsed time is known as *proportional hazards*, and this rather stringent assumption is fundamental to much of survival analysis.

Though hazard ratios may be estimated directly as in Table VI, in general it is easier to use the Cox regression model described in Section 8, where we also discuss formal tests for proportionality of hazards.

6.3. Comparison of survival between two groups

Comparison between the survival at any chosen time, say 1-year survival, is possible by computing approximate p-values based on the observed survival difference and its estimated standard error. For many other purposes some comparison of the whole survival experience of the two groups will be desirable. This requires the logrank or Cox–Mantel test.

6.4. Worked example: Logrank test for comparing survival between patients with different pa anatomy

The assessment of the statistical significance of the observed difference between K–M survival functions has been discussed in detail by Peto *et al.*,[5] but we show the layout of the data and how these calculations may be carried out in Table VII.

Table VII. Calculation of logrank statistic for comparing survival of groups

Question:		comparison of whole survival experience according to pa anatomy					
Analysis specification:		as Table VI					
Patient	Event time	paanat = 0			paanat = 1		
		at risk	observed	expected	at risk	observed	expected
5	4	20	1	0·667	10	0	0·333
15	14	19	1	0·655	10	0	0·345
21	77	19	0	0·655	10	1	0·345
19	88	19	0	0·678	9	1	0·321
3	117	16	1	0·640	9	0	0·360
2	121	16	0	0·667	8	1	0·333
30	142	16	0	0·696	7	1	0·304
29	193	16	0	0·727	6	1	0·273
14	247	16	0	0·762	5	1	0·238
17	275	14	1	0·737	5	0	0·263
12	393	13	1	0·722	5	0	0·278
28	1100	13	0	0·765	4	1	0·235
16	1791	9	1	0·692	4	0	0·308
24	2982	9	0	0·750	3	1	0·250
11	3098	5	1	0·625	3	0	0·375
			$O_1 = 7$	$E_1 = 10{\cdot}438$		$O_2 = 8$	$E_2 = 4{\cdot}561$

Hazard ratio $(8/4{\cdot}561)/(7/10{\cdot}438) = 2{\cdot}615$

Again we consider survival from presentation for the two groups defined by pulmonary artery anatomy (categories 0 and 1 of paanat). Table VII is similar to Table VI, the *observed* columns indicate the number of deaths in each group at the event time (since there are no 'ties' these are always 1 or 0) and the *expected* columns give the calculated number of deaths that would occur in each group were there no excess risk for either group; for example, at the time of the first death there were 20 at risk with small pulmonary arteries (paanat = 0) and 10 with normal size pulmonary arteries (paanat = 1), so if there were no difference between the two groups we would expect 0·67 of a death in the first and 0·33 of a death in the second group. (It may seem somewhat strange to obtain such fractions of deaths but it is their total that is important.) We then sum the observed and expected columns to give the totals denoted O_1, E_1, O_2, E_2 as shown.

If the two groups had identical risk the expected number of deaths would be close to that observed in each group. In fact there appears to be an excess of deaths in the paanat = 1 group. The statistical significance of this excess can be assessed by calculating a test statistic $\chi^2 = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2$ which will be approximately distributed as a chi-square statistic with 1 degree of freedom under the null hypothesis that the survival functions in the two groups are identical. (This approximation is conservative in that the calculated p-value may be larger than appropriate.[8]) Our statistic is $\chi^2 = 3{\cdot}72$, and consulting standard tables reveals that there is about a 6 per cent chance of observing such an extreme result if the groups had the same survival, so $p = 0{\cdot}06$. Thus there is some, but not overwhelming, evidence of a difference between the groups. An estimate of an overall hazard ratio is given by $(O_2/E_2)/(O_1/E_1) = 2{\cdot}62$.

It is worth noting that the logrank test can be made to accommodate data sets with late entry; for example, just as a breakdown of Table IV(a) according to pa anatomy has provided Tables VI

and VIII, Table IV(b) could provide a comparison of overall survival since birth with entry to the study at presentation.

7. OUTCOMES AT A FIXED TIME: MORE THAN ONE EXPLANATORY FACTOR

7.1. Adjusted odds ratios using logistic regression

Univariate, or *unadjusted*, odds ratios may be misleading if explanatory variables are strongly related to each other; for example, when considering a surgical procedure an apparent association between age and mortality might be explained by the fact that older patients have a more severe form of disease. A possible solution would be to examine whether there is still a relationship with age within each severity category, but with more than a few factors such repeated subdivisions of the data lead to numbers that are too small for meaningful analysis. When explanatory variables are themselves associated, what we are really after is a measure of the association between a factor and the outcome assuming all other measured factors are kept fixed. Logistic regression allows the required *adjusted* odds ratios for multiple factors to be estimated simultaneously, assuming such odds ratios are independent of underlying risk and the values other factors take on.

Table V shows that for our simple example, the adjusted odds ratios are slightly different from the unadjusted. The odds on mortality for a patient with paanat = 1 relative to a patient with small pulmonary arteries (paanat = 0), allowing for age at presentation (agepresx) staying constant, is now 6·94, and the 95 per cent confidence limits for paanat now exclude 1. Table V also shows a 'baseline odds' on mortality for an imaginary patient whose factors are all fixed at their baseline categories; thus a patient with paanat = 0 and agepresx = 0 has estimated odds of 0·34 on death within a year of presentation, which translates to an estimated probability of $0{\cdot}34/1{\cdot}34 = 0{\cdot}25$ or 25 per cent. (Since odds = probability/(1 − probability), we can invert the relationship to give probability = odds/(1 + odds)). By multiplying this baseline odds by the adjusted odds ratios for observed categories of explanatory variables for a specific patient, we may obtain their estimated odds on death.

In notation, we can let d_0 be the baseline odds, and d_i be the odds ratio associated with the observed category of the ith factor. If I factors are taken into account, the final odds is given by

$$\frac{p}{1-p} = d_0 \times d_1 \times \cdots \times d_I.$$

For example, a patient with both normal size pulmonary arteries (paanat = 1) and older age at presentation (agepresx = 1) would have an estimated odds on dying within one year of presentation of $0{\cdot}34 \times 6{\cdot}94 \times 0{\cdot}35 = 0{\cdot}83$, or equivalently an estimated probability of dying in that time frame of $0{\cdot}83/1{\cdot}83 = 0{\cdot}45$ or 45 per cent. The simplicity of this calculation demonstrates why working in odds ratios is advantageous when dealing with multiple explanatory variables.

7.2. Computation

Most packages will handle unadjusted and adjusted odds ratio estimation within a logistic regression framework. Care is required in handling factors with more than two categories; a variable taking on, say, values 0, 1, 2, 3 will be handled as a continuous variable by default with the implication that the odds ratio between category 0 and 1 is the same as that between categories 1 and 2 and so on. If such a specific relationship is not intended, the categorical nature of the variable must be acknowledged for appropriate analysis. If the software allows, the categorical nature can simply be 'declared', otherwise a series of (0, 1) variables, one for each

Table VIII. Logistic regression output in terms of regression coefficients (adjusted only)

Question:		as for Table V							
Analysis specification:		as for Table V							
Alternative output: Factor	Category	B	SE(B)	p-value	95% CL on B		exp(B) (= odds)	95% CL on odds	
					lower	upper		lower	upper
paanat	0						1·00		
	1	1·94	0·99	0·049	$1{\cdot}94-(1{\cdot}96\times0{\cdot}99)$ $=0{\cdot}0004$	$1{\cdot}94+(1{\cdot}96\times0{\cdot}99)$ $=3{\cdot}88$	6·95	1·01	48·4
agepresx	0 (< 1 year)						1·00		
	1 (⩾1 year)	−1·06	1·17	0·36	$-1{\cdot}06-(1{\cdot}96\times1{\cdot}17)$	$-1{\cdot}06+(1{\cdot}96\times1{\cdot}17)$	0·35	0·04	3·42
Constant		−1·08	0·58						

non-baseline category, must be created to allow the effect of each to be compared to baseline. Packages can differ in the way in which comparisons are made between categories of factors (for example, in SPSS for Windows the above standard coding is known as 'indicator'). We note that for dichotomous variables it is convenient to code the categories as 0 and 1, since it is then irrelevant whether the variable is treated as categorical or continuous.

Some packages express the results of a logistic regression in terms of odds ratios and confidence intervals, similar to Table V. Others may only give the results in terms of estimates and standard errors of individual regression coefficients related to the logarithm of the odds on death; these regression coefficients are simply the natural logarithms of the odds ratios. This relationship is demonstrated by taking natural logarithms of the formula in the previous subsection to give

$$\log\left(\frac{p}{1-p}\right) = B_0 + B_1 + \cdots + B_I$$

where $B_0 = \log d_0$ denotes the baseline log-odds, and B_1 to B_I denote the log odds-ratio $\log d_1$ to $\log d_I$. Table VIII denotes the estimate and standard error of any particular coefficient as B and SE(B), giving an approximate 95 per cent interval for the true coefficient of ($B - 1{\cdot}96$ SE(B), $B + 1{\cdot}96$ SE(B)), (since $+/-1{\cdot}96$ standard errors is a 95 per cent confidence interval assuming the estimator is normally distributed). Then the estimated odds ratio and its confidence interval are obtained by taking exponents (anti-log) of the results for B, giving an estimated odds ratio of exp(B) and 95 per cent confidence intervals of exp($B - 1{\cdot}96$ SE(B)) and exp($B + 1{\cdot}96$ SE(B)). The 95 per cent limits for the baseline odds $d_0 = \exp B_0$ may be obtained from the baseline constant in the same way, and apart from rounding errors the results of Table VIII match those of Table V.

We note that we can calculate the estimated probability of any individual surviving one year post-presentation as in Section 7.1, but using the additive coefficients rather than the multiplicative odds ratios. Thus for a patient with normal pulmonary artery size (paanat = 1) and older age at presentation (agepres = 1) we obtain a total B score (the logarithm of the odds on mortality) of $(-1{\cdot}08 + 1{\cdot}94 - 1{\cdot}06) = -0{\cdot}2$, and hence the odds are $e^{-0{\cdot}2} = 0{\cdot}82$ compared to the 0·83 found before. This equivalent result (apart from rounding errors) shows that logistic regression naturally produces a scoring system that can be used for simple risk stratification of patients. In particular, the estimated mortality probabilities for individual patients may be summed to produce an expected mortality within, say, a centre, and then may be contrasted with the observed number of deaths. The resulting comparison will serve as a fairer basis for audit of centres than naive ranking of raw mortality rates, since some adjustment has been carried out for case mix.[23]

Continuous variables are often grouped into categories and hence turned into factors. However, if kept as a continuous quantity and entered into a logistic regression, odds ratios are

interpreted as the change in odds per unit increase in the variable. It is generally useful to subtract a selected 'baseline' value, often the average in the patient sample, in order to retain the interpretation of d_0 as the odds for a baseline patient.

7.3. Caveats

We have steadily elaborated our analyses throughout this paper in an attempt to provide answers to the scientific questions being posed. Such questions may relate to estimating risks, examining associations, predictions on individuals, comparing centres, and even tentatively exploring the causal effects of interventions. The additional power to answer such questions has come through constructing a *model* for the limited data available, which attempts to provide a representation of the underlying mechanisms through making a series of assumptions. We always need to emphasize that a model is never actually *true*, but may be *useful*. The process of model construction, elaboration and criticism is possibly the most vital part of statistical analysis, although the difficulty of formulating strict rules means that it is often left out of standard statistical texts. There is inevitably a strong element of judgement required, and this is best carried out in close collaboration between statisticians and clinicians.

The data may impose limitations on the number of explanatory variables which can be usefully explored simultaneously. Even with many patients available in the database, the main constraint will relate to the number of *events* on which the logistic regression model bases its estimates. A conservative guideline proposed by Harrell *et al.*[24] is to suggest that if there are fewer than 10 times as many *events* to be predicted as there are *explanatory variables* in the model, the p-values associated with the odds related to each variable may be misleading.

In logistic regression it is assumed that odds ratios for the categories of a factor do not depend on the actual categories observed for other factors, but it is possible to specifically include such *interactions* which would allow, for example, the effect of severity of illness to differ according to the age of the patient. However, since there may be many such possible interactions, their selection should largely be based on clinical judgement.[24]

Many packages provide procedures for automatic selection of variables to be included in a model based on stepwise significance testing. Great care is required with the interpretation of the output from these techniques;[24] many significance tests have been done so neither the p-values nor the fact that a variable has been selected or removed should be taken too literally. It is better that variable selection proceeds on the basis of clinical as well as statistical considerations; in particular, the fact that a variable has an odds ratio that is not significantly different from 1 is not in itself a reason to remove it from the model (this would make the error of assuming that the odds ratio really was 1).

Measurement error in explanatory variables is an important consideration; within-individual variability in the measurement will lead to an underestimate of the true odds ratio. This is sometimes known as 'regression dilution bias'. For example, the use of a single diastolic blood pressure measurement leads to a 60 per cent underestimate of the association of diastolic blood pressure with coronary heart disease,[25] compared with the association that exists with an individual's long-term average diastolic blood pressure.

8. SURVIVAL – MANY FIXED FACTORS

8.1. Cox regression using the whole survival experience

Suppose we wish to simultaneously investigate the influence of pulmonary artery anatomy and the gender of the patient. Two survival curves for patients with `paanat` 0 and 1 have been shown

Table IX. Example of Cox regression: factors with potential to influence survival

Question:	factors with potential to influence survival?	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	presentation	
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
entry time	presentation	0
period of observation	presentation to end of follow-up	0 to followup
explanatory variables	pa anatomy	paanat
	gender	sex

Output:

Factor	Category	Univariate analysis				Multivariate analysis			
		Hazard ratio relative to baseline	95% CL on hazard ratio relative to baseline		*p*-value	Hazard ratio relative to baseline	95% CL on hazard ratio relative to baseline		*p*-value
paanat	0	1·00				1			
	1	2·68	0·96	7·42	$p = 0{\cdot}06$	2·69	0·97	7·48	$p = 0{\cdot}06$
sex	0	1·00				1			
	1	0·79	0·28	2·21	$p = 0{\cdot}66$	0·77	0·27	2·17	$p = 0{\cdot}62$

in Figure 6, and, in principle, four curves describing the survival experience of patients with each category of pulmonary artery anatomy in each category of gender could be produced. However, when there are many variables to be explored, the strategy of constantly subdividing the data set to provide comparisons will quickly limit the data available in some subgroups and summarizing the contrasts between many survival curves becomes difficult. In the same way that logistic regression provides a simplifying model that allowed estimation of odds ratios when many factors are being explored at the same time, Cox regression is the technique that provides simultaneous estimates of hazard ratios in the presence of multiple explanatory variables.[12] In logistic regression the odds ratio is assumed independent of the underlying baseline odds, and similarly in Cox regression the hazard ratio is assumed independent of the baseline hazard function, which can be of any form. We may express this by the formula

$$\text{hazard ratio at time } t = h_0(t) \times h_1 \times \cdots \times h_i$$

where $h_0(t)$ is the baseline hazard function at time t, and h_i is the hazard ratio associated with the observed category of the ith factor. If a single factor is entered into a Cox regression then unadjusted hazard ratios may be estimated and p-values calculated; these p-values will be essentially equivalent to those obtained using the logrank procedure shown previously (see Section 8.2).

By way of example, we extend the previous survival analysis in Table VI in order to explore two variables (paanat and sex) with potential to influence the survival function from presentation. Table IX provides the results; we draw attention to the strong similarities to the layout of Table V.

The adjusted hazard ratios may be interpreted as follows. Relative to a baseline patient who has small pulmonary arteries ($\mathrm{paanat} = 0$) and male gender ($\mathrm{sex} = 0$), there is no strong evidence that a similar female patient ($\mathrm{sex} = 1$) has more or less risk, although the width of the confidence interval shows considerable uncertainty as to the true effect. There is some evidence of an increase in risk with larger pulmonary arteries ($\mathrm{paanat} = 1$), with the best estimate being nearly a 2·7-fold excess death rate, but again with great uncertainly around this estimate. This analysis assumes that this excess risk persists throughout follow-up.

Just as with logistic regression, this allows an estimate of the increased hazard associated with any configuration of observed explanatory variables. For example, a patient with both $\mathrm{paanat} = 1$ and $\mathrm{sex} = 1$ would have an estimated hazard ratio of $2{\cdot}69 \times 0{\cdot}77 = 2{\cdot}07$ over a baseline patient with both factors in category 0.

8.2. Computation

Cox proportional hazards survival analysis is now available in many packages; as is the case in logistic regression, categorical variables need appropriate handling and baseline categories are chosen explicitly or by default. The output is also very similar to that of logistic regression; in particular, results are often provided in terms of the actual regression coefficients representing $\log h_i$, that have to be exponentially transformed, just as in Section 7.2, to yield estimates and intervals for the hazard ratios h_i. Usually it is possible to produce estimated survival curves for any selected configuration of explanatory variables, and estimates of the underlying hazard function are generally available.

Somewhat confusingly, *p*-values for individual factors can be obtained by three different methods – 'the likelihood ratio' procedure, the 'score test' and the Wald procedure in which the estimated coefficient divided by its standard error is compared with standard normal tables. Fortunately all three approaches generally give similar answers; our quoted *p*-values are based on the third approach.

8.3. Caveats

Cox regression is known as a *semi-parametric* procedure in which a parametric model for the relative hazard is overlaid on a non-parametric estimate of underlying hazard. With more data it is possible to carry out formal and informal checks of proportional hazards;[26] here we only consider some basic suggestions. One possibility is to divide the follow-up period into a number of *epochs*, corresponding to, say, early, middle and late mortality, and perform a Cox regression analysis separately within epochs. Comparison should then reveal whether estimated hazard ratios depend substantially on the epoch. Alternatively a time-dependent factor (see next section) can be introduced that changes the influence of an explanatory variable according to the epoch; significance of this factor relative to a constant effect would point to non-proportionality. Finally, we note that if non-proportional hazards are suspected for a factor that is not of primary interest, most software will allow this to be specified as a 'stratification factor', which means that separate underlying hazard functions are allowed for each category of that factor.

The term *relative risk* is often used interchangeably both with *hazard ratio* and with *odds ratio* (derived from logistic regression), so perhaps the term is best avoided. Hazard and odds ratios will be different for the same data set, since the odds ratios relate to a particular time while hazard ratios are concerned with the whole survival experience.

The comments in Section 7.3 concerning the dangers of automatic variable selection in logistic regression apply equally to Cox regression.

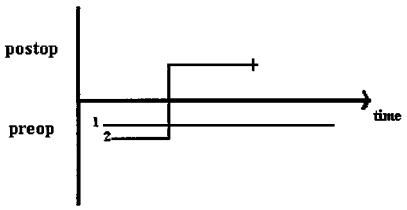

Figure 7. Patient 2 transfers from pre-operative to post-operative risk set at time of first operation

9. SURVIVAL WITH ONE TIME-DEPENDENT FACTOR

(Warning to readers: if earlier sections seemed difficult, perhaps now is the time to turn to the discussion! Section 11)

9.1. Factors which change with time

In randomized studies the intervention of interest is assumed to occur at the point of randomization, and hence the treatment groups are distinguished from the beginning of follow-up. By contrast, in observational studies we may encounter interventions that occur at any point in the period of follow-up, and yet we may be interested in making statements about the effectiveness of the intervention (relative to similar patients who have not had the intervention) in prolonging the time until a specified adverse event.

The risk associated with an intervention can only be assessed subsequent to the intervention – we may therefore consider intervention as a *time-dependent factor*, one for which a patient may change categories over time. In our example, we shall treat the factor *operation* as time-dependent with categories 'pre-op' and 'post-op'. All individuals start off as part of the 'pre-op' risk-set but some move to the 'post-op' risk set at the time of their operation (Figure 7).

Table X incorporates these transitions, showing the number of individuals in each risk set at the time of each death; the table provides the basis for comparing patients of the same elapsed time since presentation who are in the pre- and post-operative risk set. For example, from Figure 1, patient 24 shifts from the pre-op to the post-op risk set between the deaths of patients 16 and 11. In this way, we are comparing at any point the risk of those with the same survival time, who have had and did not yet have an operation. The layout of the table follows those shown previously and we may calculate estimated hazards and provide logrank test statistics as in Tables VI and VII.

The hazard ratio in the first year since presentation is estimated at $0{\cdot}74/0{\cdot}23 = 3{\cdot}2$, while from ages 1 to 10 the estimate is 3·5, suggesting that the proportional hazards assumption is realistic.

9.2. An error to avoid

In wishing to assess the influence of operation, it could be tempting to make a direct comparison between the post-presentation survival of patients who did and did not have an operation. If this approach were followed, with patients divided from entry into a 'no operation' and an 'operative' group, the results would have been somewhat different (Table XI).

The hazard in a table prepared this way would look better for those who had an operation, since their risk sets (the hazard denominator) are systematically inflated by including the pre-operative course of those who were later to have an operation. In this instance, the hazard ratio associated with operation is infancy is wrongly estimated to be $0{\cdot}413/0{\cdot}428 = 0{\cdot}96$, rather than the 3·2 obtained from the appropriate analysis in which patients switch risk sets.

Table X. Generation of pre-op and post-op hazard estimates

Question:	influence of operation on survival from presentation	
Analysis specification:		
inclusion criteria	all patients	
outcome	for pre-operative group:	`preopded = 1`
	for post-operative group:	`dead = 1`
time origin	presentation	
entry time	(a) presentation	0
	(b) at operation	`opfpres`
censoring rule	withdrawn at end of study	
survival time	for pre-operative group:	
	time from presentation until first op or to when last seen	`unopfpre`
	for post-operative group:	
	duration of follow-up	`followup`
period of observation	for pre-operative group:	
	entry: at presentation	0
	withdrawal: at operation or when last seen alive	`unopfpre`
	for post-operative group:	
	entry: at operation	`opfpres`
	withdrawal: when last seen alive	`followup`
explanatory variables (fixed)	none	
explanatory variables (time-dependent)	operation	`hadop`, changing from 0 to 1 at `opfpres`

Output:

Patient	Event time	(a) pre-operative				(b) post-operative			
		at risk	observed	estimated hazard/year	estimated SE(hazard/year)	at risk	observed	estimated hazard/year	estimated SE(hazard/year)
to 1 year post presentation									
5	4	29	1			1	0		
15	14	26	1			3	0		
21	77	20	0			8	1		
19	88	19	0	1/29 + 1/26		8	1	1/8 + 1/8	
3	117	16	1	+1/16+1/11	0·12	8	0	+ 1/8 + 1/8	0·30
2	121	14	0	= **0·23**		8	1	1/9 + 1/8	
30	142	13	0			8	1	= **0·74**	
29	193	11	0			9	1		
14	247	11	0			8	1		
17	275	11	1			7	0		
1–10 years post presentation									
12	393	10	0			7	1		
28	1100	7	0			8	1	(1/7 + 1/8	
16	1791	5	1	(1/5)/9		7	0	+1/8+1/5)/9	0·03
24	2982	2	0	= **0·02**	0·02	8	1	= **0·07**	
11	3098	2	0			5	1		

Table XI. Impact of operation on survival from presentation estimated incorrectly by division into operated and not operated groups (infancy only)

Question:	pre-operative and post-operative survival	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	death = 1
time origin	presentation	
entry time	presentation	0
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
period of observation	from presentation until death or censored censored	0 to followup
explanatory variables	operation	hadop

Output:

Patient	Event time	hadop = 0				hadop = 1			
		at risk	events	hazard/year	estimated SE(hazard/year)	at risk	events	hazard/year	estimated SE(hazard/year)
5	4	12	1			18	0		
15	14	11	1	1/12 + 1/11		18	0	1/18 + 1/17	
21	77	11	0	+ 1/9		18	1	1/15 + 1/14	
19	88	11	0	+ 1/7		17	1	1/13 + 1/12	
3	117	9	1	= **0·428**	0·048	17	0	= **0·413**	0·029
2	121	9	0			15	1		
30	142	9	0			14	1		
29	193	9	0			13	1		
14	247	9	0			12	1		
17	275	7	1			12	0		

The analysis in Table XI may appear obviously incorrect, but early studies of the benefits of heart transplantation took the time of being placed on the waiting list as the time origin and compared the survival from that origin of those who did and did not receive a transplant. Transplantation was shown to be beneficial (as we would expect since a major reason for not obtaining a transplant is early death while on the waiting list), but the errors in this method of evaluation were rapidly made clear in a classic paper.[27]

9.3. Caveats

There is, of course, a great danger in trying to draw inferences about the effectiveness of interventions from non-randomized studies, since patients have been *selected* for the treatment, though the issues surrounding selection are often not clear-cut. In order to have any confidence in the conclusions of such an analysis, we need to understand the main factors that might have influenced the choice of intervention (age, anatomy etc.) and to have explicitly controlled for them in the model; Moses[6] has recently encouraged explicit recording of the *reasons* for intervention. Thus, in order to relate subsequent changes in risk to the intervention, we would need to feel that if two patients were identical in terms of the factors included in the model at the time of intervention, the clinician's decision to intervene on one rather than the other might just as well

have been based on the toss of a coin – we shall term this an assumption of a *non-informative intervention*. Naturally, this is an ideal target, but indicates the need properly to account for severity measures. These may, however, be allowed to change with time using the methodology in this section.

This analysis has not made any attempt to account for factors determining the decision to operate.

10. COMPLEX SURVIVAL ANALYSIS

10.1. Fixed and time dependent factors using Cox modelling with late entry

Time-dependent factors can be examined in a Cox proportional hazards model, which also gives the opportunity to adjust for fixed factors and hence attempt to make the assumption of a non-informative intervention more tenable. Our final example illustrates a means of exploring the influence of operation on the 'natural history' of the disease, adjusting for a fixed risk factor. This brings together many issues demonstrated individually in the preceding sections, including that of late entry, since the time origin is now shifted back to birth and we are estimating age-specific risks. In this example, the effect of operation is 'turned on' at the time of operation, and, in addition, the post-operative phase is divided into three stages: the first 30 day period; the period between one and 6 months, and after 6 months post-operative. The model also incorporates the values of the fixed factor describing pulmonary artery anatomy (paanat), so that inferences about the hazard related to operation (or avoiding operation) can be made 'independent' of the pulmonary artery anatomy.

Table XII shows that, allowing for pulmonary artery anatomy, in the month after operation, the risk of mortality is estimated to be 12 times that of patients of the same age but not operated on. This excess risk decreases dramatically for those who survive one month, but even for those who survive six months there is still a suggestion of continuing increased risk.

10.2. Computation

From a purely technical point of view, this type of analysis requires careful attention in definition of factors and in ensuring the computer programs work correctly. There is a huge increase in the time required for computation when time-dependent covariates are included. The p-values for individual levels come from comparing estimated coefficients divided by their standard errors to standard normal tables.

10.3. Caveats

Aside from the technical problems of such an analysis, great care is required in the interpretation of the output. It is tempting to think of such analyses as obviating the need for randomized trials, since they appear to provide a means of evaluating therapeutic interventions from observational databases, while suitably adjusting for the effect of selection of cases through additional fixed and time-dependent covariates. (See Franklin *et al.*[28,29] for examples of such analyses.)

However, the plausibility of the assumption of a non-informative intervention must always be open to doubt, since it is unlikely one could ever fully control for the clinician's decision to intervene at one time rather than another. Nevertheless, it is possible to imagine a situation where similar patients might be reasonably randomized to immediate or delayed operation. With genuinely similar patients in the pre- and post-operative risk set, the kind of analysis described in

Table XII. Results of Cox model for effect of operation and pulmonary artery anatomy

Question:	influence of operation and pa anatomy on pattern of survival from birth	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead = 1
time origin	birth	
entry time	presentation	agepres
censoring rule	withdrawn at end of study	
survival time	age last seen alive	agelast
period of observation	presentation until survival time	agepres to agelast
explanatory variables	pa anatomy	paanat
(fixed)		
(time-dependent)	operation	hadop, changing from 0 to 1 at ageopl 1 to 2 at ageopl + 30 2 to 3 at ageopl + 180

Output:

Factor	Category	Hazard ratio relative to baseline	95% CI		*p*-value
operation	pre-op	1·00			
	up to 1 month post-op	12·00	1·56	92·69	0·017
	1 to 6 months post-op	1·94	0·28	13·29	0·50
	> 6 months post-op	1·43	0·28	7·02	0·66
paanat	0	1·00			
	1	1·48	0·41	5·28	0·55

this section could then supply an understanding of the role of operation which is difficult to provide with an observational study.

11. DISCUSSION

There is a very reasonable determination to maximize the value and range of inferences that can be drawn from large databases. However, it is clear that even modest inferences can only be drawn at the cost of some assumptions; these assumptions are best made explicit and ideally should be tested. Some aspects, such as the independence of the censoring mechanism, will always be untestable however large the data set, and hence there will always be some reliance on background knowledge and clinical insight.

In contrast to the value placed on the conclusions of a randomized trial, the value placed on the conclusions of an observational study will depend largely on whether all potentially relevant factors have been examined. The onus is on the designers of the observational study to make these factors explicit and ensure that they are adequately represented in the data set. Given such representation, the statistical methods demonstrated here provide some tools for meaningful inter-centre comparison for audit, identification of explanatory variables, prediction on individual cases and so on. We emphasize, however, the range of more sophisticated statistical methods that are becoming available, for example in dealing with recurrent events[30] or adjustment of predictive models for over-fitting to a database.[31]

The proliferation of databases in many branches of medicine and surgery has been partly in the expectation that they would provide a way of examining some management issues which have seemed intractable to the randomized trial approach – whether because numbers of similar patients adequate to support a trial are not available or because an ethical trial is hard to design. We have argued that inferences about how good or bad an investment is afforded by an intervention is particularly difficult to assess when simply observing the outcome of even large numbers of patients. It is here that the intellectual basis of the randomized trial is most potent. However, the careful analysis of databases might crystallize a management problem which is amenable to a randomized trial – perhaps of non-conventional design – for example, randomizing patients to alternative timing of operation. One of the most valuable products of good databases should be the increased potential to design incisive and efficient confirmatory experiments.

APPENDIX I: SPSS COMMANDS FOR ANALYSES

Provides derived variables

```
COMPUTE followup = agelast − agepres.
COMPUTE opfpres = − 1.
IF (ageop1 > 0) opfpres = ageop1 − agepres.
COMPUTE unopage = agelast.
IF (ageop1 > 0) unopage = ageop1.
COMPUTE unopfpre = ageop1 − agepres.
IF (MISSING (ageop1)) unopfpre = followup.
COMPUTE preopded = 0.
IF (MISSING (ageop1) & dead = 1) preopded = 1.
COMPUTE hadop = 1.
IF (MISSING (ageop1)) hadop = 0.
RECODE
  agepres
  (Lowest thru 365 = 0) (366 thru Highest = 1) INTO agepresx.
COMPUTE dedlyrpp = 0.
IF (dead = 1 & followup = 365) dedlyrpp = 1.
IF (adfol = 0) dedlyrpp = 2.
```

Section 3.2 and Table II. Proportion dying within one year of presentation

```
COMPUTE filter_$ = (adfol = 1).
FILTER BY filter_$.
FREQUENCIES
  VARIABLES = dedlyrpp.
FILTER OFF.
```

Section 4.5, Table III and Figure 3. Non-parametric survival from presentation. Corresponding Weibull survival prepared using EGRET

```
KM
  followup  /STATUS = dead(1)  /PRINT TABLE  /PLOT SURVIVAL.
```


Section 4.7. Table IV and Figure 4. Survival function assuming all in risk set from birth. Survival function with late entry generated in EGRET

```
KM
  agelast  /STATUS = dead(1)  /PRINT TABLE  /PLOT SURVIVAL.
```

Section 5.2 and Table V outcome after 1 year. One factor at a time.

```
COMPUTE filter_$ = (adfol = 1).
FILTER BY filter_$.
CROSSTABS
  /TABLES = paanat agepresx BY dedlyrpp
  /FORMAT = AVALUE NOINDEX BOX LABELS TABLES
  /CELLS = COUNT ROW.
```

Section 7.1 and Table V. Outcome after 1 year. More than 1 explanatory variable.

```
LOGISTIC REGRESSION dedlyrpp
  /METHOD = ENTER paanat agepresx
FILTER OFF.
EXECUTE.
```

Section 6.1 and Figure 6. Survival with one fixed explanatory variable

```
KM
  followup  /STRATA = paanat  /STATUS = dead(1)
  /PRINT TABLE
  /PLOT SURVIVAL.
```

Section 8.1 and Table IX. Cox regression using whole survival experience. First one variable at a time, then together.

```
COXREG
  folloup  /STATUS = dead(1)
  /METHOD = ENTER paanat
  /PRINT = CI (95)
  /CRITERIA = ITERATE (20).
COXREG
  followup  /STATUS = dead(1)
  /METHOD = ENTER sex
  /PRINT = CI (95)
  /CRITERIA = ITERATE (20).
COXREG
  followup  /STATUS = dead(1)
  /METHOD = ENTER paanat sex
  /PRINT = CI (95)
  /CRITERIA = ITERATE (20).
```



```
Section 9.1 and Table X, estimating hazard ratio
associated with operation: in this SPSS analysis this
ratio is assumed constant over whole period after presentation.

  First create a logical time-dependent covariate T_COV_
that is 1 if the patient had an operation
AND time-since-presentation > interval-to-operation.

TIME PROGRAM.
  COMPUTE T_COV_ = (hadop = 1) & (T_ > opfpres).
  Fit time-dependent covariate

COXREG
  followup  /STATUS = dead(1)
  /METHOD = ENTER T_COV_
  /ITERATE(20).

Analysis in 10.1, Cox with time dependent factors and late
entry performed using EGRET.
```

APPENDIX II: A NON-TECHNICAL GLOSSARY OF TERMS

Adjusted odds ratio: the odds ratio for one explanatory variable assuming other explanatory variables in the model remain fixed. Derived by logistic regression.
Baseline hazard function: the hazard function for a patient in the baseline category of all the variables entered into, say, a Cox regression analysis.
Baseline odds: the odds on the outcome of interest occurring for a patient in the baseline category of all the variables entered into a logistic regression analysis.
Censoring: withdrawal from the study before the event of interest has occurred, because the study has ended without this event occurring or for other reasons specified in the study design.
Cox regression: this technique deals with outcomes occurring over the whole survival experience and allows the generation of adjusted hazard ratios for multiple factors to be estimated simultaneously; it requires a proportional hazards assumption.
Entry time: the time when a patient starts contributing to the study. In randomized studies or observational studies where all patients have come under observation before the study starts (for example, studies of survival after surgery) the entry time and time origin of the study will be identical. However, for some observational studies, the patient may not start follow-up until after the time origin of the study and these patients contribute to the study group only after their 'late entry'.
Explanatory variables (*also risk factors*, *covariates*, *predictors*, *independent variables*): quantities which may be associated with better or worse outcome.
Factor: an explanatory variable with a limited number of states, possibly a continuous variable which has been divided up into discrete categories.
Hazard function: the instantaneous risk of a patient experiencing a particular event at each specified time.
Hazard ratio: the hazard associated with one category of patient divided by the hazard associated with another category. The hazard ratio can be estimated at an instant or averaged over an interval.

Informative censoring: when withdrawal from the study may not be independent of the current hazard; if patients at higher or lower risk than the rest are withdrawn, this will introduce bias.
Informative late entry: when time of entry is itself a predictor of survival time, perhaps because it reflects severity of the condition concerned additional to that expressed by measured risk factors.
Late entry (*left truncation*): this occurs when patients come under observation after the time origin of the study. In terms of their survival outlook, these patients may or may not be the same as those already in the risk set.
Logistic regression: this technique deals with prediction of outcome at a fixed time interval after the time origin and allows adjusted odds ratios for multiple factors to be estimated simultaneously; it assumes such odds ratios are independent of underlying risk and (unless interaction terms are fitted) the values other factors take on.
Non-informative intervention: an assumption that each patient who underwent an intervention did so for a reason which was not related to their underlying risk or, if it were related, that this relationship can be understood in terms of other associated variables entered into the analysis.
Non-parametric survival function: an estimate of the survival function that depends only on the size of the risk set at the time each event occurs, and hence the graph proceeds by downward steps.
Odds ratio (*unadjusted*, *simple*, *univariate odds ratio*): the odds associated with one category of patient divided by the odds associated with a 'baseline' category of patient.
Odds: a measure of risk defined as $p/(1-p)$, where p is the probability of the event in question.
Outcomes (*events*, *responses or dependent variables*): the endpoint of interest (outcomes dealt with in this paper have all been configured as binary events).
Parametric survival function: an assumption that the survival function is governed by a small number of parameters which are estimated from the data; the graph of the parametric survival function is smooth.
Period of observation: interval between the entry time and the occurrence of the event or censoring.
Proportional hazards: this important assumption is fulfilled if two categories of patient are being compared and their hazard ratio is constant over time (though the instantaneous hazards may vary).
Relative risk: this term can confuse as it sometimes is taken to mean a hazard ratio ('relative risk' over the whole survival experience) and sometimes an odds ratio ('relative risk' over a fixed time interval).
Risk set: the set of patients in the study at a specified time.
Semi-parametric: 'parametric' assumptions may be made about some aspects of a model, while other components may be estimated 'non-parametrically'. In the Cox regression procedure, a parametric model for the relative hazard is overlaid on a non-parametric estimate of baseline hazard.
Survival function: the probability of being free of the event at a specified time.
Survival time: interval between the time origin and the occurrence of the event or censoring.
Time-dependent factor: sometimes factors which come into play after the time origin of the study require consideration because of their possible influence on the probability of the subsequent occurrence of an adverse outcome. To compare the outcomes of patients who have had and who have not yet had this event, two risk sets are compared; patients transfer from one risk set to the other at the time of occurrence of the event of interest.
Time origin: the beginning of the story the study aims at telling. In observational studies, the patients may come under observation before or after the time origin of the study.

ACKNOWLEDGEMENT

Kate Bull is supported by the British Heart Foundation.

REFERENCES

1. Kirlin, J. W., Blackstone, E. H., Tchervenkov, C. I., Casteneda, A. R. and the Congenital Heart Surgeons Society. 'Clinical outcomes after the arterial switch operation for transposition: patient, support, procedural and institutional risk factors', *Circulation*, **86**, 1501–1515 (1992).
2. Hanley, F. L., Sade, R. M., Blackstone, E. H., Kirlin, J. W., Freedom, R. M. and Nanda, N. C., 'Outcomes in neonatal pulmonary atresia and intact ventricular septum', *Journal of Thoracic and Cardiovascular Surgery*, **105**, 406–427 (1993).
3. Byar, D. P., Simon, R. M., Friedewald, W. T. et al. 'Randomised clinical trials: perspectives on some recent ideas', *New England Journal of Medicine*, **295**, 74–80 (1976).
4. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomised clinical trials requiring prolonged observation of each patient. Introduction and design', *British Journal of Cancer* **34**, 585–612 (1976).
5. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomised cinical trials requiring prolonged observation of each patient. Analysis and examples', *British Journal of Cancer*, **35**, 1–39 (1976).
6. Moses, L. E., 'Measuring effects without randomized trials? Options, problems, challenges', *Medical Care* **33**, AS8–14 (1995).
7. D'Agostino, R. B. and Kwan, H. 'Measuring effectiveness. What to expect without a randomized control group', *Medical Care*, **33**, AS95–105 (1995).
8. Healy, M. J. R. 'Survival data', *Archives of Disease in Childhood*, **73**, 374–377 (1995).
9. Altman, D. G. *Practical Statistics for Medical Research*, Chapman and Hall, London 1991.
10. Clayton, D. and Hills, M. *Statistical Models in Epidemiology*, Oxford University Press, Oxford, 1993.
11. Fisher, L. D. and van Belle, G. *Biostatistics: a Methodology for the Health Sciences*, Wiley, New York, 1993.
12. Cox, D. R. and Oakes, D. *Analysis of Survival Data*, Chapman and Hall, London, 1984.
13. Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer, New York, 1993.
14. D'Agostino, R. B., Lee, M-L., Belanger, A. J., Cupples, L. A., Anderson, K. and Kannel, W. B. 'Relation of pooled logistic regression to time-dependent Cox regression analysis: the Framingham Heart Study', *Statistics in Medicine*, **9**, 1501–1516 (1990).
15. Sackett, D. L., Haynes, R. B. and Tugwell, P. *Clinical Epidemiology: a Basic Science for Clinical Medicine*, 2nd edn., Little, Brown, Boston, 1991.
16. Dambrosia, J. M. and Ellenberg, J. H. 'Statistical considerations for a medical data base', *Biometrics*, **36**, 323–332 (1980).
17. Bull, C., Somerville, J., Ty, E. and Spiegelhalter, D. J. 'Presentation and attrition in complex pulmonary atresia', *Journal of the American College of Cardiology*, **25**, 491–499 (1995).
18. Fleiss, J. L. *Statistical Methods for Rates and Prorportions*, 2nd edn, Wiley, New York, 1981.
19. Cnaan, A. and Ryan, L. 'Survival analysis in natural history studies of disease', *Statistics in Medicine*, **8**, 1255–1268 (1989).
20. Keiding, N. 'Independent delayed entry (with discussion)', in Wein, J. P. and Goel, P. K. (eds), *Survival Analysis: State of the Art*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1992, pp. 309–328.
21. Blackstone, E. H., Naftel, D. C., Turner, M. E. 'The decomposition of time varying hazard into phases, each incorporating a separate stream of concomitant information', *Journal of the American Statistical Association*, **81**, 615–624 (1986).
22. Tsai, W. Y. 'Testing the assumption of independence of truncation time and failure time', *Biometrika*, **77**, 169–177 (1990).
23. Rowan, K. M., Kerr, J. H., Major, E., McPherson, K., Short, A. and Vessey, M. P. 'Intensive Care Society's APACHE II study in Britain and Ireland-II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method', *British Medical Journal*, **307**, 977–981 (1993).

24. Harrell, F. E., Lee, K. L. and Mark, D. B. 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, **5**, 361–388 (1996).
25. MacMohan, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A. and Stamler, J. 'Blood pressure, stroke and coronary heart disease. Part 1. Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias', *Lancet*, **355**, 765–774 (1990).
26. Hess, K. R. 'Graphical methods for assessing violations of the proportional hazards assumption in Cox regression', *Statistics in Medicine*, **14**, 1707–1724 (1995).
27. Mantel, N. and Byar, D. P. 'Evaluation of response-time data involving transient states: an illustration using heart transplant data', *Journal of the American Statistical Association*, **69**, 81–86 (1974).
28. Franklin, R. C. G., Spiegelhalter, D. J., Anderson, R. H., Macartney, F., Rossi-Filho, R. I., Douglas, J. M., Rigby, M. L. and Deanfield, J. E. 'Double inlet ventricle presenting in infancy. iii: Outcome and potential for definitive repair', *Journal of Thoracic and Cardiovascular Surgery*, **101**, 924–934 (1991).
29. Franklin, R. C. G., Spiegelhalter, D. J., Sullivan, I. D., Anderson, R. H., Thoele, D. S., Shinebourne, E. A. and Deanfield, J. E. 'Tricuspid atresia presenting in infancy: survival and suitability for the Fontan operation', *Circulation*, **87**, 427–439 (1993).
30. Clayton, D. G. 'Some approaches to the analysis of recurrent event data', *Statistical Methods on Medicial Research*, **3**, 244–262 (1994).
31. van Houwelingen, H. C. and Thorogood, J. 'Construction, validation and updating of a prognostic model for kidney graft survival', *Statistics in Medicine*, **14**, 1999–2008 (1995).

