ELSEVIER

Journal of Statistical Planning and Inference 71 (1998) 271–285

journal of statistical planning and inference

A discrete time sequential process for analyzing censored survival data using the likelihood ratio

A. Maul*

Département Statistique et Traitement Informatique des Données, Institut Universitaire de Technologie, Université de Metz, 57000 Metz, France

Received 5 June 1997; received in revised form 9 February 1998; accepted 9 February 1998

Abstract

A sequential testing procedure for comparing survival distributions with binary responses is considered. The data are monitored according to a discrete time process of reviewing the situation at regularly spaced intervals of time by using the likelihood ratio as a test statistic. Sampling continues until either a decision can be made about the hazard rates characterizing the survival distributions to be compared or a prespecified time limit is reached. Monte Carlo simulations are used to model and estimate the power of the process. More specifically, the critical threshold which allows one to control type-I error at a given level during the whole testing procedure is also determined empirically by simulation. Particular attention is paid to the gain of efficiency resulting from the sequential approach. The better understanding of the relative incidence of the parameters defining the experimental conditions on the power of the process is shown to be helpful in planning a proper experimental design for a wide range of comparative studies (e.g. clinical trials, environmental health studies). Two examples referring to survival data of couples trying to begin a pregnancy and patients with bladder cancer, including environmental or technical factors, are presented as an illustration.

Keywords: Survival analysis; Sequential trials; Likelihood ratio; Censored data; Monte Carlo simulation

1. Introduction

The statistical methods which are used in a great variety of biomedical or environmental health investigations are often dealing with survival data analysis since the principal interest during follow-up concerns the occurrence or non-occurrence of a particular event (e.g. death, local recurrence or any other clinical observation). Several parametric and non-parametric approaches are available to design, and ultimately analyze the results of comparative studies or trials conducted in order to compare several

* Tel.: 33 03 87 31 51 67; fax: 33 03 87 54 71 62; e-mail: maul@iut.univ-metz.fr.

0378-3758/98/$19.00
PII: S0378-3758(98)00086-X

survival distributions with respect to longevity (Bernstein and Lagakos, 1978; Gehan, 1961; Freedman, 1982; George and Desu, 1974; Peto et al., 1977).

In most of the conventional statistical procedures the duration of the study is predetermined and the total number of individuals required is calculated so as to ensure a sufficient number of events to be observed. This is done in order to provide adequate power to the comparative test at the scheduled termination of the trial. Following this principle, a number of tables and monograms have been published (Freedman, 1982; Shuster, 1993; Casagrande et al., 1978; Cochran and Cox, 1957; Machin and Campbell, 1987; Peto et al., 1977; Schoenfeld and Richter, 1982) for planning the size and duration of comparative tests and clinical trials.

Most of the previous methods emphasize the sample size, though the duration of the study should also be considered an aspect of major importance when designing a comparative test in many situations which are encountered in practice. On the one hand, inferences about the parameters of interest or the power of the tests of comparison obviously tend to improve as the available information which is accumulated over time increases. On the other hand, considering the possible implications for public health intervention, prevention policy or ethical aspects, it is desirable that the outcome of the study be stated within the shortest possible period of time. A compromise between the two previous antagonistic time constraints can be reached by reviewing the situation in a sequential way at successive chronological terms. However, a crucial problem, when performing interim analyses, is to control the type-I error rate at some prespecified level of significance.

The purpose of the present paper is to outline a sequential procedure to perform comparative analysis of survival distributions. It should be pointed out that sequential analysis of survival data with prolonged observation of each individual has been proposed by Whitehead (1992), among others. Several methods to derive exact stopping boundaries for group sequential clinical trials were developed recently (Mehta et al., 1994; Lin et al., 1991; Pawitan and Hallstrom, 1990). These methods are based on the exact joint permutation distribution of rank statistics observed across all the monitoring times. Lan and Zucker (1993) present a unified conceptual framework for sequential monitoring covering a wider variety of clinical settings. Reviews of some currently used sequential methods in clinical trials are given by Fleming and DeMets (1993) and Lee (1994).

The method suggested here introduces a new dimension of flexibility into the analysis of survival data within the framework of a stepwise assessment by taking explicit account of the discrete nature of the data. The method is based on a discrete time expression for the hazard while examining the different study subjects individually. It is therefore appropriate to deal with data sets which may contain a large number of censored values and/or tied failure times as a result of the sequential process of reviewing the situation when carrying out a follow-up study with staggered entries. Moreover, the critical threshold in the testing procedure is determined so as to maintain type-I error risk under a prespecified level during the whole experiment. To this end, the stopping criterion and decision rule of the sequential procedure are based on the

distribution of the likelihood ratio statistic under the null hypothesis. In practice, the critical threshold of the test or the parameters of interest (e.g. duration and power of the process, mean number of events, proportion of inconclusive trials) at termination of the sequential procedure can be estimated by Monte Carlo simulations which are performed for a number of different experimental designs. Conversely, the results of this empirical approach for estimating the parameters are useful to define appropriate designs, in the sense that they will provide adequate power to detect a given difference in the hazard rates associated with the groups to be compared. This is to be done within the shortest possible period of time.

The different aspects of the method presented here are illustrated by means of two numerical examples from the survival literature. These examples are concerned with: (i) studying the effect of smoking on the fecundability of couples trying to begin a pregnancy, and (ii) a trial on superficial bladder cancer.

2. Statistical methods

2.1. The model

Let us consider g groups of individuals and let p_{ij}^t be the hazard rate corresponding to the jth individual ($j=1,\ldots,n_i$) of the ith group ($i=1,\ldots,g$) at time t. Note that the hazard is given as a discrete expression of the time i.e., the hazard function is supported on the integers ($t=1,2,\ldots$). Clearly, p_{ij}^t is the probability for individual j in group i to fail at time t provided that the individual was still at risk at time $t-1$. The random variable associated with the survival corresponding to this individual is denoted as Y_{ij}. If Y_{ij} is censored on one side, the observed survival time of the corresponding individual will be denoted as y_{ij}^c. Note that both right or left single censoring may be considered hereafter.

As a preliminary approach, the further development will be focussed on the special case which arises when the hazard rate characterizing all the individuals belonging to the same group is time-fixed, i.e. $p_{ij}^t = p_i$ ($t=1,2,\ldots;j=1,\ldots,n_i$). In other words, this means that the survival times are exponentially distributed. The exponential model is often considered an appropriate underlying hypothesis and it is therefore commonly used in survival data analysis (Bernstein and Lagakos, 1978; Maul, 1994; Prentice, 1973; Schoenfeld and Richter, 1982). Assessment of the adequacy of the exponential model is easy to perform on partially censored data by using the cumulative hazard function corresponding to each lifetime (Maul, 1994; Lawless, 1982). Hence, the further developed technique will be referred to as discrete time Bernoulli likelihood ratio sequential (DTBLRS) method.

We have

$$\begin{aligned} &\Pr(Y_{ij}=y_{ij})=(1-p_i)^{y_{ij}-1}p_i,\\ &\Pr(Y_{ij}>y_{ij}^c)=(1-p_i)^{y_{ij}^c}, \quad (i=1,\ldots,g;\ j=1,\ldots,n_i;\ y_{ij} \text{ or } y_{ij}^c=1,2,\ldots). \end{aligned} \tag{1}$$

However, in order to be of practical use (e.g. randomized controlled clinical trials, environmental health or biomedical follow-up studies) the results obtained for the DTBLRS method are shown to be applicable to much less restrictive situations. In particular, assessing the critical threshold and choosing a sampling scheme so as to control type-I error under a prespecified level is applicable, more generally, to situations involving: (i) staggered entries, that is when the subjects arrive in sequence, (ii) any type of single censoring, and/or (iii) time-dependent hazard rates.

2.2. Sequential comparison of survival distributions

2.2.1. Test statistic

Testing the equality of the g survival distributions under the underlying assumption of an exponential distribution of survival times can be performed by testing the equality of the g hazard rates in $\underline{p}=(p_1,\ldots,p_g)$, i.e. $H_0: p_1=p_2=\cdots=p_g$.

The observed survival times which are available for group i $(i=1,\ldots,g)$ at stage t $(t=1,2,\ldots)$ in the sequential procedure are: $\underline{y}_i^t=(y_{i1},\ldots,y_{i,n_i-k_i^t},y_{i,n_i-k_i^t+1}^t,\ldots,y_{i,n_i}^t)$ where the last k_i^t observations are censored (i.e. individuals still at risk at time t). Note that t is then taken as the censoring value.

The likelihood function of the sample $\underline{y}^t=(\underline{y}_1^t,\ldots,\underline{y}_g^t)$ at stage t is given by

$$L(\underline{y}^t\mid\underline{p})=\prod_{i=1}^{g}\left\{\prod_{j=1}^{n_i-k_i^t}[(1-p_i)^{y_{ij}-1}p_i]\prod_{j=n_i-k_i^t+1}^{n_i}(1-p_i)^{y_{ij}^t}\right\}. \tag{2}$$

The test statistic, $-2\ln\Lambda^t$, calculated at stage t of the sequential process is expressed in terms of the ln likelihood ratio. Hence,

$$-2\ln\Lambda^t=2\sum_{i=1}^{g}\left\{\ln\left(\frac{1-\hat{p}_i^t}{1-\hat{p}^t}\right)\sum_{j=1}^{n_i}y_{ij}^*+(n_i-k_i^t)\ln\left[\frac{\hat{p}_i^t(1-\hat{p}^t)}{\hat{p}^t(1-\hat{p}_i^t)}\right]\right\}, \tag{3}$$

where y_{ij}^* is for y_{ij}^t or y_{ij} according to whether Y_{ij} has been censored at y_{ij}^t or not, respectively. In Eq. (3), $\hat{p}_i^t$ is the maximum-likelihood estimate of p_i under H_1 whereas $\hat{p}^t$ is the maximum-likelihood estimate of the common hazard rate under H_0. Note that if the number of events in group i is 0 then the contribution of group i to the likelihood is $2\ln(1/(1-\hat{p}^t))\sum_{j=1}^{n_i}y_{ij}^*$.

2.2.2. Critical threshold

Performing the test in a sequential way requires one to adjust the critical threshold so as to control type-I error at a specific level during the entire procedure. In this regard, we address settings in which the data are monitored continuously in the sense that t is increased by one at each step of the process up to some predetermined practical limit $t_{\max}$. Moreover, let D be the random variable in the sequential procedure associated with the smallest integer t such that $-2\ln\Lambda^t$ is significant, regarding a given type-I error rate, α, and provided that $t\leqslant t_{\max}$. This means that D is the duration of the sequential procedure under the condition it does not exceed $t_{\max}$. The critical threshold

will be denoted by $s(n_1,\ldots,n_g,\alpha,t_{\max},H_0)$ since it depends on the sample size of the different groups to be compared, the level of type-I error, the practical time limit and the level of the common hazard rate as stated in H_0. Its value is determined so as to fulfill the following constraint:

$$\alpha = \Pr(D \leqslant t_{\max} \mid H_0) = \Pr(-2\ln \Lambda_{\max} > s(n_1,\ldots,n_g,\alpha,t_{\max},H_0) \mid H_0), \tag{4}$$

where $-2\ln\Lambda_{\max}$ represents the maximum value of $-2\ln\Lambda$ which is observed on the discrete time interval $[1,\ldots,t_{\max}]$.

2.2.3. Stopping criterion

The stopping boundaries are derived from the exact distribution of $-2\ln\Lambda_{\max}$ observed across all the monitoring times, given $n_1,\ldots,n_g,\alpha,t_{\max}$ and H_0. The decision rule at stage t $(t=1,2,\ldots,t_{\max})$ is as follows:

- if $-2\ln\Lambda^t \leqslant s(n_1,\ldots,n_g,\alpha,t_{\max},H_0)$ and $t \leqslant t_{\max}$, then continue;
- if $-2\ln\Lambda^t > s(n_1,\ldots,n_g,\alpha,t_{\max},H_0)$ and $t \leqslant t_{\max}$, stop and reject H_0 (i.e. the hazard rates are not equal). Then, the value taken by D is t;
- the trial is declared inconclusive if $-2\ln\Lambda^t$ has not reached a significant value by $t = t_{\max}$.

Thus, α may be considered the probability of observing a significant outcome before the practical limit, $t_{\max}$, of the sequential procedure is reached, provided H_0 is true. In fact, the value of α is larger than the actual type-I error risk in case a decision is made and the sequential process is stopped before $t_{\max}$. In practical terms, this is an error in the right direction since our approach leads to more conservative tests.

2.2.4. Simulation procedure

Although the results presented in this work can be easily generalized to any number of groups by following the simulation procedure which is outlined hereafter, the scope of this paper will be focussed on the comparison of two survival distributions only. The objective is to determine whether one treatment is better than the other with respect to longevity. Monte Carlo simulations were carried out to assess:

(1) the critical threshold corresponding to a significant outcome of the comparative test;

(2) the incidence of both the parameters of the experimental design and type-I error rate on the duration and the power of the DTBLRS method.

The curves and surfaces in this paper were estimated empirically by simulation. The graphs have been interpolated between observations for convenience in plotting. Each point estimate which was used for the construction of the curves corresponds to $r = 5000$ or $10\,000$ distinct sets of simulation runs. Power analysis showed that the power of the sequential testing procedure reached its maximum level in the case of a balanced allocation of the individuals in the treatment groups to be compared. All the sequential comparative tests performed in this paper are therefore based on a balanced randomization of the n individuals in the groups. The survival of the individuals at

risk within each simulated trial was examined at each step of the sequential procedure by following a Bernoulli random process. The parameter of the Bernoulli distribution was set equal to the hazard rate characterizing the group to which the individual under consideration belonged.

3. Results and discussion

Setting the value of $s(n_1,\ldots,n_g,\alpha,t_{\max},H_0)$ at a level which is appropriate to ensure a type-I error risk equal to α amounts to determining the quantile of order $1-\alpha$ in the distribution of $-2\ln\Lambda_{\max}$ for any combination of $t_{\max}$ and hazard rates as in H_0. Stating the null hypothesis as $p_1=p_2=p$, several sets of $r=5000$ trials were generated for a wide range of values of p and n. The maximum value of $-2\ln\Lambda$, observed on the discrete time interval $[1,\ldots,t_{\max}]$, was computed for each trial.

The results of these simulations showed that the distribution of $-2\ln\Lambda_{\max}$ depends on $t_{\max}$ and H_0, only. In particular, it should be emphasized that, under H_0, the distribution of $-2\ln\Lambda_{\max}$ may be considered approximately independent of the sample size n, provided $np\geqslant 1$. This condition is easily fulfilled in consideration of the range of values of n which is currently used in practice. Fig. 1 displays the variation of the critical threshold (i.e. the appropriate quantile in the distribution of $-2\ln\Lambda_{\max}$) in a three-dimensional plot. The surface in Fig. 1a shows the variation of the quantile at level 0.95 in the empirical distribution of $-2\ln\Lambda_{\max}$ under H_0, as a function of $t_{\max}$ and p. The visual information on the 3-D scatterplot shown in Fig. 1a can be greatly enhanced (see Fig. 1b) by using a multivariate smoothing procedure such as locally weighted regression (Cleveland and Devlin, 1988). As was to be expected, the higher the value of $t_{\max}$, the higher the quantiles. Moreover, it is interesting to note that the surface shown in Fig. 1a and b is relatively flat provided that $t_{\max}$ and p are not too close to zero. This property concerning the general shape of the surface applies to any other quantile. It is also important to note that the quantiles and the common hazard rate in the null hypothesis vary in the inverse order. Clearly, if $p<p'$ then $s(n_1,n_2,\alpha,t_{\max},H_0(p))>s(n_1,n_2,\alpha,t_{\max},H_0(p'))$. Thus, in consideration of all these remarks it is always possible to set the value of the critical threshold so as to control type-I error under any prespecified level during the whole study. This is actually one of the major benefits of the procedure presented in the present work.

The conditional expectation of the duration time in a balanced experiment is presented in Fig. 2, which is concerned with studying the variations of $E[D\,|\,D\leqslant t_{\max}]$, $(t_{\max}=200)$, as a function of the sample size n. Taking $p_1=0.01$, the different curves represented in Fig. 2 correspond to survival rates in the second group fixed at $p_2=0.015$, $p_2=0.02$, $p_2=0.03$, $p_2=0.04$ and $p_2=0.05$. From Fig. 2 it is clear that $E[D\,|\,D\leqslant t_{\max}]$ decreases as (a) the hazard ratio p_2/p_1 increasingly diverges from one, or (b) the number of individuals involved in the study increases. A similar statement can be made if type-I error rate becomes larger.

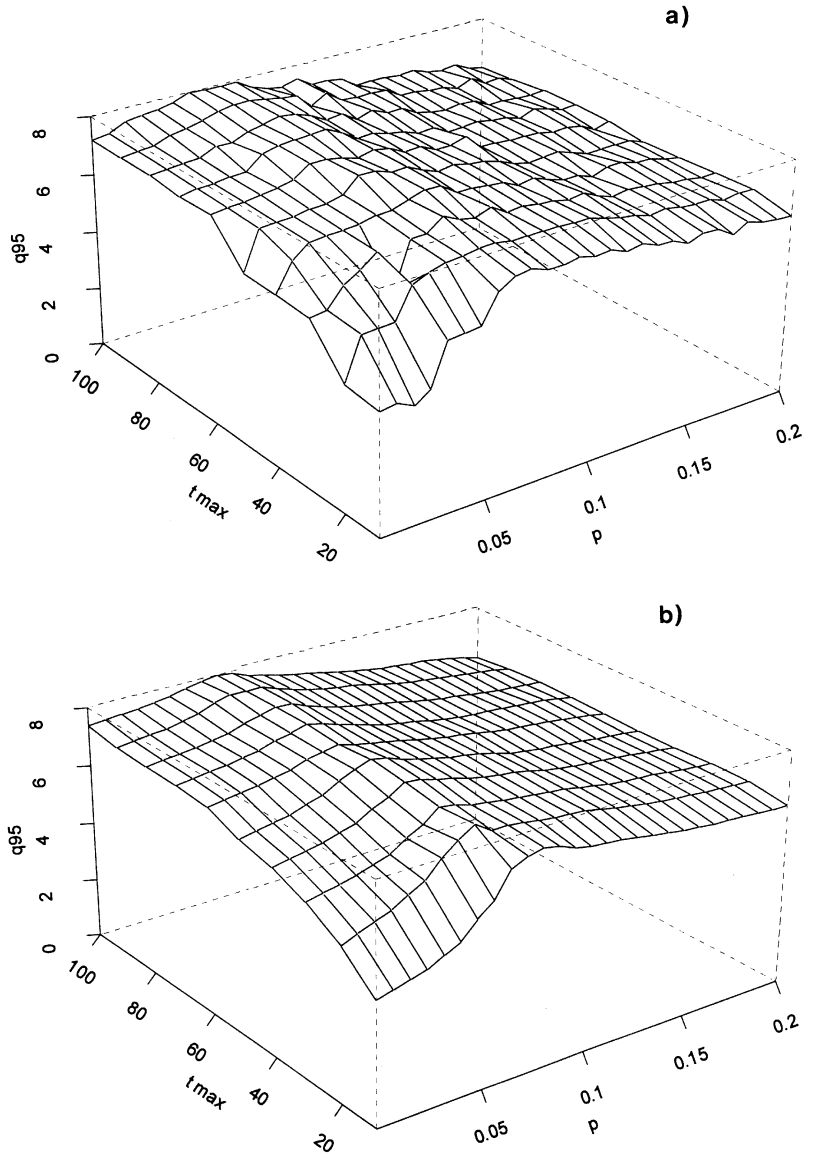

Fig. 1. Estimated quantiles at level 0.95 in the distribution of $-2\ln\Lambda_{\max}$ under H_0 ($p=p_1=p_2$) as a function of the common hazard rate and $t_{\max}$ ($n=100; r=10\,000$). (a) Simulated data (b) smoothed surface by using locally weighted regression.

We define the power of the sequential test at time $t, \pi(t)$, as the probability of finding a significant difference, if it exists, by time t. In other words, $\pi(t)=\Pr(D\leqslant t)$ is the probability of stopping by t. Then, the power at time t depends on the expectation of the maximum value of the likelihood ratio statistic, $-2\ln\Lambda_{\max}$, which is associated with the discrete time interval $[1,\ldots,t]$. The variation of the conditional expectation

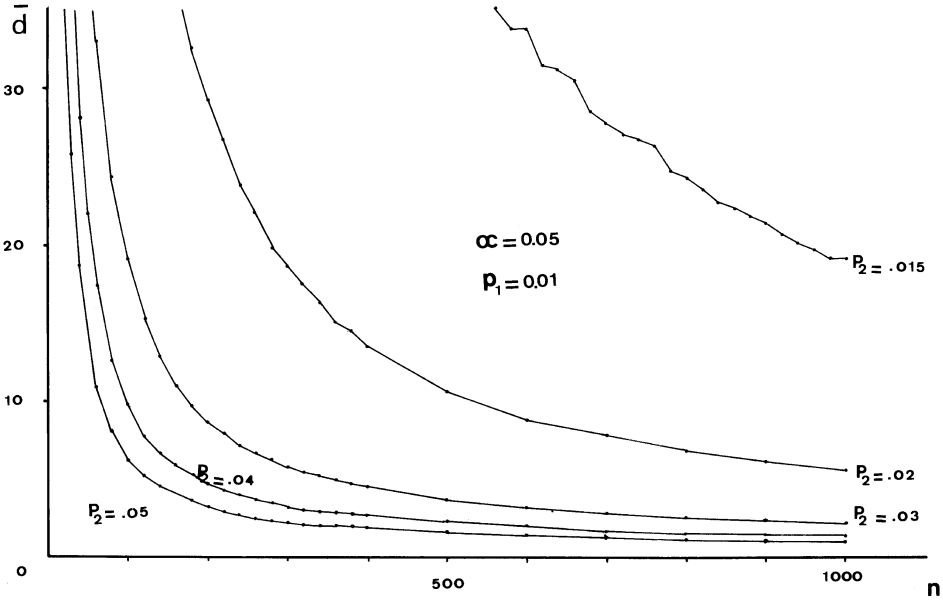

Fig. 2. Expectation of D (conditional on $D \leqslant 200$) as a function of the sample size n ($r = 5000$).

of $-2 \ln \Lambda_{\max}$, as a function of the sample size n and the hazard ratio p_2/p_1, can be obtained by Monte Carlo simulation, conditional on the event (time $\leqslant t$). A graphical illustration of power analysis is given in Fig. 3. From this figure, it appears that the conditional expectation of $-2 \ln \Lambda_{\max}$, and thus indirectly the power of the test, may be considered an increasing linear function of n. Moreover, the slope of the straight line increases as the hazard ratio departs further from unity. Notwithstanding the increasing relation between the cumulative power and time, it is also clear from Fig. 3 that power increases as p_2/p_1, n and α increase. This remark suggests the use of power analysis, namely, power curves, for determining the number of individuals needed for carrying out a survival comparative test. The number of individuals to be involved in the study increases as the hazard ratio becomes closer to one. In particular, the straight lines in Fig. 3 give some quantitative grounds to the fact that a sufficient number of individuals is required in order to avoid the possibility of ending on an inconclusive experiment if there is a difference in the survival rates indeed. This is also the reason why a practical time limit ($t_{\max}$) has to be stated before starting a sequential test.

Regarding practical purposes, it is possible to give more flexibility to the DTBLRS approach. For instance, one may adjust the interval of time between two consecutive tests in the sequential procedure so as to render the different hazard rates concordant with the values which are used in this paper. Moreover, the discrete time expression for the hazard while examining the different study subjects individually allows one to carry out the sequential procedure in situations of any type of single censored data

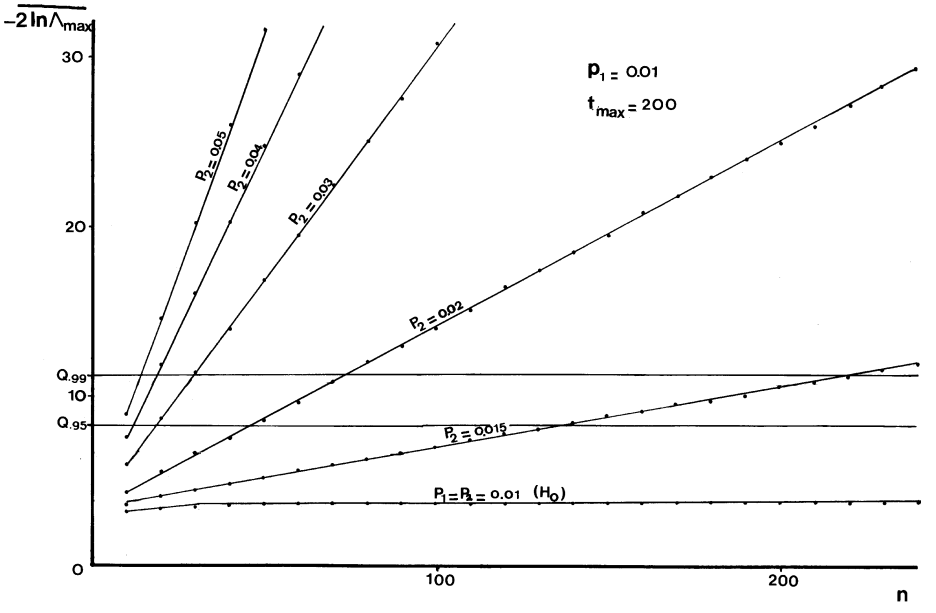

Fig. 3. Estimated expectation of the maximum value of the likelihood ratio statistic (conditional on $D \leqslant t_{\max}$) as a function of the sample size n ($r = 5000$).

involving staggered entries; for example, with a fresh block of individuals entering the study from one monitoring point to the next. This is a consequence of the monotonicity of the expectation of $-2\ln\Lambda$ which, under H_1, is an increasing function of both time and the number of individuals (see Fig. 3). Clearly, the expectation of the likelihood ratio statistic, in case the subjects arrive in sequence, is in the preferred direction of slightly underestimating the value of this expectation corresponding to a simultaneous entrance of all the subjects from the beginning of the study. Furthermore, some of the previous results allow the sequential procedure to be applicable to the analysis of survival data with time-dependent hazards which may incorporate a variety of shapes. For instance, the two-sample comparison test may be extended to a discrete time proportional hazards model, that is, assuming a constant hazard ratio as the underlying alternative hypothesis. Indeed, the surface in Fig. 1 shows that type-I error can still be maintained at any given level. To this end, one needs to consider the highest value of the critical threshold corresponding to the range of values of the hazard rate, which is expected to be encountered on the whole period of the study.

4. Examples

The first example is concerned with studying the effect of smoking on fecundability (Weinberg and Gladen, 1986). The duration variable is the time taken by couples

Table 1
Data on the number of menstrual cycles to pregnancy (adapted from Weinberg and Gladen, 1986)

Time interval	S or NS	Number of menstrual cycles												
		1	2	3	4	5	6	7	8	9	10	11	12	>12
1	S	5	1	2				1						1
	NS	21	5	7	2	3					1	1		1
2	S	1		2	1		1		1					
	NS	26	17	7	3		3		1	1		1	2	
3	S	3	3	1	1	1								2
	NS	16	12	4	1		1	2	2	1			2	3
4	S	2	2	2			2			1			1	
	NS	19	11	10	5	5	3	1	1	1		1	1	1
5	S	2		2			1	1				1		1
	NS	28	5	5	8	1	2	1		1				1
6	S	3	1	2		2	1		3					
	NS	16	14	2	4	2			1					
7	S	2	2	1				1						
	NS	20	16	2	2	3	2	1						
8	S	5	2	3	1		1	1	1					2
	NS	14	5	6	2	1	6	2	2		1	1		3
9	S	2	3		1						1		1	
	NS	24	6	4	5		4		1			1		1
10	S	4	2	2			3						1	1
	NS	14	16	8	6	3	1		1	1	1	1	1	2

S: Smokers.
NS: Non-smokers.

who were attempting to conceive, until pregnancy results. To this end, the number of menstrual cycles to pregnancy was recorded for each of the 586 couples involved in the study. Couples are designated as "smokers" if the female partner smoked. The data set used to illustrate the DTBLRS method is presented in Table 1. Thus, the entire duration of the study was divided into ten intervals (months) and each observation (couple) of the original data was allocated at random to one of these intervals of time. The data in Table 1 were subjected to a prospective sequential analysis. As a further justification for this type of study, one can find its interest within the framework of a trial aiming at studying, for instance, the effect of a new treatment against sterility. The outcome of the testing procedure at each step of the DTBLRS method is given in Table 2. The total number of events and the successive maximum-likelihood estimates of the probability of conception at each cycle are also presented in Table 2. These estimates were calculated (i) within the two groups of subjects (H_1), or (ii) by assuming there is no difference between the smokers and non-smokers (H_0), with respect to fecundability. The P-values displayed in the last column of Table 2 required previous Monte Carlo simulations ($r = 5000$) to determine the empirical distribution of $-2 \ln \Lambda_{\max}$ ($t_{\max} = 10$) under H_0 (i.e. $p_1 = p_2 = 0.30$). It becomes clear from the results in Table 2 that fecundability is affected by smoking. Thus, the depreciative effect of

Table 2
Sequential analysis of the effect of smoking on the number of menstrual cycles to pregnancy

Time interval	S or NS	Total number of events	$\hat{p}$ (under H_1)	Common $\hat{p}$ (under H_0)	$-2\ln \Lambda^t$
	S	5	0.5000		
1				0.5098	0.0048
	NS	21	0.5122		$P<0.8546$
	S	7	0.3333		
2				0.4126	0.6509
	NS	52	0.4262		$P<0.7122$
	S	12	0.2927		
3				0.4047	2.6245
	NS	92	0.4259		$P<0.2794$
	S	19	0.2879		
4				0.3823	3.0873
	NS	132	0.4012		$P<0.2131$
	S	25	0.2717		
5				0.3773	5.4507
	NS	181	0.3987		$P<0.0473$
	S	31	0.2500		
6				0.3521	7.2090
	NS	213	0.3743		$P<0.0146$
	S	39	0.2500		
7				0.3501	8.7784
	NS	260	0.3725		$P<0.0051$
	S	48	0.2449		
8				0.3474	11.7587
	NS	306	0.3718		$P<0.0007$
	S	56	0.2258		
9				0.3378	18.4780
	NS	348	0.3671		$P<0.0001$
	S	69	0.2363		
10				0.3295	15.2038
	NS	385	0.3545		$P<0.0001$

S: Smokers.
NS: Non-smokers.

smoking on fecundability could be stated from the seventh interval of time onwards ($P<0.01$).

The relative efficiency of the DTBLRS method over the conventional statistical practices, which are still currently used for the analysis of clinical trials dealing with survival data, is illustrated by means of a second example referring to a trial on superficial bladder cancer (Freedman, 1982). In this example, it was assumed that with the current method of treatment of superficial bladder cancer (i.e. resection of tumor at cystoscopy) the recurrence-free rate was 50% at 2 yr. The problem is to design a clinical trial which is appropriate to show an increase in the previous rate to at least 70% using intravesical chemotherapy immediately after surgery by the time of cystoscopy.

If we assume that there is an exponential distribution of survival times, then the hazard rates corresponding to the previous situations are: $p_2 = 0.02847$ and $p_1 = 0.01475$ per month, respectively. Rescaling the time from months to 21-day periods results in 35 time-periods (over 24 months). This is done for convenience to obtain values of the hazards rates in agreement with those used in the present work. Hence, the values of the hazard rates become $p_2 = 0.01934$ (say 0.02) and $p_1 = 0.01$ if the time unit is changed to 20.57 days. The characteristics of the procedure suggested in this paper are compared with (a) the approach based on the exponential model, assuming an approximate normal distribution of the ln maximum-likelihood estimate of the hazard rates (Schoenfeld and Richter, 1982; Bernstein and Lagakos, 1978), and (b) the non-parametric method using the logrank test (Freedman, 1982) under the assumption that the hazard function is expressed as the well-known semi-parametric proportional hazards model (Cox, 1972). To facilitate comparison between the different methods, all the patients are assumed to have been entered in the study at the same time, i.e. $t = 0$. Statistical analysis of the results by following methods (a) and (b) is assumed to occur after a follow-up time which is fixed at $t_{\max} = 24$ months.

The benefit of the sequential approach over the other two methods is shown in Table 3. It can be assessed by comparing the power and/or the total expected number of events needed before a decision is planned to be made (methods (a) and (b)) or can be made (DTBLRS method). However, when doing such a comparison between the different methods one should be aware that the proportional hazards model is used in the approach developed by Freedman (1982) whereas the simple but also more restrictive exponential model is used in the other two methods. The mean number of events, which is mentioned in Table 3, was calculated by the end of each sequential procedure (conditional on $t \leqslant t_{\max}$) from $r = 5000$ simulated experiments.

The results presented in Table 3 show that the DTBLRS procedure allows a decision to be made at a time, and subsequently a number of events, which are both considerably smaller than those by the other two approaches. These two methods, in turn, are comparable in terms of efficiency. Nevertheless, it is interesting to note that the power characterizing the sequential process at the time limit of the comparative study (i.e. $t_{\max} = 24$) is less than the corresponding power as calculated for the other two methods. The difference is more marked when the power is low. This may be considered as the counterpart to the possibility of stopping the sequential procedure at any step before the time limit of the trial is reached while maintaining type-I error under a prespecified level during the whole study. Seen another way, it must be emphasized that there might be a marked difference in the hazard rates one is not necessarily aware of at the beginning of the experiment. In this case, the possibility for early termination of the trial constitutes the chief advantage of the sequential procedure over the fixed-time test.

It should be pointed out that if there are only two groups to be compared, a one-tailed test might be more appropriate. Other simulation methods similar to those which are used in this paper could be used to obtain a one-sided sequential test. These tests would be more powerful than the DTBLRS method to handle the examples presented in this work.

Table 3
Efficiency of the sequential procedure compared with two other commonly used approaches in planning a comparative test

Number of individuals	Method	Hazard rates corresponding to a 21 days – period of time: $p_1 = 0.01$ and $p_2 = 0.02$			
		$\alpha = 0.05$		$\alpha = 0.01$	
	Normal	$\pi = 0.323$	$m = 20$	$\pi = 0.142$	$m = 20$
	Non-parametric	$\pi = 0.309$	$m = 20$	$\pi = 0.125$	$m = 20$
$n = 50$					
		$\pi(6) = 0.040$		$\pi(6) = 0.006$	
	Sequential	$\pi(12) = 0.097$	$\bar{m} = 12.62$	$\pi(12) = 0.032$	$\bar{m} = 14.02$
		$\pi(24) = 0.189$		$\pi(24) = 0.079$	
	Normal	$\pi = 0.565$	$m = 40$	$\pi = 0.325$	$m = 40$
	Non-parametric	$\pi = 0.563$	$m = 40$	$\pi = 0.310$	$m = 40$
$n = 100$					
		$\pi(6) = 0.097$		$\pi(6) = 0.030$	
	Sequential	$\pi(12) = 0.204$	$\bar{m} = 23.44$	$\pi(12) = 0.083$	$\bar{m} = 26.37$
		$\pi(24) = 0.393$		$\pi(24) = 0.202$	
	Normal	$\pi = 0.852$	$m = 80$	$\pi = 0.665$	$m = 80$
	Non-parametric	$\pi = 0.861$	$m = 80$	$\pi = 0.667$	$m = 80$
$n = 200$					
		$\pi(6) = 0.202$		$\pi(6) = 0.084$	
	Sequential	$\pi(12) = 0.426$	$\bar{m} = 42.61$	$\pi(12) = 0.232$	$\bar{m} = 50.65$
		$\pi(24) = 0.713$		$\pi(24) = 0.507$	
	Normal	$\pi = 0.989$	$m = 160$	$\pi = 0.953$	$m = 160$
	Non-parametric	$\pi = 0.992$	$m = 160$	$\pi = 0.959$	$m = 160$
$n = 400$					
		$\pi(6) = 0.427$		$\pi(6) = 0.239$	
	Sequential	$\pi(12) = 0.763$	$\bar{m} = 65.42$	$\pi(12) = 0.574$	$\bar{m} = 82.70$
		$\pi(24) = 0.967$		$\pi(24) = 0.898$	

π: Power of the test at termination of the trial (i.e. $t_{\max} = 24$ months).
m: Expected number of events needed to be observed.
$\pi(t)$: Power of the sequential procedure at t months.
$\bar{m}$: Mean total number of events observed by the end of the test ($t \leqslant 24$) obtained by simulation ($r = 5000$).

5. Concluding remarks

The sequential procedure presented in this paper provides a particularly convenient and useful way for designing appropriate comparative tests. Namely, the approach can be used for comparing survival distributions in a wide class of biomedical (e.g. clinical trials) or environmental health (e.g. risk analysis) investigations. The critical threshold which is used when stating the decision rule in the DTBLRS approach can be determined empirically by simulation. The criterion for early stopping of the process is obtained by computing the exact distribution of the likelihood ratio statistic under the null hypothesis. The value of the threshold is thus determined on the basis of satisfying statistical grounds so as to control type-I error at a specific level during the

entire sequential testing procedure. In a comparative study between two populations the gain of efficiency resulting from the sequential approach may be particularly marked, especially if the actual hazard ratio between the groups departs from unity more than expected when planning the study. In such a situation, both the duration and the number of events are substantially smaller than one could expect had one used a current method based on the analysis of the data observed at a predetermined value of the time. Power analysis shows that planning the size and duration of a comparative study can be achieved on the basis of: (i) the combination of power and level of significance to be attained, and (ii) the prior knowledge of the hazard rates in each group or, similarly, the smallest difference in the relative hazard rates one wishes the trial to be able to detect reliably. The simulation procedure used while performing the DTBLRS test may be considered a reference method for computing the critical threshold which can be applicable to a large class of survival distributions. Finally, the great generality of the statistical model considered accommodates the possibility of handling more than two treatment groups with time-dependent hazard rates and staggered entries comprising high rates of single censored values and/or tied failure times.

Acknowledgements

The author would like to thank Dr. P. Wild for many useful comments on a previous draft of the manuscript.

References

Bernstein, D., Lagakos, S., 1978. Sample size and power determination for stratified clinical trials. J. Stat. Comput. and Simul. 8, 65–73.

Casagrande, J.T., Pike, M.C., Smith, P.G., 1978. The power function of the "exact" test for comparing two binomial distributions. Appl. Statist. 27, 176–180.

Cochran, W.G., Cox, G.M., 1957. Experimental Design, 2nd ed. Wiley, New York, pp. 24–25.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Statist. Assoc. 83, 596–610.

Cox, D.R., 1972. Regression models and life-tables (with discussion). J.R. Statist. Soc. Series B 34, 187–220.

Fleming, T.R., DeMets, D.L., 1993. Monitoring of clinical trials: issues and recommendations. Control. Clin. Trials 14, 183–197.

Freedman, L.S., 1982. Tables of the number of patients required in clinical trials using the logrank test. Statist. Med. 1, 121–129.

Gehan, E.A., 1961. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. J. Chron. Dis. 13, 346–353.

George, S.L., Desu, M.M., 1974. Planning the size and duration of a clinical trial studying the time to some critical event. J. Chron. Dis. 27, 15–24.

Lan, K.K., Zucker, D.M., 1993. Sequential monitoring of clinical trials: the role of information and Brownian motion. Statist. Med. 12, 753–765.

Lawless, J.F., 1982. Statistical Models and Methods for Lifetime Data. Wiley, New York.

Lee, J.W., 1994. Group sequential testing in clinical trials with multivariate observations: a review. Statist. Med. 13, 101–111.

Lin, D.Y., Wei, L.J., DeMets, D.L., 1991. Exact statistical inference for group sequential trials. Biometrics 47, 1399–1408.

Machin, D., Campbell, M.J., 1987. Statistical Tables for the Design of Clinical Trials. Blackwell, Oxford.
Maul, A., 1994. A discrete time logistic regression model for analyzing censored survival data. Environmetrics 5, 145–157.
Mehta, C.R., Patel, N., Senchaudhuri, P., Tsiatis, A., 1994. Exact permutational tests for group sequential clinical trials. Biometrics 50, 1042–1053.
Pawitan, Y., Hallstorm, A., 1990. Statistical interim monitoring of the cardiac arrhythmia suppression trial. Statist. Med. 9, 1081–1090.
Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1977. Design and analysis of clinical trials requiring prolonged observation of each patient, I. Introduction and design. Br. J. Cancer 34, 585–612. II. Analysis. Br. J. Cancer 35, 1–39.
Prentice, R.L., 1973. Exponential survivals with censoring and explanatory variables. Biometrika 60, 279–288.
Schoenfeld, D., Richter, J., 1982. Monograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. Biometrics 68, 163–170.
Shuster, J.J., 1993. Fixing the number of events in large comparative trials with low event rates: a binomial approach. Control. Clin. Trials 14, 198–208.
Weinberg, C.R., Gladen, B.C., 1986. The beta-geometric distribution applied to comparative fecundability studies. Biometrics 42, 547–560.
Whitehead, J., 1992. The Design and Analysis of Sequential Clinical Trials, 2nd ed. Ellis Horward, New York.