In Computing Science and Statistics (1992)
Proceedings of the 22nd Interface Conference, 1990
(eds C. Page & R. LePage) Springer-Verlag, 89-97.

Bootstrap tests and confidence intervals for a hazard ratio when the number of observed failures is small, with applications to group sequential survival studies

BY CHRISTOPHER JENNISON

School of Mathematical Sciences,
University of Bath,
Bath, BA2 7AY, U.K.

ABSTRACT

We present a small sample approximation to the distribution of the efficient score statistic for testing the null hypothesis $\lambda=\lambda_0$, where λ is the hazard ratio in a proportional hazards model. Here, "small sample" refers to a small number of failures in the observed data. Our basic approximation is to the conditional distribution of observations' group memberships, given the observed number of failures and number of censored observations between successive failures; the implied distribution of the score statistic is found by simulation. Our method can be incorporated into sequential procedures for monitoring survival data and it successfully overcomes inaccuracies in the usual normal approximations that arise when only a few subjects have failed at early analyses.

A simulation study to assess the accuracy of our approximation required bootstrap simulation nested within experimental replications. Here, the computational demands were alleviated substantially by conducting the bootstrap tests themselves sequentially; using an innovative form of stochastic curtailment reduced the required computation by a factor of up to 25, 50 or even more.

1. INTRODUCTION

Consider a sequential clinical trial in which subjects arrive over a period of time, each patient is randomly allocated to one of two treatments and his or her subsequent progress is followed until death. Suppose that survival time is the major endpoint of interest and a proportional hazards model with hazard ratio λ between the two treatments is assumed. The logrank statistic is commonly used in such situations. Asymptotic theory (Tsiatis 1981, 1982, Harrington, Fleming & Green 1982) shows that the joint distribution of the sequence of logrank statistics observed at successive interim analyses is approximately multivariate normal. In the asymptotic setting, results are obtained in the limit as the rate of accrual of subjects to the study increases, with calendar times of analyses held fixed. The accuracy of the normal approximation for a single analysis depends primarily on the total number of observed failures; at early analyses, when only a small number of patients have been accrued and very few have failed, it may be quite poor. Nevertheless, in extreme cases, for example, if 11 out of 12 failures have occurred on the same treatment arm, one would still like to have the opportunity of reaching an early decision.

Inaccuracies in the normal approximation for the logrank statistic and more general score statistics are greater when $\lambda\neq1$. In order to implement sequential procedures based on repeated confidence intervals (Jennison & Turnbull, 1984, 1989) it is important to have reliable group sequential tests of null hypotheses $H_0: \lambda=\lambda_0$ for all λ_0. A good approximation to the joint distribution of the sequence of logrank statistics under values of $\lambda\neq1$ is also needed to construct confidence intervals for the hazard ratio following a conventional group sequential test. Our objective in this paper is to develop tests of $H_0: \lambda=\lambda_0$ for general λ_0, which achieve a specified type I error divided equally between the two sides of the alternative hypothesis $\lambda\neq\lambda_0$.

In § 2 we describe the sequential survival problem and in § 3 we introduce our proposed small sample

approximation. Implementation of this approximation for significance tests and confidence intervals is discussed in § 4. In § 5 we report on simulation studies carried out to assess the accuracy of our approximation; we also describe a form of sequential curtailment of the bootstrap test which led to large reductions in the computation time of these simulations. In § 6 we make some concluding remarks and indicate a topic for future research.

2. THE SURVIVAL PROBLEM

We consider the problem of comparing two groups of patients in a clinical trial. We assume that survival times for the two groups follow a proportional hazards model with hazard rates $h(t)$ and $\lambda h(t)$ in groups 1 and 2 respectively. Suppose subjects enter the study at staggered intervals and they are also subject to competing risk censoring which is assumed to be independent of survival time and of treatment group, for example, loss to follow up due to a subject's moving to a different part of the country. Suppose also that the survival study is monitored sequentially, interim analyses being conducted at predetermined calendar times. Our objective is to derive sequential tests of the null hypothesis $H_0: \lambda=\lambda_0$ against a two sided alternative, achieving a specified type I error, α.

Let d_k denote the number of exact failures observed at the time of the kth analysis and $t_1 < t_2 < \ldots < t_{d_k}$ the ordered values of these failure times; the failure time for each subject is measured from his or her time of entry to the study. Note that some subjects may not yet have entered the study and others will be subject to end-of-study censoring in addition to competing risk censoring. The efficient score statistic for testing $H_0 : \lambda = \lambda_0$, based on the partial likelihood (Cox, 1972) of the data available at the time of the kth analysis, is

$$L_k(\lambda_0) = \sum_{i=1}^{d_k} \delta_i - \frac{\lambda_0 r_{i2}}{r_{i1}+\lambda_0 r_{i2}} \quad , \qquad (2.1)$$

where r_{i1} and r_{i2} are the numbers at risk in groups 1 and 2 respectively just before time t_i and δ_i is an indicator variable taking the value 1 if the failure at time t_i is in group 2; the variables δ_i, r_{i1} and r_{i2} all depend on k but this dependence is suppressed to simplify the notation. If $\lambda_0=1$, this statistic reduces to the well-known logrank statistic.

Harrington *et al.* (1982) prove that the sequence of statistics $\{L_k(\lambda_0)\}$, $k\geq 1$, has, asymptotically, a multivariate normal joint distribution with independent increments. A consistent estimate of the asymptotic variance is available and the correspondingly standardised statistic,

$$\frac{L_k(\lambda_0)}{\{\sum_i \lambda_0 r_{i1} r_{i2}/(r_{i1}+\lambda_0 r_{i2})^2\}^{1/2}} \quad , \qquad (2.2)$$

is approximately $N(0,1)$. Equivalently, one can treat the joint distribution of the sequence of unstandardised statistics $\{L_k(\lambda_0)\}$ as that of a sequence of zero-mean normal variables with the variance of $L_k(\lambda_0)$ equal to

$$I_k = \sum_i^{d_k} \frac{\lambda_0 r_{i1} r_{i2}}{(r_{i1}+\lambda_0 r_{i2})^2}$$

and independent increments, i.e., the same joint distribution as a standard Brownian motion observed at times I_k in the Brownian motion timescale. This representation is particularly convenient for the construction of group sequential tests of $H_0: \lambda=\lambda_0$. We shall follow the approach of Slud & Wei (1982) and specify a group sequential test with K analyses by a sequence of probabilities $\pi_1, \ldots, \pi_K$, summing to α, where π_k denotes the amount of type I error to be "spent" at the kth analysis. Thus, $H_0: \lambda=\lambda_0$ is to be rejected at the kth analysis if $|L_k(\lambda_0)| \geq c_k$, where the critical values $c_1, \ldots, c_K$ are chosen to satisfy

$$P(|L_1(\lambda_0)| < c_1, \ldots, |L_{k-1}(\lambda_0)| < c_{k-1},$$

$$|L_k(\lambda_0)| \geq c_k) = \pi_k \qquad k=1, \ldots, K. \qquad (2.3)$$

Under the above normal approximation to the joint distribution of the sequence $L_1(\lambda_0), L_2(\lambda_0), \ldots$, values $c_1, c_2, \ldots$ can be calculated successively, using numerical integration to evaluate the left hand side of (2.3). Note that only the values $I_1, \ldots, I_k$, which are known at the time of the kth analysis, are needed to find c_k.

Gail, DeMets & Slud (1982) and DeMets & Gail (1985) have studied the normal approximation to the joint distribution of sequentially calculated logrank statistics. They find this approximation to work well for $\lambda=1$ but note that differences from the anticipated power of sequential tests occur at $\lambda=2$. Jennison & Turnbull (1989) report on a simulation study investigating the adequacy in sequential tests of the normal approximation for general score statistics. They note that it works well in many cases but that problems do arise when only a small number of failures, e.g., 20 or 30, have occurred at the time of one or more early analyses. Inaccuracies are greatest when the values of π_k for small k are reasonably large, for example, if a repeated significance test at constant nominal level (Pocock, 1977), adapted to unequal increments in I_k, is used. In typical examples, achieved

one-sided error rates range from 0.035 to 0.06 compared with their intended value of $\alpha/2 = 0.05$. Closer inspection of these examples reveals that the empirical probability of first rejecting H_0 at analysis k differs substantially from the nominal value, π_k, when the average number of failures occurring by analysis k is small.

Our proposal for improving agreement with nominal error rates is to retain the general approach outlined above, calculating critical values c_k by solving (2.3) under the normal approximation. However, if no more than 30 failures have been observed at analysis k the criterion for rejecting H_0 is replaced by a two-sided significance test of size $2\{1-\Phi(c_k/\sqrt{I_k})\}$ and this test is conducted using a good small sample approximation to the distribution of $L_k(\lambda_0)$. This method of calculating a sequence of nominal significance levels appropriate to normal data but implementing individual tests using exact or nearly exact distributions of non-normal statistics appears to work well quite generally; Jennison & Turnbull (1991) note that it gives good results for sequential t- and χ^2-tests.

3. A SMALL SAMPLE APPROXIMATION TO THE DISTRIBUTION OF $L_k(\lambda_0)$

We introduce the small sample approximation in the non-sequential case. Notation is as in § 2, but subscripts k denoting the number of the interim analysis are omitted. Let n denote the total number of observations available. Arrange the set of n survival times, censored or exact, in ascending order and let $J(i)$, $i=1,\ldots,n$, take the value 0 if the ith time is censored and 1 if it is exact. We shall approximate the conditional distribution of $L(\lambda_0)$ under $\lambda=\lambda_0$ given $\{J(i); i=1,\ldots,n\}$ or, equivalently, given the number of exact failures d and the numbers of censored observations between each pair of successive failure times. We first approximate the conditional joint distribution of observations' group memberships, 1 or 2, given $\{J(i); i=1,\ldots,n\}$. This approximating distribution is most easily described by means of an algorithm and this also provides a straightforward method for generating realisations from the distribution. Start at time $t=0$, with numbers n_1 and n_2 at risk in groups 1 and 2 respectively where n_1 and n_2 are the actual group sizes; if the next observation is censored allocate it to group 1 with probability $n_1/(n_1+n_2)$ and to group 2 otherwise, if the next observation is exact allocate it to group 1 with probability $n_1/(n_1+\lambda_0 n_2)$ and to group 2 otherwise, decrease either n_1 or n_2 by 1 as appropriate and proceed to the next observation; continue this procedure until all observations have been allocated a group label. The value of $L(\lambda_0)$ is calculated by applying the formula (2.1) to the new data set; the distribution of such values forms our approximation to the conditional distribution of $L(\lambda_0)$ given $\{J(i); i=1,\ldots,n\}$ under $\lambda=\lambda_0$.

To see that this method is only approximate, consider the labelling of the first censored observation: if the number of subsequent exact failures is relatively high and $\lambda_0>1$, there is a greater conditional probability that the censored observation is from group 1, since this would leave more members of group 2 who are more likely to fail before being censored. Here "relatively" high depends on the baseline hazard rate $h_0(t)$ and the censoring distribution and, since these are unknown, it follows that the conditional distribution of $L(\lambda_0)$ cannot be found exactly. However, our approximate conditional distribution is exactly correct either if $\lambda_0=1$ or in the absence of censoring. It can also be shown, by application of the Martingale convergence theorem (Brown, 1971, Theorem 2), that the implied distribution for the standardised statistic (2.2) converges to the standard normal distribution. Thus, our approximation is guaranteed to be reliable for large samples; in addition, the fact that it is supported on the same set of values as $L(\lambda_0)$ and its exactness in the above special cases promises a better performance in small samples than the usual normal approximation.

We shall refer to tests of $H_0: \lambda=\lambda_0$ in which the reference distribution for $L(\lambda_0)$ is obtained by simulating from the distribution described above as "bootstrap" tests. Usage of the term "bootstrap" now extends to tests which would previously have been described simply as Monte Carlo tests but our tests do possess other features more closely related to the ideas introduced by Efron (1979). Specifically, the reference conditional distribution used in simulations is only an approximation to the true distribution, but it is asymptotically equivalent. However, we have deliberately chosen not to generate "bootstrap" samples by resampling from the original survival times (Efron, 1981) or by sampling from Kaplan-Meier survival estimates (Reid, 1981); we believe the dependence on λ_0 of the bootstrap distribution for testing $H_0: \lambda=\lambda_0$ to be a key factor in the accuracy of our method, in particular, it is essential in order for the approximation to be exact in the two special cases mentioned above.

4. IMPLEMENTATION

The basic method for Monte Carlo testing was suggested by Barnard (1963). To perform a size α, equal tailed test of the null hypothesis $H_0: \lambda=\lambda_0$, generate $N-1$ values from the approximating conditional distribution for $L(\lambda_0)$ as described in § 3. Combining these with the observed

value gives a sample of N observations. If the approximating distribution were exactly correct then, under H_0, the observed value would be equally likely to be any one of the N values in this sample, so a satisfactory test is to reject H_0 if the observed value is one of the $N\alpha/2$ smallest or $N\alpha/2$ largest values.

Marriott (1979) discusses how large N should be in a test of this form. Our recommendation is to take N extremely large, e.g., 10000, 100000 or a million. Although Barnard's test provides a clever way of incorporating the randomness of the simulations into the type I error statement, small values of N do sacrifice power; since survival data are usually expensive to collect but simulations are easy and fast, it is only right to make as much use as possible of the available data. It is also advisable, for obvious reasons, to minimise the chance that two statisticians conducting the "same" bootstrap test reach opposite conclusions. Our philosophy here is really to use simulation to perform Monte Carlo integration with N chosen to ensure that the numerical error is negligible.

Computation of a confidence interval for λ is of interest in its own right; it is also a necessary step in constructing a sequence of repeated confidence intervals for λ. Computation of a $1-\alpha$ level confidence interval for λ requires the inversion of a family of hypothesis tests. Equivalently, we seek $\underline{\lambda}$ and $\bar{\lambda}$ such that $L(\underline{\lambda})$ is at the $1-\alpha/2$ quantile of the bootstrap distribution for $\lambda=\underline{\lambda}$ and $L(\bar{\lambda})$ is at the $\alpha/2$ quantile of the bootstrap distribution under $\lambda=\bar{\lambda}$. Substantial savings in computation can be made by modelling

$$p(\lambda) = P\{\text{Bootstrap } L \text{ under } \lambda \geq \text{observed } L(\lambda)\}$$

as a function of λ in the vicinity of the endpoints of the confidence interval, initial estimates of which can be obtained using the normal approximation. We illustrate this approach with an example.

The following display represents the rank data for a sample of 40 survival times, twenty in group 1 and twenty in group 2. The display shows the group memberships of the forty observations arranged in increasing order with asterisks denoting censored observations.

$$1^*, 2, 2, 1^*, 2^*, 2, 1^*, 1, 2, 2^*, 2, 1, 2,$$

$$1^*, 2, 1^*, 2, 1, 2^*, 2, 1, 2, 1^*, 2, 1, 1^*, 1,$$

$$2^*, 1^*, 1, 2^*, 1^*, 1, 2^*, 1^*, 1^*, 2, 2^*, 1^*, 2^*.$$

The normal approximation yields (0.714, 4.112) as a 95% confidence interval for λ. Performing 10000 bootstrap simulations at each of a set of λ values near 0.714 gave the following results:

$\log \lambda$	*Number out of* 10000 *bootstrap values* > *observed* $L(\lambda)$
-0.40	190
-0.39	209
-0.38	238
-0.37	244
-0.36	241
-0.35	267
-0.34	258
-0.33	272
-0.32	284
-0.31	303
-0.30	332

Fitting a logistic regression model, we obtain

$$\log\{p/(1-p)\} = -1.993 + 4.712 \log\lambda$$

and solve to find $\underline{\lambda}=0.702$ for $p=0.025$. Repeating the same exercise at values of λ near 4.112 gave a fitted model

$$\log\{p/(1-p)\} = -2.211 + 3.936 \log\lambda$$

and, hence, $\bar{\lambda}=4.450$ for $p=0.975$. Thus our 95% confidence interval is (0.702, 4.450). Note that we fit two separate models and use each one only in the vicinity of those λ values used to fit it. Examination of the deviance of each fitted model provides a check on its chosen form. Also, variances of the roots, $\underline{\lambda}$ and $\bar{\lambda}$, can be obtained from variances and covariances of the parameters of the logistic regression model. In this example standard errors for $\underline{\lambda}$ and $\bar{\lambda}$ are 0.003 and 0.036 respectively; if these are deemed too large, further simulations should be performed.

Many authors use pivotal or nearly pivotal quantities to construct bootstrap confidence intervals. It is unlikely that a really good pivot exists in this problem since the distribution of $L(\lambda_0)$ is discrete and often "clumpy" — values depend primarily on $\Sigma\delta_i$, the number of failures in group 2, with small perturbations arising from the second term $\Sigma\lambda_0 r_{i2}/(r_{i1}+\lambda_0 r_{i2})$. In any case, we would argue that to investigate $p(\lambda_0)$ by simulating directly under $\lambda=\lambda_0$ removes one level of approximation and a possible source of error.

5. ASSESSMENT

5.1 *Design of simulation studies*

We have examined the accuracy of our proposed method in producing equal tailed size α tests (or, equivalently, $1-\alpha$ confidence intervals). In a simulation study, we applied our method to M data sets simulated under the null hypothesis and observed the empirical error rate. To estimate a true error rate of around 0.05 with a standard error of 0.0015, say, requires $M=20000$ replicates; since $N-1$ bootstrap simulations are nested within each replication, the computational task is substantial.

If our approximation to the conditional distribution of $L(\lambda_0)$ is good, the standard argument for Barnard's (1963) Monte Carlo test, as described in § 4, implies that the error probabilities of this test should be close to their nominal values for all N, as long as $N/2$ is an integer. However, if the approximation is poor, the achieved error rates for small N are complex functions of the true and approximating distributions and they might still turn out to be close to their nominal values. Thus, studies with large N, of the order of the value to be used in practice, are necessary to confirm fully the accuracy of the approximation. In our simulations, we started with $N=20$ and 100 and, since these gave promising results, moved on to $N=1000$. In sequential tests, the bootstrap test is part of a larger procedure and Barnard's argument for small N does not apply. Thus, a large value of N must be used from the outset. We have used $N=1000$, which is considerably less than the value we recommend for analysing real data. However, a heuristic argument in which the randomness of the bootstrap test is equated with additional variance in $L(\lambda_0)$ suggests that this should affect the error rates only slightly.

5.2 *Sequential curtailment of bootstrap tests*

In our bootstrap test, the value $L(\lambda_0)$ is calculated from the data, $N-1$ "bootstrap" values of L are simulated and $H_0: \lambda=\lambda_0$ is rejected if the number of bootstrap values greater than $L(\lambda_0)$ is $\leq C$ or $\geq N-C-1$ for some integer C. (For the present discussion we shall ignore the possibility that the bootstrap value is exactly equal to $L(\lambda_0)$.) Let G_n denote the number out of the first n bootstrap values which exceed $L(\lambda_0)$, $0 \leq n \leq N-1$. Then H_0 is rejected if and only if $G_{N-1} \leq C$ or $G_{N-1} \geq N-C-1$. If, after n bootstrap values have been computed, $G_n > C$ and $G_n < n-C$ eventual acceptance of H_0 is inevitable and the bootstrap simulations can be curtailed at this point. Also, if $G_n \leq C+n+1-N$ or $G_n \geq N-C-1$, simulation can be curtailed as it is inevitable that H_0 will be rejected. These rules define a sequential test with the stopping boundary shown in Figure 1a. This sequential rule could be used in analysing a set of data, but only if acceptance or rejection

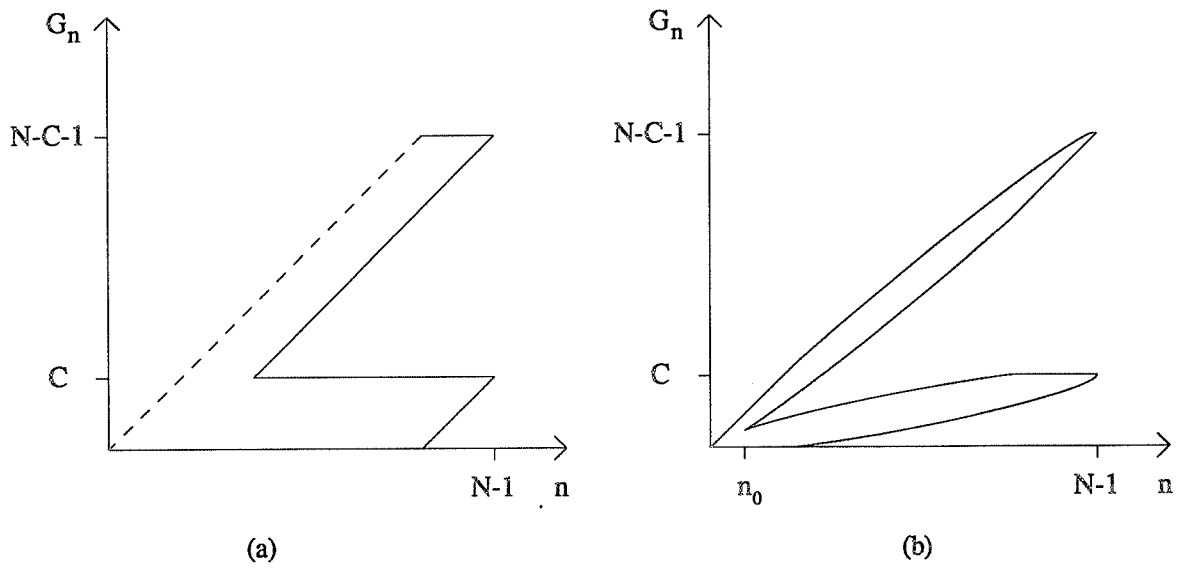

Figure 1. *Boundaries of sequential bootstrap tests designed to reject H_0 if $G_{N-1} \leq C$ or $G_{N-1} \geq N-C-1$. Here G_n is the number out of the first n bootstrap statistics exceeding the observed value. In* (a) *the test is curtailed only if the final result is completely inevitable. Much narrower boundaries are achieved in* (b) *by allowing a small probability of disagreement with the fixed sample test using all $N-1$ bootstrap values.*

of H_0 at a single significance level were the sole question of interest. Its value in a simulation study assessing the accuracy of an approximation to the null distribution is much clearer, since there, concentration on a single significance level is quite appropriate.

Note that here, the question being tested sequentially is whether or not $C+1 \le G_{N-1} \le N-C-2$, i.e., the intention is to produce the same outcome as the fixed sample bootstrap test with the full $N-1$ simulated values. An alternative approach, which we shall not pursue here, would be to test sequentially the null hypothesis that $L(\lambda_0)$ lies between the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution.

The deterministic curtailment described above does not take full advantage of the sequential setting. If we allow a very small probability of reaching the opposite conclusion from a fixed sample test using all $N-1$ bootstrap values, substantially greater reductions in the numbers of bootstraps can be achieved; for the methods we consider these probabilities are of the order of 10^{-5}, so their effect on estimated error probabilities with standard errors of around 10^{-3} really is negligible. We shall consider tests of the following general form, a typical example of which appears in Figure 1b. At stage n,

if $G_n \le a_n$	*conclude* $G_{N-1} \le C$
if $G_n \ge d_n$	*conclude* $G_{N-1} \ge N-C-1$
if $n \ge n_0$ and $b_n \le G_n \le c_n$	*conclude* $C < G_{N-1} < N-C-1$
else	*continue by generating the* $(n+1)$*th bootstrap.*

Here $a_{N-1}=C$, $b_{N-1}=C+1$, $c_{N-1}=N-C-2$ and $d_{N-1}=N-C-1$. We restrict attention to tests which are symmetric in that $a_n+d_n=b_n+c_n=n$ for each $n=1,\ldots,N-1$. Note the role of n_0, the minimum number of bootstraps required to conclude $C < G_{N-1} < N-C-1$.

To design and evaluate such tests, we consider the distribution of $\{G_n; n=1,\ldots,N-2\}$ given G_{N-1}. Marginally, each G_n has a hypergeometric distribution with parameters $N-1$, G_{N-1} and n; if G_{N-1}, $N-1-G_{N-1}$, n and $N-1-n$ are all large, this can be approximated by the normal distribution,

$$N\left(\frac{nG_{N-1}}{N-1}, \frac{G_{N-1}(N-1-G_{N-1})\,n\,(N-1-n)}{(N-1)^2\,(N-2)}\right). \quad (5.1)$$

The joint distribution of $\{G_n; n=1,\ldots,N-1\}$ given G_{N-1} forms a random walk from $G_0=0$ to $G_{N-1}=G_{N-1}$ with

$$P(G_{n+1} = G_n+1) = (G_{N-1}-G_n)/(N-1-n),$$
$$P(G_{n+1} = G_n) = 1-(G_{N-1}-G_n)/(N-1-n). \quad (5.2)$$

Approximate calculations could be based on the convergence of this process to a Brownian bridge as $N\to\infty$. However, such approximation is unnecessary since (5.2) allows exact numerical evaluation of any given sequential test using standard techniques as described in, for example, Schultz *et al.* (1973). It is easily seen that the largest conditional probability, given G_{N-1}, of a sequential test reaching the opposite conclusion from that of the fixed sample test with all $N-1$ bootstraps is the maximum of $P(\text{Conclude } G_{N-1} \le C \mid G_{N-1}=C+1)$ and $P(\text{Conclude } G_{N-1} > C \mid G_{N-1}=C)$; note that probabilities relating to the conclusion $G_{N-1} \ge N-C-1$ are equal by symmetry and we ignore probabilities such as $P(\text{Conclude } G_{N-1} \ge N-C-1 \mid G_{N-1}=C)$ which are smaller by an order of magnitude. This maximum is then an upper bound on the overall probability that the sequential test and fixed sample test do not agree.

The tests used in our simulation studies had the following form, which we shall refer to as "conditional repeated significance tests". Values of a_n are set at extreme points of the lower tail of the marginal distribution of $G_n \mid G_{N-1}=C+1$; if the mean of this distribution, $n(C+1)/(N-1)$, is less than 200 or greater than $C+1-200$, a_n is set at the 10^{-6} point of the hypergeometric distribution, for other cases a_n is the 10^{-7} point of the approximating normal distribution (5.1), the more extreme tail point being used to allow for any error in the normal approximation. Values of b_n, c_n and d_n are the corresponding upper tail points of $G_n \mid G_{N-1}=C$, lower tail points of $G_n \mid G_{N-1}=N-C-1$ and upper tail points of $G_n \mid G_{N-1}=N-C-2$, respectively; n_0 is the smallest value of n for which $c_n \ge b_n$. Exact calculations, using (5.2), give the worst case probability that the sequential test reaches a different conclusion from the fixed sample test, as described above. In our examples, using this form of boundary with $N=1000$, the worst case error was less than 10^{-5}. The construction of this test is in the same spirit as the stochastically curtailed tests introduced by Lan, Simon & Halperin (1982), however, there are important differences and the much narrower boundaries of our test for small n lead to major improvements in performance.

Table 1. *Average number of bootstrap evaluations under H_0 for sequential tests of the "conditional repeated significance test" form described above. The reference fixed sample test rejects H_0 if the observed statistic is one of the $N\alpha/2$ lowest or $N\alpha/2$ highest values in the combined sample of observed statistic plus $N-1$ bootstrap values.*

	N		
	100	1000	10000
$\alpha/2=0.05$	29	182	688
$\alpha/2=0.01$	9	84	295

Table 1 shows the average number of bootstrap evaluations under H_0 for conditional repeated significance tests with two-sided type I error probabilities α and number of bootstraps $N-1$ in the reference fixed sample test. Note that for $\alpha/2=0.01$ and $N=10000$, less than 3% of the fixed sample size, $N-1$, is required on average and this proportion continues to decrease as N increases. These reductions in average sample size are well in excess of those usually achieved by sequential tests. The main reason for this is that, under H_0, G_{N-1} is typically well away from the borderline values C and $N-C-1$. Since this is the situation usually encountered in simulation studies, and since a sequential stopping rule is easily added to a simulation program, we strongly recommend the routine use of sequential methods in bootstrap simulation studies. The above form of boundary was chosen for convenience and other boundaries could certainly be used instead. However, our experience suggests that any test which provides sufficient opportunity for very early stopping to accept or reject H_0 will have a very similar performance.

5.3 *Simulation results*

We first tested our proposed methods in the fixed sample setting. Two sets of survival times were generated under a proportional hazards model with hazard ratio λ_0 and it was noted whether the null hypothesis $H_0 : \lambda=\lambda_0$ was rejected in favour either of $\lambda>\lambda_0$ or of $\lambda<\lambda_0$. Results from one example are given in Table 2. In this case there were 30 observations in each group, survival was exponential with the geometric mean of the two median survival times equal to 1 and censoring was uniform on [0,1]. The average number of observed failures was around 17 for each λ_0.

It is seen from Table 2 that the normal approximation is satisfactory when $\lambda_0=1$ but deteriorates as λ_0 moves away from 1. The substantial differences between achieved and nominal error rates for $\alpha/2=0.01$ are a particular cause of concern since tests at about this level are often used at intermediate stages of sequential tests. The small sample approximation, which was implemented here with a bootstrap sample size of $N=1000$, works very well and is

Table 2. *Empirical error rates for single sample tests of* $H_0 : \lambda=\lambda_0$ *using the score statistic* (2.1) *with* (a) *the normal approximation and* (b) *the new small sample approximation described in* § 3 *with bootstrap* $N = 1000$. *The table shows separately the estimated probabilities that* H_0 *is rejected in favour of larger or smaller values of* λ *when the intended error rate is* $\alpha/2$. *Results are based on a simulation study with* 20000 *replications; standard errors are* 0.0015 *for* $\alpha/2=0.05$ *and* 0.0007 *for* $\alpha/2=0.01$.

		$\alpha/2=0.05$		$\alpha/2=0.01$	
		$\lambda>\lambda_0$	$\lambda>\lambda_0$	$\lambda>\lambda_0$	$\lambda>\lambda_0$
$\lambda_0=1$	*Normal approximation*	0.050	0.050	0.0097	0.0097
	New approximation	0.049	0.049	0.0097	0.0097
$\lambda_0=1.5$	*Normal approximation*	0.046	0.053	0.0080	0.0107
	New approximation	0.049	0.050	0.0098	0.0090
$\lambda_0=2$	*Normal approximation*	0.045	0.056	0.0067	0.0134
	New approximation	0.050	0.050	0.0094	0.0101
$\lambda_0=3$	*Normal approximation*	0.042	0.061	0.0055	0.0144
	New approximation	0.049	0.052	0.0112	0.0094

within one or two standard errors of the nominal error rates in all cases. The same comparisons have been made for examples with different survival and censoring distributions and different sample sizes, and in all cases results were similar to those reported here.

Table 3 shows simulation results for a sequential survival study. The experimental design is typical of many clinical trials, with a total duration of 5 years and subjects entering over an initial two year accrual period according to a Poisson process with rate 100 per year. On entry, each subject is randomly allocated to one of two treatment groups. The results of Table 3 are for the cases of exponential survival times and Weibull survival times with shape parameter $\rho=3$. The survival distributions follow the proportional hazards model with a hazard ratio λ_0 and the geometric mean of the two median survival times equal to 2.5 years. In addition to right censoring at each interim analysis, competing risk censoring is present with a hazard rate of 0.1. Up to 10 interim analyses take place at intervals of 6 months and these are carried out using all information available at that time on those subjects already entered into the study; thus, the amount of information available at each analysis varies from one simulation to another. In testing $H_0 : \lambda=\lambda_0$, the Slud and Wei (1982) method, as described in § 2, was used with error probabilities to be "spent" at each analysis, π_k, $k=1, \ldots, 10$, taking values corresponding to a Pocock (1977) repeated significance test with 10 equally sized groups of observations; if no failures at all had occurred at an early analysis the allocated error probability was carried forward to the next analysis. Tests using the new small sample approximation to the distribution of the score statistic (2.1) were implemented with a bootstrap sample size of 1000 at interim analyses at which 30 or fewer failures had occurred but if more than 30 failures were observed the normal approximation was used instead.

The lack of accuracy of the normal approximation, which the results for fixed sample tests had portended, is seen clearly in Table 3. In the case of Weibull failure times, for which the number of early failures is very small, the normal approximation is even unreliable for $\lambda_0 = 1$. In contrast, the new small sample approximation leads to error rates very close to their nominal levels in all cases. Similar findings for other experimental designs suggest that the new approximation is reliable quite generally and provides a widely applicable approach to testing at early analyses in sequential survival studies.

Table 3. *Empirical error rates for sequential tests of* $H_0 : \lambda=\lambda_0$ *using the score statistic* (2.1) *with* (a) *the normal approximation and* (b) *the new small sample approximation described in* § 3 *with* $N=1000$. *The table shows estimated probabilities that* H_0 *is rejected in favour of larger or smaller values of* λ *when the intended error rate is* $\alpha/2 = 0.05$. *Results are based on* 20000 *replications; standard errors are* 0.0015.

	Empirical error rates					
	$\lambda_0 = 1$		$\lambda_0 = 2$		$\lambda_0 = 3$	
	$\lambda > \lambda_0$	$\lambda < \lambda_0$	$\lambda > \lambda_0$	$\lambda < \lambda_0$	$\lambda > \lambda_0$	$\lambda < \lambda_0$
Exponential survival						
Normal approximation	0.046	0.046	0.036	0.060	0.032	0.067
New approximation	0.052	0.052	0.049	0.051	0.049	0.051
Weibull, $\rho = 3$, *survival*						
Normal approximation	0.036	0.036	0.024	0.057	0.021	0.073
New approximation	0.048	0.048	0.048	0.050	0.047	0.052

6. DISCUSSION

We have presented a small sample approximation to the distribution of the efficient score statistic for testing a hypothesised hazard ratio, λ_0, in the proportional hazards model. The method has been shown to yield error probabilities close to their nominal values in both fixed sample and sequential tests. Even when the number of observed failures is as low as, say, 5 the method still manages to achieve near nominal error rates, despite the discrete nature of the sample space (the number of failures in either group must be an integer between 0 and 5). The reason for this is the slight difference in the second term of the score statistic (2.1) produced by different patterns of censoring. There is something rather worrying here since a distinct ordering is created of points in the sample space for which the likelihood, as a function of λ, is almost identical. In certain situations one might well prefer not to separate such outcomes, for example, one would probably wish to avoid finding a two-sided significance level of 0.015 at an early analysis of a survival study where only one subject had failed and six subjects, all in the same group as the failure, had been censored before the failure time! We intend to address this problem in greater depth in a subsequent paper.

In implementing and assessing our proposed method, we have made use of "bootstrap" methods and we conclude with a few remarks on this topic. The numerical results of § 5.3 show our method to be much more accurate than many bootstrap tests. We attribute this to the fact that our bootstrap samples are actually generated under the null hypothesis $\lambda=\lambda_0$; it is our belief that, in situations where this is possible, such an approach will be preferable to direct resampling from the observed data plus allowance for the possibly non-null value of λ by pivotal techniques. For similar reasons, when constructing a bootstrap confidence interval for λ, we advocate modelling as a function of λ the probability that a bootstrap sample exceeds the observed statistic. Finally, we note the very substantial savings that can be achieved using sequential methods in conjunction with bootstrap sampling and recommend their use more generally.

REFERENCES

Barnard, G. A. (1963) Discussion of paper by M. S. Bartlett. *J. R. Statist. Soc.*, B 25, 294.

Brown, B. M. (1971) Martingale central limit theorems. *Ann. Math. Statist.*, 42, 59-66.

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

DeMets, D. L. and Gail, M. H. (1985) Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics*, 41, 1039-1044.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7, 1-26.

Efron, B. (1981) Censored data and the bootstrap. *J. Amer. Statist. Assoc.*, 76, 312-319.

Gail, M. H., DeMets, D. L. and Slud, E. V. (1982) Simulation studies on increments of the two-sample logrank score test for survival data, with application to group sequential boundaries. In *Survival Analysis*, Monograph Series 2, (J. Crowley and R. Johnson eds), 287-301. Hayward, California: IMS Lecture Notes.

Harrington, D. P., Fleming, T. R. and Green, S. J. (1982) Procedures for serial testing in censored survival data. In *Survival Analysis*, Monograph Series 2, (J. Crowley and R. Johnson eds), 269-286. Hayward, California: IMS Lecture Notes.

Jennison, C. and Turnbull, B. W. (1984) Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, 5, 33-45.

Jennison, C. and Turnbull, B. W. (1989) Interim analyses: the repeated confidence interval approach (with discussion). J. Roy. Statist. Soc., B, 51, 305-361.

Jennison, C. and Turnbull, B. W. (1991). Exact calculations for sequential t, χ^2 and F tests. To appear in *Biometrika*.

Lan, K. K. G., Simon, R. and Halperin, M. (1982) Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis*, 1, 207-219.

Marriott, F. H. C. (1979) Barnard's Monte Carlo test: how many simulations? *Appl. Statist.*, 28, 75-77.

Pocock, S. J. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191-199.

Reid, N. (1981) Estimating the median survival time. *Biometrika*, 68, 601-608.

Schultz, J. R., Nichol, F. R., Elfring, G. L. and Weed, S. D. (1973) Multiple-stage procedures for drug screening. *Biometrics*, 29, 293-300.

Slud, E. V. and Wei, L. J. (1982) Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Statist. Assoc.*, 77, 862-868.

Tsiatis, A. A. (1981) The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, 68, 311-315.

Tsiatis, A. A. (1982) Group sequential methods for survival analysis with staggered entry. In *Survival Analysis*, Monograph Series 2, (J. Crowley and R. Johnson eds), 257-268. Hayward, California: IMS Lecture Notes.