SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK

TECHNICAL REPORT NO. 623

February 1984
Revised August 1985

REPEATED CONFIDENCE INTERVALS FOR THE MEDIAN SURVIVAL TIME

by

Christopher Jennison

Bruce W. Turnbull

Typed by Kathy King

Repeated confidence intervals for the median survival time

By Christopher Jennison
Department of Mathematical Sciences
University of Durham, England

and Bruce W. Turnbull[1]
School of Operations Research and Industrial Engineering
Cornell University, Ithaca, New York 14853, USA

SUMMARY

We describe methods for the construction of a confidence interval for median survival time based on right-censored data. These methods are extended to the construction of repeated confidence intervals for the median, based on accumulating data; here, the overall probability that all intervals contain the true median is guaranteed at a fixed level. The use of repeated confidence intervals for median survival time in Phase IV clinical trials (post-marketing surveillance) is discussed.

Some key words: Censored data; Confidence intervals; Group sequential tests; Interim analyses; Kaplan-Meier estimator; Median survival time; Phase IV clinical trials.

[1]Corresponding author

1. INTRODUCTION

A confidence interval for median survival time provides a useful summary of the survival experience of a group of patients. The recent papers of Brookmeyer and Crowley (1982), Efron (1981), Emerson (1982), Reid (1981), Simon and Lee (1982) and Slud, Byar and Green (1984) describe competing nonparametric confidence intervals for the median, based on right-censored data. In §2 we propose a new form of single sample confidence interval; this interval has asymptotically correct coverage probability and Monte Carlo simulations suggest that it is superior to its competitors for small sample sizes.

In survival studies data usually accumulate over a long period of time and confidence intervals (CI's) for the median survival time may be calculated on several occasions. When these CI's are used in a decision making process there is the problem of 'multiple looks': the probability that <u>at least</u> one interval fails to contain the median may be much higher than the error rate for a single interval and the probability of an incorrect decision of a corresponding hypothesis test increases accordingly. Jennison and Turnbull (1984) have proposed methods appropriate to such situations: a maximum number of analyses, K, is fixed in advance and repeated confidence intervals (RCI's) are constructed, for which the probability of <u>all</u> K intervals containing the true median equals a chosen confidence level; since a decision may be taken to stop the experiment after less than K analyses, these RCI's are in fact conservative.

In §3 we extend the results of Meier (1975) to obtain properties of the joint distribution of Kaplan-Meier survival probability

estimates, based on accumulating data. An alternative derivation of these results, using martingale techniques, is given in Jennison and Turnbull (1985). Repeated confidence intervals for the median are derived in §4 and their small sample size performance is assessed by Monte Carlo simulation. An application of the use of RCI's is given in §5.

Although we discuss CI's for the median only, all the methods considered can be modified to give CI's for other quantiles; also, there are related methods for calculating CI's for the survival probability at a fixed time.

2. SINGLE SAMPLE CONFIDENCE INTERVALS

Slud, Byar and Green (1984), (hereafter referred to as SBG), compare various methods of obtaining CI's for median survival time. They survey asymptotic results and, in the case of no censoring, exact small sample properties; in the censored case they examine small sample properties by Monte Carlo simulation. In this section, we describe some of the competing CI's and discuss the results of SBG. In particular we note that the 'Simple Reflected' and 'Transformed Reflected' intervals have highly unequal error rates in the two tails. We propose a new form of CI which we shall subsequently adapt to the sequential setting.

Suppose that observed survival data (T_i,Δ_i), $i = 1,\ldots,N$, are generated from independent pairs of death and censoring times (X_i,Y_i) where $T_i = \min(X_i,Y_i)$ and $\Delta_i = 1$ if $X_i \leq Y_i$, 0 if $X_i > Y_i$. We write $S(t) = P(X > t)$ and assume that $S(\cdot)$ is continuous with cumulative hazard $\Lambda(\cdot)$; nothing is assumed about

the censoring times Y_i, except that they are independent of the death times. There will be no tied death times or simultaneous death and censoring times and the median μ of X_i is uniquely defined as $S^{-1}(1/2) = \inf\{t: S(t) \leq 1/2\}$.

Let r_j be the number of T_i's which are $\geq T_j$, $i = 1,\ldots,N$, that is the size of the risk set at T_j. Let $\hat{p}_j = 1$ if $\Delta_j = 0$, and $(r_j-1)/r_j$ if $\Delta_j = 1$. Also let $\hat{q}_j = 1-\hat{p}_j$. The nonparametric estimator of the survival function due to Kaplan and Meier (1958) is

$$\hat{S}(t) = \prod_{j:T_j \leq t} \hat{p}_j, \qquad t > 0.$$

This is the generalized maximum likelihood estimate of $S(\cdot)$--Johansen (1978), Scholtz (1980). Assuming some regularity conditions on the $\{Y_i\}$, Meier (1975) has shown that the variance of $\hat{S}(t)$ is consistently estimated by 'Greenwood's formula'

$$\hat{\phi}_G(t) = \hat{S}(t)^2 \sum_{j:T_j \leq t} \frac{\hat{q}_j}{r_j \hat{p}_j};$$

an alternative estimate for t near the median is

$$\tilde{\phi}_G(t) = \frac{1}{4} \sum_{j:T_j \leq t} \frac{\hat{q}_j}{r_j \hat{p}_j}.$$

We define an estimate of cumulative hazard due to Nelson (1972)

$$\hat{\Lambda}(t) = \sum_{j:T_j \leq t} \Delta_j r_j^{-1}.$$

The variance of $\hat{\Lambda}(t)$ can be estimated (see SBG) by $\hat{\phi}_G(t)/\hat{S}(t)^2$ which equals $4\tilde{\phi}_G(t)$. The inverses of the right-continuous step-functions $\hat{S}$ and $\hat{\Lambda}$ are taken to be $\hat{S}^{-1}(x) = \inf\{t: \hat{S}(t) \leq x\}$ and $\hat{\Lambda}^{-1}(x) = \inf\{t: \hat{\Lambda}(t) \geq x\}$ and we estimate the median μ by $\hat{\mu} = \hat{S}^{-1}(1/2)$.

We consider the problem of obtaining a $1-2\alpha$ level, equal tailed confidence interval for μ. This may be done by inverting a family of tests of the hypotheses $S(t) = 1/2$, for $t > 0$. Brookmeyer and Crowley (1982) invert tests which use the approximation $\hat{S}(t) \sim N(S(t), \hat{\phi}_G(t))$ to obtain the interval

$$\{t: (\hat{S}(t) - 1/2)^2 \leq \chi^2_{1,2\alpha}\, \hat{\phi}_G(t)\}. \qquad (2.1)$$

Here, $\chi^2_{1,2\alpha}$ denotes the $(1-2\alpha)$ point of a χ^2_1 distribution. (Although pathological examples can be constructed for which this does not give an interval this problem is unlikely to occur in practice and we shall ignore it.) Brookmeyer and Crowley show that this method gives asymptotically correct coverage probabilities but, for small N, error rates tend to be too high. SBG argue that Efron's (1981) bootstrapped confidence interval for μ should agree closely with that of Brookmeyer and Crowley.

The methods of Reid (1981) and Emerson (1982) approximate the distribution of $\hat{S}(t)$ by a binomial distribution and their intervals are asymptotically anti-conservative in the presence of

censoring. The simulation results of SBG show that this is a serious effect even when N is as low as 41.

Slud, Byar and Green distinguish between so called 'Test based' and 'Reflected' intervals. In fact, these 'Reflected' intervals do correspond to families of tests: Emerson approximates $N\hat{S}(t)$ as binomial $(N,S(t))$ and tests $S(t) = 1/2$; the 'Simple Reflected' interval described by SBG corresponds to tests of $S(t) = 1/2$ using a normal approximation for $\hat{S}(t)$, with the variance of $\hat{S}(t)$ estimated by $\tilde{\phi}_G(\hat{\mu})$; SBG's 'Transformed Reflected' interval corresponds to tests of $\Lambda(t) = \Lambda(\mu)$ $(= \log_e 2)$, with $\Lambda(\mu)$ estimated by $\hat{\Lambda}(\hat{\mu})$ and a normal approximation for $\hat{\Lambda}(t)$ with the variance of $\hat{\Lambda}(t)$ estimated by $4\tilde{\phi}_G(\hat{\mu})$. Of course, the error rates of these intervals are simply the Type I error probabilities of the corresponding tests at $t = \mu$.

SBG show that the Simple Reflected intervals and their Transformed Reflected intervals have asymptotically correct coverage probabilities and they investigate by simulation the error rates for $N = 21$ and 41 under various censoring schemes. From these simulation results, it appears that the Simple Reflected and Transformed Reflected CI's have close to the desired coverage probabilities but SBG do not distinguish between errors in the two tails. Our own simulations show that, in fact, these procedures exhibit a great imbalance between error rates in the two tails when censoring is present (see Tables 2.1 and 2.2). This is not acceptable in situations where the consequences of reporting intervals above and

below the true median are different; this behaviour also shows that the approximations underlying the CI's are poor (insofar as, if the approximations were good both upper and lower tail error rates would be close to their nominal values), and this casts doubt on the performance of the CI's in other situations.

Of the intervals discussed so far, those of Brookmeyer and Crowley are the most appealing although they are anti-conservative in small samples. One possible reason for this anti-conservatism is that $\hat{\phi}_G(\mu)$ underestimates the variance of $\hat{S}(\mu)$ when $\hat{S}(\mu) \neq 1/2$ --this is certainly the case in the absence of censoring when $N\hat{S}(\mu)$ ~ binomial $(N,1/2)$ and $\hat{\phi}_G(\mu) = \hat{S}(\mu)(1-\hat{S}(\mu))/N$. When testing the hypothesis $S(t) = 1/2$, $\text{var}\{\hat{S}(t)\}$ should be estimated under the assumption $S(t) = 1/2$. Let $N(t)$ be the number of $T_i \geq t$, $i = 1,\ldots,N$ and let $\mathcal{N}(t) = E\{N(t)\}$. Breslow and Crowley (1974) show that if the Y_i's are identically distributed, $\text{var}\{\hat{S}(t)\} \sim -S(t)^2 \int_0^t dS(u)/\{\mathcal{N}(u)S(u)\}$ as $\mathcal{N}(t) \to \infty$ and they claim that this result still holds under different censoring schemes. (See also more recent results by Gill (1983).) We therefore wish to estimate $\text{var}\{\hat{S}(t)\}$ by an estimate of

$$S(t)^2 \int_0^t \frac{-dS(u)}{\mathcal{N}(u)S(u)} . \tag{2.2}$$

In Greenwood's formula this is accomplished by substituting the estimates $\hat{S}(t)$ for $S(t)$, $\hat{q}_j/\hat{p}_j$ for $-dS(T_j)/S(T_j)$ and r_j for (T_j). Under the constraint $\hat{S}(t) = 1/2$, the generalized nonparametric maximum likelihood estimate of $S(u)$, $(0 < u < t)$, is $\tilde{S}(u) = \Pi_{j:T_j \leq u} \tilde{p}_j$, where $\tilde{p}_j = 1$ if $\Delta_j = 0$, $(r_j+\lambda-1)/(r_j+\lambda)$

if $\Delta_j = 1$ and λ is chosen so that $\tilde{S}(t) = 1/2$. (See Thomas and Grunkemeier, 1975). We define $\tilde{q}_j = 1-\tilde{p}_j$. Estimating $N(T_j)$ by $r_j\tilde{S}(T_j)/\hat{S}(T_j)$ compensates for the occurrence of higher or lower than typical survival up to time T_j; this estimate may be derived more rigorously using the Markov property of the process $\{N(u), u \geq 0\}$ described by Meier (1975, Section 2.1). Substituting these constrained estimates into (2.2) gives the 'constrained' estimate of $\text{var}\{\hat{S}(t)\}$

$$\hat{\phi}_c(t) = \frac{1}{4} \sum_{j:T_j \leq t} \frac{\hat{S}(T_j)\tilde{q}_j}{r_j\tilde{S}(T_j)\tilde{p}_j} .$$

This estimate has been used by Thomas and Grunkemeier (1975) to obtain CI's for $S(t)$ at fixed time t; it has the desirable property that it gives the correct binomial variance of $\hat{S}(\mu)$, namely $1/4n$, in the absence of censoring.

Using $\hat{\phi}_c(t)$ as an estimate of $\text{var}\{\hat{S}(t)\}$ in the normal approximation gives the CI for μ:

$$\{t: (\hat{S}(t) - 1/2)^2 \leq \chi^2_{1,2\alpha}\hat{\phi}_c(t)\}. \qquad (2.3)$$

We shall refer to this as the Constrained Variance interval. The error rates of this CI are those of the test which rejects the null hypothesis $S(\mu) = 1/2$ when $(\hat{S}(\mu) - 1/2)^2 > \chi^2_{1,2\alpha}\hat{\phi}_c(\mu)$: the error rates are asymptotically correct and small sample properties are given in Tables 2 and 3 of Thomas and Grunkemeier (1975); for censored samples the error rates are considerably nearer to their

nominal values than those of the test using $\hat{\phi}_G(\mu)$ as an estimate of $\text{var}\{\hat{S}(\mu)\}$, which corresponds to Brookmeyer and Crowley's CI for μ.

Anderson, Bernstein and Pike (1982) have proposed CI's for survival quantiles. Their 'Rothman-Wilson' interval is obtained by replacing $\hat{\phi}_G(t)$ in (2.1) or $\hat{\phi}_c(t)$ in (2.3) by

$$\hat{\phi}_R(t) = \frac{1/2 \cdot 1/2}{\hat{S}(t)(1-\hat{S}(t))} \sum_{j:T_j \leq t} \frac{\hat{q}_j}{r_j \hat{p}_j} .$$

This gives asymptotically correct error rates and $\hat{\phi}_R(t)$ is the correct binomial variance in the absence of censoring. However, Anderson, Bernstein and Pike report that neither this nor any of the related methods which they consider, show any improvement in error rates for testing $S(\mu) = 1/2$ over a test which uses $\hat{\phi}_G(\mu)$ as an estimate of $\text{var}\{\hat{S}(\mu)\}$. If this is so, then the error rates of their CI's for μ are no better than those of the Brookmeyer and Crowley CI.

Tables 2.1 to 2.3 show the results of simulation experiments comparing four of the above methods. The survival and censoring distributions are those used by Slud, Byar and Green. Survival distributions are exponential with hazard rate 1, denoted by Exp(1); Weibull with scale parameter 1 and shape parameter 0.7 denoted by Weib(1,.7); and Weib(1,1.5). Censoring distributions are Exp(1); uniform on the interval 0 to 2, denoted Unif(0,2); Exp(.3); and Unif(0,4.5). The entries of Tables 2.1 and 2.2 are the relative frequencies with which the intervals failed to contain the true median

of the survival distribution for sample sizes 21 and 41 respectively; these errors are broken down according to whether the interval was above or below the true median. Each entry of Tables 2.1 and 2.2 is based on 6000 replications. Table 2.3 gives the average observed length of CI's for sample sizes 21 and 41; calculation of a CI requires considerably more work than testing that it contains the true median and these entries are based on 1000 replications.

As in SBG (Section 4) to avoid degeneracy when $r_j = 1$, we have replaced (r_j-1) by r_j in the final term of the summation for $\tilde{\phi}_G(\hat{\mu})$ used in the Simple Reflected and Transformed Reflected intervals. In choosing λ to satisfy $\tilde{S}(t) = 1/2$ a bisection method was used with initial interval $(0,N(t_1))$, where t_1 is the smallest observed death time (min T_i for which $\Delta_i = 1$), when $\hat{S}(t) < 1/2$, and $(2\text{-}N(t_m),0)$, where t_m is the largest observed death time $\leq t$, when $\hat{S}(t) > 1/2$. A number of problems arise when the sample size is small or censoring is heavy: the Kaplan-Meier estimate $\hat{S}(t)$ is undefined for $t > \max_i(T_i)$ and $\hat{\phi}_c(t)$ is undefined for $t <$ the smallest observed death time. If $\hat{S}(t) = 1$, one may test $S(t) = 1/2$ by referring $N\hat{S}(t)$ to a binomial $(N(t),1/2)$ distribution; for small sample sizes this can give a CI with lower end point = 0. If the hypothesis $S(t_{max}) = 1/2$, where t_{max} is the largest observed death time, is accepted then the upper end point of a CI for μ should be ∞. To simplify calculations we have adopted the approach of SBG who treat the highest survival time as uncensored, whatever its true nature; we have also imposed their restriction that CI's lie

between the lowest observed death time and the largest observed time, T_i. The effect of these modifications on the error rates is negligible in the examples considered (the resulting increase in observed error rates is approximately 0.001 in the worst case). Strictly speaking, all methods give CI's with infinite expected length when censoring is present; but with the imposed restriction, the average observed length is a simple and fair summary of the distribution of the length of a CI.

From Tables 2.1, 2.2 we see that the Constrained Variance interval gives both upper and lower tail error probabilities closest to the nominal α level. From Table 2.3, we see that the expected lengths of the Constrained Variance intervals are generally shorter than of either the Simple Reflected or Transformed Reflected intervals. As previously mentioned, these two latter methods can give severely asymmetric error probabilities in the two tails. The Brookmeyer-Crowley intervals yield the shortest intervals, but this is to be expected since they are severely anti-conservative for these sample sizes (N = 21,41).

Based on the superior performance of the Constrained Variance intervals found in the simulation experiments, we shall concentrate most of our attention on them in our discussion of sequential confidence intervals that follows in the next sections.

3. ASYMPTOTIC JOINT DISTRIBUTION OF KAPLAN-MEIER SURVIVAL ESTIMATES BASED ON ACCUMULATING DATA

Breslow and Crowley (1974) derive the asymptotic joint distribution of $\hat{S}(t)$, $t > 0$, for a single sample experiment with independent identically distributed censoring times. Meier (1975) treats the case of fixed censoring times but his results hold under any form of random censoring which is independent of survival and for which the realized censoring times satisfy the appropriate asymptotic conditions with probability 1. We shall extend Meier's methods to obtain the asymptotic joint distribution of $\{\hat{S}(t), t > 0\}$ over repeated analyses of accumulating data; at successive analyses there will be both new observations and additional information on previously right censored observations.

Suppose a total of K analyses are performed and N subjects are involved. Let $Y_{i,k}$ $(1 \leq i \leq N, 1 \leq k \leq K)$ be the potential censoring time for the i'th subject at the k'th analysis, with $Y_{i,k} = 0$ if the i'th subject has not yet entered the study; although they may have arisen in a random manner both N and the $Y_{i,k}$ can be regarded as fixed. Both end of study and competing risk censoring may be present so, for a fixed i, $Y_{i,k}$ will increase with k up to a maximum value (possibly infinite) equal to the competing risk censoring time. Let X_i be the survival time of the i'th subject and define $T_{i,k} = \min(X_i, Y_{i,k})$ and $\Delta_{i,k} = 1$ if $X_i \leq Y_{i,k}$, 0 if $X_i > Y_{i,k}$ $(1 \leq i \leq N, 1 \leq k \leq K)$. As in §2, suppose that X_i's are independent with continuous survival function $S(\cdot)$ and

median μ; nothing is assumed of the $Y_{i,k}$ except that they are independent of the X_i's. For each k $(1 \leq k \leq K)$ we define the following functions of time t:

$$M_k(t) = \text{number of } Y_{i,k} \geq t, \quad 1 \leq i \leq N$$
$$N_k(t) = \text{number of } T_{i,k} \geq t, \quad 1 \leq i \leq N$$
$$N_k(t) = E\{N_k(t)\} = M_k(t)S(t).$$

Suppose there are m distinct non-zero values taken by the $Y_{i,k}$ and label these $L_1,\ldots,L_m$ where $0 < L_1 < L_2 < \ldots < L_m$. For a given value of t, let h be the index such that $L_h < t \leq L_{h+1}$ where $L_0 = 0$ and $L_{m+1} = \infty$. Let $I_1,\ldots,I_{h+1}$ be the half-open intervals

$$I_1 = (0,L_1], \; I_2 = (L_1,L_2],\ldots,I_h = (L_{h-1},L_h], \; I_{h+1} = (L_h,t],$$

and let

$$p_1 = S(L_1), \; p_2 = S(L_2)/S(L_1),\ldots,p_h = S(L_h)/S(L_{h-1}), \; p_{h+1} = S(t)/S(L_h)$$

i.e. p_j = conditional probability of survival over the interval I_j, given survival to the start of that interval. Let $q_j = 1-p_j$. It follows that

$$S(t) = \prod_{j=1}^{h+1} p_j.$$

We define also, for each k $(1 \leq k \leq K)$: $N_{j,k} = N_k(L_{j-1}+)$; $D_{j,k}$ = number of $T_{i,k} \in I_j$ with $\Delta_{i,k} = 1$ $(1 \leq i \leq N)$; and $N_{j,k} = E\{N_{j,k}\}$.

At the k'th analysis, the Kaplan-Meier estimate of $S(\cdot)$ is

$$\hat{S}_k(t) = \begin{cases} \prod_{i=1}^{h+1} \hat{p}_{i,k} & t \leq \max_i (T_{i,k}) \\ 0 & t > \max_i (T_{i,k}) \end{cases}$$

where

$$\hat{p}_{i,k} = \frac{N_{i,k} - D_{i,k}}{N_{i,k}} .$$

Letting

$$C_k(t) = \sum_{i=1}^{h+1} \frac{q_i}{N_{i,k} p_i} ,$$

Meier (1975, page 78) has shown that if $C_k(t) e^{N_k(t)} \to \infty$ as $N_k(t) \to \infty$, then

$$\lim_{N_k(t) \to \infty} E\left\{ \left[\frac{\hat{S}_k(t) - S(t)}{S(t)\sqrt{C_k(t)}} \right] 2 \right\} = 1;$$

and (page 82) that if $N_k^2(t) C_k(t) \to \infty$, then the random variable

$$\frac{\hat{S}_k(t) - S(t)}{S(t)\sqrt{C_k(t)}}$$

converges in distribution to $N(0,1)$ as $N_k(t) \to \infty$.

Consider a sequence of experiments of increasing size, for which $N_k(t) \to \infty$ for each $1 \leq k \leq K$ and there exist constants γ_{ij} $(1 \leq i,j \leq K)$ such that

$$\frac{\sqrt{C_{k_2}(t)}}{\sqrt{C_{k_1}(t)}} \to \gamma_{k_1 k_2} \quad (1 \leq k_1, k_2 \leq K).$$

We show in the Appendix that if, in this asymptotic setting, $C_K(t)e^{N_1(t)} \to \infty$, then

$$\lim E\left\{\frac{(\hat{S}_{k_1}(t)-S(t))(\hat{S}_{k_2}(t)-S(t))}{S^2(t)C_{k_2}(t)}\right\} = 1, \quad \text{for all} \quad 1 \le k_1 \le k_2 \le K \tag{3.1}$$

also, if $N_k^2(t)C_k(t) \to \infty$ for all $1 \le k \le K$, the random vector $\underset{\sim}{Z}$ defined by

$$Z_k = \frac{\hat{S}_k(t)-S(t)}{S(t)\sqrt{C_k(t)}}, \quad 1 \le k \le K, \tag{3.2}$$

converges to a multivariate normal distribution with mean 0 and covariance matrix Σ, where $\Sigma_{ij} = \gamma_{ij}$ for all $1 \le i \le j \le K$.

Asymptotic properties of repeated significance tests of the hypothesis $S(t) = \theta$, based on the estimates $\hat{S}_k(t)$, may be derived from the above results. For these purposes, and for a fixed value of t, the assumption $\underset{\sim}{Z} \sim N(0,\Sigma)$ is equivalent to assuming the joint distribution of the $\hat{S}_k(t)/C_k(t)$, $(1 \le k \le K)$ to be that of a Brownian motion, with drift $S(t)$ and variance $S^2(t)$ per unit time, observed at times $1/C_k(t)$, $(1 \le k \le K)$.

4. REPEATED CONFIDENCE INTERVALS FOR THE MEDIAN

In a single sample experiment a confidence interval is obtained by inverting a family of hypothesis tests; similarly, in a

sequential setting, repeated confidence intervals can be obtained by inverting a family of sequential tests (see Jennison (1982), Jennison and Turnbull (1984)). In this Section we derive RCI's for the median from families of group sequential tests of the hypotheses $S(t) = 1/2, \quad t > 0$. RCI's for $S(t)$, for a fixed t, can be obtained in a similar way by inverting group sequential tests of the hypotheses $S(t) = \theta, \quad 0 \leq \theta \leq 1$.

Consider testing sequentially the hypothesis $S(t) = 1/2$ for a fixed value of t. According to the asymptotic approximations of §3, we can regard the sequence $\hat{S}_k(t)/C_k(t), \quad (1 \leq k \leq K)$ as being generated by observing a Brownian motion at times $C_k(t)^{-1}$. Thus, the repeated significance tests proposed by Pocock (1977) and O'Brien and Fleming (1979) are applicable if the increments in $C_k(t)^{-1}$ between successive analyses are equal. Both tests are of the form: Reject H_0: $S(t) = 1/2$ at analysis k if H_0 is rejected under a fixed sample size test at significance level α_k', $1 \leq k \leq K$; K is a fixed number and the $\{\alpha_k'\}$ are chosen to give a specified overall error rate; in Pocock's test the $\{\alpha_k'\}$ are all equal, whereas the O'Brien/Fleming $\{\alpha_k'\}$ increase with k. (See Tables 4.3(a), (b) for $\alpha = .05$ and $k = 5,10$). We shall use the fixed sample test based on the asymptotic normality of $\hat{S}_k(t)$ and the constrained estimate of $\text{var}\{\hat{S}_k(t)\}$, described in §2, which we shall call $\hat{\phi}_{c,k}(t)$. The repeated significance tests with approximate overall error rates 2α are then

Pocock: reject $S(t) = 1/2$ in favour of $\hat{S}(t) > 1/2$, at analysis k, if

$$\hat{S}_k(t) - 1/2 > Z_p(K,\alpha)\sqrt{\hat{\phi}_{c,k}(t)},$$

reject $S(t) = 1/2$ in favour of $\hat{S}(t) < 1/2$, at analysis k, if

$$\hat{S}_k(t) - 1/2 < -Z_p(K,\alpha)\sqrt{\hat{\phi}_{c,k}(t)};$$

O'Brien/Fleming: reject $S(t) = 1/2$ in favour of $\hat{S}(t) > 1/2$, at analysis k, if

$$\hat{S}_k(t) - 1/2 > Z_B(K,\alpha)\sqrt{\hat{\phi}_{c,k}(t)/k},$$

reject $S(t) = 1/2$ in favour of $\hat{S}(t) < 1/2$, at analysis k, if

$$S_k(t) - 1/2 < -Z_B(K,\alpha)\sqrt{\hat{\phi}_{c,k}(t)/k}.$$

Values of $Z_p(K,\alpha)$ and $Z_B(K,\alpha)$ for $K = 2,3,5,10$ and $\alpha = 0.005$, 0.025, 0.05 are given in Table 4.1 (taken from Jennison and Turnbull (1984, Table 1).)

[Table 4.1 about here.]

Suppose that one form of test and a fixed value of K are selected before the start of an experiment. Let $[\underline{\mu}_k, \overline{\mu}_k) = \{t$: do not reject the hypothesis $S(t) = 1/2$ at analysis $k\}$, $1 \leq k \leq K$. Then $\{[\underline{\mu}_k, \overline{\mu}_k]; 1 \leq k \leq K\}$ is a sequence of RCI's for μ and the property

$$P\{\mu \in [\underline{\mu}_k, \overline{\mu}_k] \text{ for each } 1 \leq k \leq K\} \cong 1-2\alpha$$

is inherited from the original sequential tests. In fact, since the probability of both $\mu < \underline{\mu}_{k_1}$ and $\mu > \overline{\mu}_{k_2}$ for some $1 \leq k_1, k_2 \leq K$ is negligible (a typical value is 0.0001 when $\alpha = 0.05$), we have

$$P\{\mu < \underline{\mu}_k \text{ for any } 1 \leq k \leq K\} \cong \alpha \tag{4.1}$$

and

$$P\{\mu > \overline{\mu}_k \text{ for any } 1 \leq k \leq K\} \cong \alpha, \tag{4.2}$$

If an experiment may stop early, without all k RCI's being calculated, the probability of observing an interval which fails to contain the true median is reduced. The decision to terminate an experiment early may in fact be based on one of the intervals $[\underline{\mu}_k, \overline{\mu}_k]$ but (4.1) and (4.2) will still hold with $\lesssim$ in place of $\cong$, whatever system of early stopping is used.

The above development requires equal increments in $C_k(t)^{-1}$ $(1 \leq k \leq K)$ between successive analyses. Since the error rates for a sequence of RCI's are those of the corresponding sequential test of $S(t) = 1/2$ at $t = \mu$, Equations (4.1) and (4.2) will hold if the increments in $C_k(\mu)^{-1}$ are equal. Suppose subjects enter a study

at a steady rate, either at equal time intervals or according to a Poisson process, and suppose each subject is exposed to the same form of competing risk censoring. Let the calendar times of analyses, measured from the start of the study, be $\tau_0+k\tau_1$, $1 \leq k \leq K$; end of study censoring will occur at each analysis in the obvious way. If $t < \tau_0$ then $N_k(t) = a(t) + kb(t)$ for some constants $a(t)$ and $b(t)$. If $\mu < \tau_0$ and $\tau_1 \gg \tau_0$, so that $b(t) \gg a(t)$ for $0 < t \leq \mu$, then $N_k(t)$ is approximately proportional to k for $0 < t \leq \mu$ and it follows that $C_k(\mu)^{-1}$ is approximately proportional to k. Pocock (1977, page 197) notes that group sequential procedures may be robust to variations in the group size, which correspond, in this setting, to unequal increments in $C_k(\mu)^{-1}$; thus, in the above situation, with $\tau_1 \gg \tau_0$, Equations (4.1) and (4.2) will hold approximately. In fact, the results of Monte Carlo simulations suggest these equations hold quite accurately under much weaker conditions and a simple but adequate practical approach would be to take $\tau_0 = \mu_0$ and $\tau_1 = \mu_0$ or $\mu_0/2$, say, where μ_0 is some initial estimate of μ.

We have defined RCI's based on the Constrained Variance method; however, all of the CI methods described in §2 correspond to hypothesis tests and can therefore be used to define RCI's in a completely analogous fashion.

Tables 4.2(a)-(d) give simulation results for overall error rates of 5- and 10-look procedures with α = .005, .025, .05, for the Constrained Variance, Brookmeyer-Crowley, Simple Reflected and

Transformed Reflected methods. The survival distributions are the same as those used by SBG and in §2, namely Exp(1), Weib(1,.7), Weib(1,1.5) with no competing risk censoring. (Of course at each look, there is "end-of-study" censoring present). For the Exp(1) survival distribution, results are also shown for the case of competing risk censoring with an Exp(.3) distribution. The tables are based on 5000 replications and assume interim results are calculated at times $t_k = 1 + (.5)k$ for $k = 1,2,\ldots,K$ where $K = 5$ or 10. The accrual rate is 20 subjects per unit time. This led to moderate sample sizes which might be typical for a clinical trial. More extensive simulations for different situations were performed and results were similar; details are available from the authors upon request. Tables 4.3(a),(b) show the estimated error rates of the CI's at each interim analysis, for the first situation considered in Table 4.2, namely Exp(1) survival distribution with no competing risk censoring. Results for the other situations were similar.

The implications of these results are similar to those of §2 where the case $K = 1$ was discussed. The Brookmeyer-Crowley RCI's tend to be anti-conservative especially for the Pocock construction where non-normality at the early analyses (smaller sample sizes) affects the overall error rates more than for the O'Brien and Fleming method (See Tables 4.3). The Transformed Reflected and to a lesser extent the Simple Reflected RCI's give quite unequal error rates in the two tails. On the other hand, the Constrained Variance RCI's give upper and lower

tail error rates close to the nominal value in all cases shown, and in addition the error rates are close to their nominal values, α_k', at each individual analysis--see Tables 4.3. This fact suggests that the procedure is robust to non-equal increments in $\{C_k(t)^{-1},\ 1 < k < K\}$.

In a recent paper, Lan and DeMets (1983) have suggested an alternative approach to group sequential hypothesis tests. This approach is a more flexible one, in which the maximum number of analyses K and their times t_k, $1 \leq k \leq K$, need not be determined in advance. The approach also avoids the technical problem of requiring equal increments in $\{C_k(t)^{-1}\}$.

5. APPLICATION

Repeated confidence intervals for the median or other quantile of a distribution are useful in a study with a single treatment group. Such studies are conducted in post-marketing surveillance, or "Phase IV" clinical trials, both to check the efficacy of a drug and to investigate the level of adverse reactions. The distribution of overall survival time or duration of remission may be compared to that observed in earlier controlled studies, to establish the treatment's efficacy for the wider patient population; the level of an adverse reaction may be assessed by RCI's for the median, or other quantile, of time until its onset.

The use of RCI's when accumulating data is analyzed repeatedly avoids the 'multiple looks' problem which can lead to too many falsely

significant findings. In the broader decision making process, RCI's can be used to summarize information on a particular feature and since their properties do not rely on the rigid observation of a stopping rule, they allow more flexibility than traditional sequential tests (see Jennison and Turnbull, 1984).

We illustrate the use of RCI's for the median with data from a clinical trial in the treatment of carcinoma of the oropharynx (Kalbfleisch and Prentice, 1980, Data Set II, pages 225-229). We use the data purely to provide a numerical example of our methods; thus, although patients were randomized between two treatments, we regard all patients as a single group. A total of 195 patients entered the study over a five year period. The time of entry and survival time at the end of the study are recorded for each patient; both end of study censoring and loss to follow up censoring are present. We have calculated RCI's for median survival time as if the accumulating data had been examined repeatedly as the experiment progressed; at each analysis, end of study censoring was imposed according to the time of that analysis. Calculations were performed under several sequential schemes and the results are summarized in Table 5. This table shows Constrained Variance RCI's with an overall confidence level of 90% under procedures with 5 and 10 analyses; RCI's based on both the Pocock and O'Brien/Fleming sequential tests are given. In the final column are the CI's which would be obtained if a fixed sample analysis, unadjusted for 'multiple looks', were performed at each stage. At each analysis, the number of patients who had entered the

study by that time and the number of deaths observed before that time are shown; patients who had entered the study but not died at the time of analysis are censored at that analysis. Note that the Pocock RCI's are narrower than the O'Brien/Fleming RCI's at the early analyses but the reverse is true at later analyses; in both the sequential analyses the final O'Brien/Fleming RCI is only slightly wider than the CI (376,532) from the single analysis at 2160 days.

Had this data arisen in a post-marketing surveillance study of a drug, RCI's for median survival time could have been used to summarize the survival experience of the patients. Suppose, for example, that observed survival times were lower than expected: an RCI for the median would provide a numerical summary of this fact to be used in conjunction with information on, say, the level of adverse reactions in reaching a decision on whether or not to withdraw the drug. Such a decision must be taken in a sequential setting but the complexity of the situation precludes the use of a sequential procedure with a rigid stopping rule; the use of repeated confidence intervals with their inherent flexibility overcomes this problem.

This research was supported in part by the National Institutes of Health.

APPENDIX

The results stated in Section 3 are obtained by applying the methods of Meier (1975) in a sequential setting. For computational convenience we introduce the 'extended product limit estimator' $S_k^*(\cdot)$, defined by

$$S_k^*(t) = \prod_{i=1}^{h+1} p_{i,k}^*, \quad \text{where} \quad p_{i,k}^* = \frac{N_{i,k}^* - D_{i,k}^*}{N_{i,k}^*},$$

$$N_{i,k}^* = \begin{cases} N_{i,k} & \text{if } N_{i,k} > 0 \\ 1 & \text{if } N_{i,k} = 0, \end{cases}$$

$$D_{i,k}^* = \begin{cases} D_{i,k} & \text{if } N_{i,k} > 0 \\ d_{i,k} & \text{if } N_{i,k} = 0 \end{cases}$$

and $d_{i,k}$ is a $\{0,1\}$ random variable, independent of the sample, with $P\{d_{i,k} = 0\} = p_i$. $S^*(t)$ is identical with $\hat{S}(t)$, except when $t > \max_i T_{i,k}$, and $S^*(t)$ is unbiased for every t. In deriving (3.1) we make the following modification: if $k_1 < k_2$ and $N_{i,k_1} = 0$, set $p_{i,k_1}^* = p_{i,k_2}^*$ --this ensures that $\mathrm{cov}(p_{i,k_1}^*, p_{i,k_2}^* | N_{i,k_1}, N_{i,k_2}) = \mathrm{var}(p_{i,k_2}^* | N_{i,k_2})$.

Heuristically, using the 'δ method' (Rao, 1973, page 388) twice gives

$$\operatorname{cov}(S^*_{k_1}(t),S^*_{k_2}(t)) \simeq S(t)^2 \sum_{i,j=1}^{h+1} \frac{\operatorname{cov}(p^*_{i,k_1},p^*_{j,k_2})}{p_i p_j}$$

$$= S(t)^2 \sum_{i=1}^{h+1} \frac{\operatorname{var}(p^*_{i,k_2})}{p_i^2}$$

$$\simeq \operatorname{var}(S^*_{k_2}(t)) \simeq S(t)^2 C_{k_2}(t), \quad \text{for} \quad k_1 < k_2.$$

However, in the asymptotic setting, h, the number of censoring times $< t$, will usually increase to ∞ as the sample size increases. The same problem arises in a single sample experiment; Meier ignores this in his proofs but they can be corrected by a few simple modifications. Meier's proof of proposition (c) on page 76 of his paper can be extended to give a rigorous proof that if $0 < S(t) < 1$,

$$\lim_{N_{k_2}(t)\to\infty} E\left\{\frac{(S^*_{k_1}(t)-S(t))(S^*_{k_2}(t)-S(t))}{S^2(t)C_{k_2}(t)}\right\} = 1, \quad \text{for} \quad k_1 < k_2.$$

A simple calculation shows $E\{p^*_{i,k_1},p^*_{i,k_2}|N_{i,k_1},N_{i,k_2}\} = p_i^2/\{1+q_i/(N^*_{i,2}p_i)\}$. It is then necessary to show that $\max_{1\le i\le h} N_{i,k_2} P\{N^*_{i,k_2} < (1-\varepsilon)N_{i,k_2}\} \to 0$ as $N_{k_2}(t) \to \infty$, even if $h \to \infty$; but, $P\{N^*_{i,k_2} < (1-\varepsilon)N_{i,k_2}\}$ is of the order $\exp(-\alpha N_{i,k_2})$ as $N_{i,k_2} \to \infty$, for some $\alpha > 0$ and the result follows. Together with the fact $P\{S^*_k(t) \neq \hat{S}(t)\} \le \exp\{-N_k(t)\}$, this establishes (3.1).

To prove that $\underset{\sim}{Z}$ defined by (3.2) converges to a multivariate normal distribution we extend Meier's proof (pages 80-83) that $(\hat{S}(t)-S(t))/S(t)\sqrt{C(t)} \rightarrow N(0,1)$. According to the Multivariate Central Limit Theorem (Rao, 1973, page 128) it is sufficient to show that $\sum_{k=1}^{K} \lambda_k Z_k$ converges to the appropriate univariate normal distribution for any $\lambda_1,\ldots,\lambda_K$. The joint distribution of the Z_k $(1 \leq k \leq K)$ is approximated by that of the random variables

$$\frac{1}{\sqrt{C_k(t)}} \sum_{i=1}^{h+1} \sum_{r=1}^{N'_{i,k}} \frac{X_{ir}-p_i}{N_{i,k}p_i}, \quad 1 \leq k \leq K,$$

where X_{ir} are independent binomial $(1,p_i)$ random variables and $N'_{i,k}$ = the integer part of $N_{i,k}+1$. Convergence of a linear combination of these random variables to the correct normal distribution is proved using the Liapunov Central Limit Theorem.

We note that these results can be extended one stage further using the same method of proof. Let $x \wedge y$ and $x \vee y$ denote $\min(x,y)$ and $\max(x,y)$ respectively. If $S(T) > 0$ and $C_K(t)e^{N_1(T)} \rightarrow \infty$ for all $0 < t < T$, then for any $0 < u,v \leq T$ with $S(u),S(v) > 0$

$$\lim \left[\frac{\mathrm{Cov}\{\hat{S}_{k_1}(u),\hat{S}_{k_2}(v)\}}{S(u)S(v)C_{k_1 \vee k_2}(u \wedge v)} \right] = 1.$$

Also, if $0 < t_1,t_2,\ldots,t_K < T$ and the random vector $\underset{\sim}{W}$ is defined by

$$W_k = \frac{\hat{S}_k(t_k)-S(t_k)}{S(t_k)\sqrt{C_k(t_k)}}, \quad 1 \leq k \leq K,$$

then, if $N_k^2(t)C_k(t) \to \infty$ for all $1 \leq k \leq K$ and $0 < t \leq T$, W_k converges to a multivariate normal distibution with mean 0 and covariance matrix Σ, where

$$\Sigma_{k_1 k_2} = \lim \frac{C_{k_1 \vee k_2}(t_{k_1} \wedge t_{k_2})}{\sqrt{C_{k_1}(t_{k_1})}\sqrt{C_{k_2}(t_{k_2})}},$$

provided this limit exists. If the W_k correspond to estimates $\hat{S}(t_k)$, for different values t_k, but for a single analysis these results reduce to those of Meier (part (b) of the second Theorem on page 78 and the Corollary on page 83).

REFERENCES

ANDERSON, J.R., BERNSTEIN, L. & PIKE, M.C. (1982). Approximate confidence intervals for probabilities of survival and quantiles in life-table analysis. Biometrics 38, 407-416.

BRESLOW, N. & CROWLEY, J. (1975). A large sample study of the life table and product limit estimates under censorship. Ann. Statist. 437-453.

BROOKMEYER, R. & CROWLEY, J. (1982). A confidence interval for the median survival time. Biometrics 38, 29-41.

EFRON, B. (1981). Censored data and the bootstrap. J. Am. Statist. Assoc. 76, 312-319.

EMERSON, J. (1982). Nonparametric confidence intervals for the median in the presence of right censoring. Biometrics 38, 17-27.

GILL, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. Ann. Statist. 11, 49-58.

JENNISON, C. (1982). Sequential methods for medical experiments. Ph.D. thesis, Cornell University.

JENNISON, C. & TURNBULL, B.W. (1984). Repeated confidence intervals for group sequential clinical trials. Contr. Clinical Trials 5, 33-45.

JENNISON, C. & TURNBULL, B.W. (1985). Repeated confidence intervals for the median survival time. To appear in Biometrika 72.

JOHANSEN, S. (1978). The product limit estimator as maximum likelihood estimator. Scand. J. Statist. 5, 195-199.

KALBFLEISCH, J.D. & PRENTICE, R.L. (1980). The statistical analysis of failure time data. New York: Wiley.

KAPLAN, E.L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. J. Am. Statist. Assoc. 53, 457-481.

LAN, K.K.G. & DEMETS, D.L. (1983). Discrete sequential boundaries for clinical trials. Biometrika 70, 659-663.

MEIER, P. (1975). Estimation of a distribution function from incomplete observations. In Perspectives in Probability and Statistics. Papers in Honour of M.S. Bartlett. (Ed. J. Gani), pp. 67-87. New York: Academic Press.

NELSON, W. (1972). Theory and applications of hazard plotting for censored failure data. Technometrics 14, 945-966.

O'BRIEN, P.C. & FLEMING, T.R. (1979). A multiple testing procedure for clinical trials. Biometrics 35, 549-556.

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.

SCHOLTZ, F.W. (1980). Towards a unified definition of maximum likelihood. Canad. J. of Statist. 8, 193-203.

SIMON, R. & LEE, Y.J. (1982). Nonparametric confidence limits for survival probabilities and median survival time. Cancer Treatment Reports 66, 37-42.

SLUD, E.V., BYAR, D.P. & GREEN, S.B. (1984). A comparison of reflected versus test-based confidence intervals for the median survial time based on censored data. Biometrics 40, 587-600.

THOMAS, D.R. & GRUNKEMEIER, G.L. (1975). Confidence interval estimation of survival probabilities for censored data. J. Am. Statist. Assoc. 70, 865-871.

Table 2.1(a). Empirical error rates[1] for (1-2α) level confidence intervals. Sample size 21, α = .005

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	.007,.004	.020,.014	.000,.004	.000,.014
	Unif(0,2)	43	.006,.005	.018,.013	.000,.004	.000,.014
	Exp(.3)	23	.003,.006	.014,.015	.001,.004	.000,.015
	Unif(0,4.5)	22	.005,.004	.016,.010	.002,.003	.000,.010
Weib(1,.7)	Exp(1)	48	.006,.004	.018,.012	.000,.005	.000,.015
	Unif(0,2)	42	.005,.005	.015,.010	.000,.004	.000,.011
	Exp(.3)	25	.004,.006	.015,.013	.001,.003	.000,.012
	Unif(0,4.5)	25	.006,.005	.016,.013	.001,.003	.000,.011
Weib(1,1.5)	Exp(1)	53	.007,.008	.024,.022	.000,.008	.000,.019
	Unif(0,2)	44	.006,.005	.016,.014	.000,.005	.000,.015
	Exp(.3)	22	.004,.006	.013,.012	.002,.004	.000,.011
	Unif(0,4.5)	20	.005,.004	.013,.011	.002,.003	.000,.011

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 6000 simulations. Standard error for each entry is approximately .001.

Table 2.1(b). Empirical error rates[1] for (1-2α) level confidence intervals. Sample size 21, α = .025

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	.027,.024	.047,.038	.001,.023	.000,.034
	Unif(0,2)	43	.029,.025	.043,.038	.006,.024	.000,.033
	Exp(.3)	23	.024,.028	.040,.040	.013,.024	.004,.035
	Unif(0,4.5)	22	.024,.019	.040,.032	.014,.018	.005,.029
Weib(1,.7)	Exp(1)	48	.026,.023	.045,.034	.002,.022	.000,.032
	Unif(0,2)	42	.025,.021	.040,.033	.003,.019	.000,.029
	Exp(.3)	25	.023,.023	.042,.037	.012,.022	.003,.031
	Unif(0,4.5)	25	.024,.023	.046,.034	.012,.021	.005,.031
Weib(1,1.5)	Exp(1)	53	.030,.029	.051,.043	.002,.024	.000,.034
	Unif(0,2)	44	.024,.024	.041,.036	.004,.021	.001,.034
	Exp(.3)	22	.024,.023	.039,.034	.013,.019	.004,.029
	Unif(0,4.5)	20	.022,.023	.036,.031	.012,.019	.004,.029

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 6000 simulations. Standard error for each entry is approximately .002.

Table 2.1(c). Empirical error rates[1] for (1-2α) level confidence intervals. Sample size 21, α = .05

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	.051,.051	.071,.063	.006,.046	.003,.051
	Unif(0,2)	43	.053,.052	.070,.061	.013,.048	.007,.054
	Exp(.3)	23	.052,.053	.064,.062	.032,.047	.016,.054
	Unif(0,4.5)	22	.050,.049	.060,.057	.036,.041	.022,.048
Weib(1,.7)	Exp(1)	48	.052,.051	.066,.059	.006,.045	.003,.049
	Unif(0,2)	42	.049,.049	.062,.059	.010,.044	.005,.049
	Exp(.3)	25	.054,.051	.062,.060	.031,.045	.016,.050
	Unif(0,4.5)	25	.059,.050	.064,.057	.040,.038	.022,.044
Weib(1,1.5)	Exp(1)	53	.056,.054	.078,.066	.011,.045	.002,.049
	Unif(0,2)	44	.049,.051	.067,.062	.016,.047	.007,.049
	Exp(.3)	22	.049,.048	.062,.062	.032,.043	.020,.052
	Unif(0,4.5)	20	.049,.048	.062,.061	.032,.040	.020,.050

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 6000 simulations. Standard error for each entry is approximately .003.

Table 2.2(a). Empirical error rates[1] for (1-2α) level confidence intervals. Sample size 41, α = .005

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	.003,.005	.010,.009	.000,.006	.000,.015
	Unif(0,2)	43	.004,.006	.010,.010	.001,.007	.000,.014
	Exp(.3)	23	.004,.004	.010,.007	.002,.004	.000,.009
	Unif(0,4.5)	22	.004,.006	.009,.010	.002,.006	.001,.012
Weib(1,.7)	Exp(1)	48	.004,.004	.010,.008	.000,.006	.000,.013
	Unif(0,2)	42	.004,.004	.010,.007	.000,.004	.000,.010
	Exp(.3)	25	.004,.006	.008,.007	.002,.005	.000,.010
	Unif(0,4.5)	25	.003,.006	.006,.008	.002,.005	.000,.011
Weib(1,1.5)	Exp(1)	53	.004,.003	.011,.008	.000,.005	.000,.012
	Unif(0,2)	44	.006,.007	.014,.010	.001,.006	.000,.014
	Exp(.3)	22	.005,.005	.009,.010	.002,.005	.000,.012
	Unif(0,4.5)	20	.004,.006	.008,.010	.003,.006	.001,.013

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 6000 simulations. Standard error for each entry is approximately .001.

Table 2.2(b). Empirical error rates[1] for (1-2α) level confidence intervals. Sample size 41, α = .025

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	.024,.028	.037,.033	.003,.030	.000,.036
	Unif(0,2)	43	.023,.026	.032,.030	.008,.027	.003,.034
	Exp(.3)	23	.027,.023	.033,.029	.019,.021	.012,.029
	Unif(0,4.5)	22	.026,.027	.034,.033	.019,.026	.012,.034
Weib(1,.7)	Exp(1)	48	.025,.026	.036,.030	.004,.027	.002,.037
	Unif(0,2)	42	.028,.022	.034,.028	.009,.023	.004,.031
	Exp(.3)	25	.023,.025	.031,.032	.014,.025	.009,.032
	Unif(0,4.5)	25	.020,.024	.028,.031	.015,.024	.008,.030
Weib(1,1.5)	Exp(1)	53	.023,.024	.038,.032	.005,.024	.002,.031
	Unif(0,2)	44	.030,.027	.040,.033	.011,.029	.006,.036
	Exp(.3)	22	.022,.023	.028,.029	.016,.022	.010,.030
	Unif(0,4.5)	20	.026,.028	.034,.034	.019,.026	.011,.034

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 6000 simulations. Standard error for each entry is approximately .002.

Table 2.2(c). Empirical error rates[1] for (1-2α) level confidence intervals. Sample size 41, α = .05

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	.050,.053	.064,.056	.015,.054	.008,.058
	Unif(0,2)	43	.048,.051	.057,.057	.025,.051	.021,.055
	Exp(.3)	23	.052,.046	.059,.053	.043,.044	.034,.047
	Unif(0,4.5)	22	.053,.054	.061,.060	.045,.053	.038,.054
Weib(1,.7)	Exp(1)	48	.052,.052	.063,.055	.016,.053	.009,.058
	Unif(0,2)	42	.051,.049	.060,.055	.025,.050	.024,.054
	Exp(.3)	25	.047,.048	.053,.052	.038,.047	.031,.051
	Unif(0,4.5)	25	.049,.051	.056,.059	.035,.051	.032,.053
Weib(1,1.5)	Exp(1)	53	.051,.050	.066,.056	.018,.048	.010,.051
	Unif(0,2)	44	.054,.053	.065,.060	.032,.051	.025,.055
	Exp(.3)	22	.044,.047	.051,.052	.036,.044	.029,.045
	Unif(0,4.5)	20	.054,.055	.060,.061	.045,.053	.037,.053

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 6000 simulations. Standard error for each entry is approximately .003.

Table 2.3(a). Empirical average observed lengths of 99% confidence intervals (α = .005) for sample sizes N = 21 and 41, respectively.[1]

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	1.47,1.30	1.30,1.21	1.61,1.29	1.55,1.26
	Unif(0,2)	43	1.28,1.06	1.15,0.99	1.37,1.06	1.38,1.04
	Exp(.3)	23	1.57,0.93	1.30,0.85	1.65,0.95	1.50,0.91
	Unif(0,4.5)	22	1.49,0.95	1.23,0.87	1.56,0.96	1.44,0.92
Weib(1,.7)	Exp(1)	48	1.79,1.68	1.64,1.62	1.85,1.61	1.78,1.52
	Unif(0,2)	42	1.44,1.26	1.33,1.21	1.49,1.25	1.47,1.21
	Exp(.3)	25	2.25,1.24	1.85,1.13	2.29,1.24	1.93,1.15
	Unif(0,4.5)	25	2.02,1.24	1.69,1.14	2.08,1.23	1.81,1.14
Weib(1,1.5)	Exp(1)	53	1.16,0.98	1.01,0.87	1.33,1.02	1.33,1.02
	Unif(0,2)	44	1.08,0.82	0.92,0.75	1.18,0.84	1.23,0.84
	Exp(.3)	22	1.06,0.71	0.88,0.64	1.13,0.73	1.14,0.72
	Unif(0,4.5)	20	1.01,0.68	0.84,0.63	1.09,0.70	1.09,0.70

[1]Results based on 1000 simulations. Standard errors are, approximately, .02 and .01 for sample sizes 21 and 41, respectively.

Table 2.3(b). Empirical average observed lengths of 95% confidence intervals (α = .025) for sample sizes N = 21 and 41, respectively.[1]

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	1.20,0.95	1.09,0.93	1.34,0.98	1.36,0.98
	Unif(0,2)	43	1.05,0.78	0.97,0.75	1.14,0.80	1.15,0.79
	Exp(.3)	23	1.13,0.68	0.99,0.65	1.17,0.70	1.13,0.69
	Unif(0,4.5)	22	1.07,0.69	0.95,0.66	1.12,0.71	1.08,0.70
Weib(1,.7)	Exp(1)	48	1.50,1.22	1.38,1.21	1.58,1.22	1.58,1.22
	Unif(0,2)	42	1.21,0.98	1.12,0.95	1.28,0.98	1.27,0.97
	Exp(.3)	25	1.52,0.88	1.32,0.84	1.56,0.89	1.46,0.88
	Unif(0,4.5)	25	1.44,0.88	1.28,0.83	1.46,0.88	1.40,0.87
Weib(1,1.5)	Exp(1)	53	0.94,0.69	0.84,0.66	1.10,0.75	1.15,0.76
	Unif(0,2)	44	0.83,0.59	0.75,0.57	0.92,0.62	0.96,0.62
	Exp(.3)	22	0.76,0.52	0.69,0.49	0.83,0.54	0.82,0.54
	Unif(0,4.5)	20	0.74,0.50	0.66,0.48	0.78,0.52	0.78,0.52

[1]Results based on 1000 simulations. Standard errors are, approximately, .02 and .01 for sample sizes 21 and 41, respectively.

Table 2.3(c). Empirical average observed lengths of 90% confidence intervals ($\alpha = .05$) for sample sizes N = 21 and 41, respectively.[1]

Survival Distribution	Censoring Distribution	Expected % Censored	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
Exp(1)	Exp(1)	50	1.02,0.77	0.95,0.76	1.17,0.81	1.20,0.83
	Unif(0,2)	43	0.89,0.64	0.84,0.63	0.98,0.67	1.01,0.67
	Exp(.3)	23	0.88,0.56	0.83,0.54	0.94,0.58	0.95,0.58
	Unif(0,4.5)	22	0.86,0.57	0.81,0.55	0.90,0.59	0.90,0.59
Weib(1,.7)	Exp(1)	48	1.28,0.97	1.22,0.97	1.39,1.00	1.43,1.01
	Unif(0,2)	42	1.04,0.80	1.01,0.78	1.13,0.82	1.15,0.82
	Exp(.3)	25	1.18,0.72	1.12,0.70	1.25,0.74	1.25,0.73
	Unif(0,4.5)	25	1.12,0.72	1.07,0.69	1.20,0.73	1.19,0.73
Weib(1,1.5)	Exp(1)	53	0.79,0.56	0.73,0.54	0.94,0.61	0.98,0.63
	Unif(0,2)	44	0.69,0.48	0.64,0.47	0.78,0.51	0.81,0.52
	Exp(.3)	22	0.62,0.43	0.57,0.42	0.66,0.45	0.67,0.45
	Unif(0,4.5)	20	0.60,0.42	0.56,0.41	0.63,0.44	0.64,0.44

[1]Results based on 1000 simulations. Standard errors are, approximately, .02 and .01 for sample sizes 21 and 41, respectively.

Table 4.1. Parameters $Z_P(N,\alpha)$ and $Z_B(N,\alpha)$ for multiple testing procedures of Pocock and O'Brien/Fleming respectively

	$Z_P(N,\alpha)$			$Z_B(N,\alpha)$		
N \ α	0.005	0.025	0.05	0.005	0.025	0.05
2	2.772	2.178	1.876	3.626	2.803	2.391
3	2.873	2.289	1.993	4.417	3.438	2.953
5	2.986	2.413	2.122	5.847	4.555	3.929
10	3.117	2.555	2.270	8.328	6.689	5.732

Table 4.2(a). Empirical error rates[1] for (1-2α) level repeated confidence intervals based on Constrained Variance method

Survival Distribution Competing Risk Censoring		Exp(1) None	Weib(1,.7) None	Weib(1,1.5) None	Exp(1) Exp(.3)
Pocock					
α = .005	K = 5	.004,.003	.005,.004	.004,.003	.003,.005
	K = 10	.004,.004	.004,.003	.005,.006	.004,.003
α = .025	K = 5	.019,.020	.021,.019	.024,.020	.021,.021
	K = 10	.024,.025	.020,.019	.024,.026	.025,.023
α = .050	K = 5	.040,.043	.044,.043	.042,.043	.047,.044
	K = 10	.049,.046	.046,.036	.047,.050	.047,.045
O'Brien/Fleming					
α = .005	K = 5	.005,.003	.004,.003	.006,.004	.005,.004
	K = 10	.005,.009	.005,.005	.006,.007	.006,.005
α = .025	K = 5	.021,.022	.022,.022	.024,.020	.024,.024
	K = 10	.025,.026	.021,.019	.024,.028	.023,.021
α = .050	K = 5	.045,.042	.047,.042	.049,.048	.051,.048
	K = 10	.054,.052	.045,.043	.052,.051	.046,.044

[1]Entries show overall lower and upper tail error probabilities, respectively, estimated from 5000 simulations. Intervals are calculated at times $t_k = 1 + (.5)k$ for $1 \leq k \leq K$ and the accrual rate is 20 subjects per unit time. Standard errors are approximately .001, .002, .003 for α = .005, .025, .050, respectively.

Table 4.2(b). Empirical error rates[1] for (1-2α) level repeated confidence intervals based on Brookmeyer-Crowley method

Survival Distribution Competing Risk Censoring		Exp(1) None	Weib(1,.7) None	Weib(1,1.5) None	Exp(1) Exp(.3)
Pocock					
α = .005	K = 5	.012,.010	.013,.009	.016,.010	.014,.011
	K = 10	.013,.012	.011,.009	.014,.015	.012,.011
α = .025	K = 5	.034,.030	.034,.029	.040,.031	.042,.029
	K = 10	.039,.036	.036,.028	.041,.038	.043,.033
α = .050	K = 5	.059,.056	.057,.054	.062,.054	.070,.056
	K = 10	.065,.060	.062,.053	.067,.066	.068,.058
O'Brien/Fleming					
α = .005	K = 5	.009,.007	.007,.006	.010,.006	.009,.007
	K = 10	.008,.010	.006,.006	.008,.011	.009,.009
α = .025	K = 5	.028,.026	.029,.027	.031,.026	.032,.029
	K = 10	.031,.031	.024,.024	.029,.034	.028,.026
α = .050	K = 5	.056,.050	.055,.049	.058,.054	.061,.053
	K = 10	.058,.057	.051,.048	.057,.058	.054,.050

[1]Entries show overall lower and upper tail error probabilities, respectively, estimated from 5000 simulations. Intervals are calculated at times $t_k = 1 + (.5)k$ for $1 \leq k \leq K$ and the accrual rate is 20 subjects per unit time. Standard errors are approximately .001, .002, .003 for α = .005, .025, .050, respectively.

Table 4.2(c). Empirical error rates[1] for (1-2α) level repeated confidence intervals based on Simple Reflected method

Survival Distribution Competing Risk Censoring		Exp(1) None	Weib(1,.7) None	Weib(1,1.5) None	Exp(1) Exp(.3)
Pocock					
α = .005	K = 5	.001,.004	.001,.003	.002,.004	.000,.006
	K = 10	.002,.005	.001,.003	.002,.007	.001,.006
α = .025	K = 5	.008,.022	.007,.021	.011,.019	.004,.024
	K = 10	.011,.026	.010,.020	.012,.027	.007,.025
α = .050	K = 5	.019,.045	.021,.044	.024,.042	.014,.046
	K = 10	.028,.049	.024,.038	.029,.051	.021,.048
O'Brien/Fleming					
α = .005	K = 5	.002,.004	.002,.004	.003,.003	.001,.005
	K = 10	.004,.009	.003,.005	.005,.008	.004,.006
α = .025	K = 5	.015,.022	.015,.022	.018,.020	.013,.026
	K = 10	.022,.027	.017,.020	.023,.028	.016,.022
α = .050	K = 5	.035,.042	.034,.043	.040,.047	.028,.047
	K = 10	.048,.053	.039,.044	.047,.051	.035,.047

[1]Entries show overall lower and upper tail error probabilities, respectively, estimated from 5000 simulations. Intervals are calculated at times $t_k = 1 + (.5)k$ for $1 \leq k \leq K$ and the accrual rate is 20 subjects per unit time. Standard errors are approximately .001, .002, .003 for α = .005, .025, .050, respectively.

Table 4.2(d). Empirical error rates[1] for (1-2α) level repeated confidence intervals based on Transformed Reflected Method

Survival Distribution Competing Risk Censoring		Exp(1) None	Weib(1,.7) None	Weib(1,1.5) None	Exp(1) Exp(.3)
Pocock					
α = .005	K = 5	.000,.016	.000,.015	.000,.016	.000,.020
	K = 10	.000,.021	.000,.016	.000,.024	.000,.021
α = .025	K = 5	.003,.041	.003,.041	.003,.041	.001,.043
	K = 10	.016,.048	.005,.039	.005,.051	.002,.047
α = .050	K = 5	.010,.063	.011,.063	.012,.060	.006,.063
	K = 10	.015,.074	.013,.063	.014,.076	.010,.073
O'Brien/Fleming					
α = .005	K = 5	.000,.011	.001,.010	.001,.009	.000,.014
	K = 10	.002,.014	.001,.009	.002,.014	.001,.012
α = .025	K = 5	.010,.033	.009,.033	.010,.035	.007,.039
	K = 10	.016,.037	.012,.031	.015,.038	.010,.034
α = .050	K = 5	.029,.055	.028,.055	.030,.058	.023,.058
	K = 10	.038,.064	.031,.055	.039,.064	.030,.056

[1]Entries show overall lower and upper tail error probabilities, respectively, estimated from 5000 simulations. Intervals are calculated at times $t_k = 1 + (.5)k$ for $1 \leq k \leq K$ and the accrual rate is 20 subjects per unit time. Standard errors are approximately .001, .002, .003 for α = .005, .025, .050, respectively.

Table 4.3(a). Empirical error rates[1], at each analysis, of $(1-2\alpha)$-repeated confidence intervals based on Pocock multiple testing procedure with $\alpha = .05$

Look k	Nominal error rate α'_k	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
5-Look Procedure					
1	.0169	.0148,.0170	.0270,.0252	.0008,.0180	.0000,.0262
2	.0169	.0150,.0166	.0210,.0228	.0030,.0164	.0004,.0252
3	.0169	.0134,.0168	.0210,.0204	.0066,.0184	.0022,.0232
4	.0169	.0136,.0164	.0180,.0188	.0092,.0166	.0060,.0212
5	.0169	.0158,.0156	.0194,.0182	.0114,.0158	.0072,.0222
10-Look Procedure					
1	.0116	.0118,.0098	.0222,.0178	.0004,.0114	.0000,.0212
2	.0116	.0132,.0112	.0200,.0160	.0018,.0110	.0000,.0192
3	.0116	.0100,.0098	.0152,.0128	.0034,.0094	.0012,.0172
4	.0116	.0110,.0082	.0150,.0120	.0060,.0086	.0020,.0140
5	.0116	.0114,.0098	.0148,.0128	.0074,.0098	.0028,.0156
6	.0116	.0122,.0108	.0160,.0138	.0092,.0112	.0044,.0176
7	.0116	.0104,.0096	.0128,.0116	.0082,.0100	.0044,.0138
8	.0116	.0110,.0082	.0130,.0120	.0096,.0086	.0062,.0140
9	.0116	.0098,.0102	.0130,.0120	.0084,.0110	.0046,.0152
10	.0116	.0108,.0096	.0118,.0112	.0090,.0104	.0054,.0148

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 5000 simulations. Intervals are calculated at times $t_k = 1 + (.5)k$ for $1 \leq k \leq K$ and the accrual rate is 20 subjects per unit time. Survival distribution is Exp(1) with no competing risk censoring. Standard errors are less than .002.

Table 4.3(b). Empirical error rates[1], at each analysis, of $(1-2\alpha)$-repeated confidence intervals based on O'Brien and Fleming multiple testing procedure with $\alpha = .05$

Look k	Nominal error rate α'_k	Constrained Variance	Brookmeyer-Crowley	Simple Reflected	Transformed Reflected
5-Look Procedure					
1	.00004	.0000,.0000	.0012,.0014	.0000,.0000	.0000,.0022
2	.0027	.0020,.0028	.0052,.0052	.0002,.0026	.0000,.0078
3	.0117	.0084,.0122	.0132,.0158	.0026,.0122	.0012,.0198
4	.0248	.0208,.0238	.0272,.0272	.0138,.0228	.0100,.0290
5	.0395	.0384,.0352	.0440,.0400	.0314,.0354	.0268,.0414
10-Look Procedure					
1	.00001	.0000,.0000	.0000,.0000	.0000,.0000	.0000,.0000
2	.00003	.0000,.0000	.0004,.0000	.0000,.0000	.0000,.0004
3	.0005	.0008,.0000	.0014,.0002	.0000,.0000	.0000,.0030
4	.0021	.0018,.0010	.0030,.0024	.0010,.0012	.0002,.0044
5	.0052	.0034,.0038	.0064,.0068	.0020,.0048	.0008,.0094
6	.0096	.0098,.0090	.0126,.0110	.0070,.0094	.0032,.0148
7	.0152	.0134,.0118	.0178,.0142	.0112,.0124	.0064,.0176
8	.0213	.0220,.0174	.0240,.0208	.0178,.0182	.0130,.0248
9	.0280	.0276,.0260	.0284,.0282	.0230,.0266	.0186,.0316
10	.0349	.0298,.0314	.0324,.0358	.0274,.0328	.0236,.0370

[1]Entries show lower and upper tail error probabilities, respectively, estimated from 5000 simulations. Intervals are calculated at times $t_k = 1 + (.5)k$ for $1 \leq k \leq K$ and the accrual rate is 20 subjects per unit time. Survival distribution is Exp(1) with no competing risk censoring. Standard errors are less than .002.

Table 5. Constrained variance 90% repeated confidence intervals for median survival time.

Data from Kalbfleisch and Prentice (1980, Data Set II, pages 225-229).

Time of analysis (days)	Patients entered	Patients died	Patients censored	Pocock RCI	O'Brien/ Fleming RCI	Fixed[1] Sample CI's
Procedure with 5 analyses						
720	83	27	56	(279,631)	(245,631)	(324,631)
1080	126	58	68	(338,631)	(324,661)	(374,560)
1440	174	91	83	(376,631)	(374,631)	(404,560)
1800	195	129	66	(370,548)	(374,546)	(376,544)
2160	195	142	53	(369,545)	(374,541)	(376,532)
Procedure with 10 analyses						
540	61	18	43	(235,∞)	(144,∞)	(245,∞)
720	83	27	56	(279,631)	(243,631)	(324,631)
900	104	43	61	(301,631)	(273,∞)	(327,575)
1080	126	58	68	(328,631)	(324,661)	(374,560)
1260	148	77	71	(338,575)	(327,631)	(374,548)
1440	174	91	83	(374,631)	(374,631)	(404,560)
1620	188	105	83	(382,637)	(404,637)	(432,599)
1800	195	129	66	(369,548)	(372,546)	(376,544)
1980	195	137	58	(363,546)	(370,544)	(374,532)
2160	195	142	53	(363,546)	(372,541)	(376,532)

[1]Not adjusted for multiple looks.