Statistica Sinica **8**(1998), 411-428

REPEATED SIGNIFICANCE TESTING WITH CENSORED RANK STATISTICS IN INTERIM ANALYSIS OF CLINICAL TRIALS

Minggao Gu and Tze Leung Lai

McGill University and Stanford University

Abstract: A simple class of stopping rules is introduced for time-sequential rank tests to compare time to failure between two treatment groups, such as in the case of a clinical trial in which patients enter serially and in which interim analyses of the data are performed periodically so that the trial may be stopped early when one treatment is found to be significantly better than the other. These time-sequential rank tests are shown to achieve both savings in study duration and increase in power over their nonsequential counterparts, and provide a simple but statistically efficient method to circumvent the difficulty of "calendar time" versus "information time" in the design of group sequential trials with failure-time endpoints.

Key words and phrases: Censored survival data, clinical trials, group sequential tests, rank statistics, stopping rules, use function.

1. Introduction

In many chinical trials a primary objective is to compare time to failure between two treatment groups X and Y. Suppose that the failure times $X_1,\ldots,X_{n'}$ are independent having a common distribution function F and the failure times $Y_1,\ldots,Y_{n''}$ are independent having a common distribution function G. Let $n = n' + n''$. To test the null hypothesis $H_0 : F = G$ or $H_0' : F \le G$, a commonly used method is to evaluate the ranks R_i of $X_i (i = 1,\ldots,n')$ in the combined sample $X_1,\ldots,X_{n'},Y_1,\ldots,Y_{n''}$ and to use rank statistics of the form $\ell_n = \sum_{i=1}^{n'} \varphi(R_i/n)$, where $\varphi : (0,1] \to (-\infty,\infty)$. However, because of withdrawal from the study and the need to terminate the trial by some scheduled date, the X_i and Y_j may be censored and one cannot compute ℓ_n in these situations. As noted in Gu, Lai and Lan (1991), a natural extension of ℓ_n to censored data is the censored rank statistic of the form

$$S_n = \sum_{k=1}^{K} \psi(H_n(Z_{(k)}))(z_k - m_k/\#_k), \tag{1.1}$$

where $Z_{(1)} \le \cdots \le Z_{(K)}$ denote the ordered uncensored observations in the combined sample, $z_k = 1$ if $Z_{(k)}$ is an X and $z_k = 0$ if $Z_{(k)}$ is a Y, $\#_k$ (resp. m_k)

denotes the number of observations (resp. X's) in the combined sample that are $\geq Z_{(k)}, H_n$ is the Kaplan-Meier curve (or an asymptotically equivalent variant thereof) based on the combined sample, and ψ is related to φ by the relation

$$\psi(u) = \varphi(u) - (1-u)^{-1} \int_u^1 \varphi(t)dt, \quad 0 < u < 1. \tag{1.2}$$

Taking $\psi(u) = (1-u)^\rho$ $(\rho \geq 0)$ yields the G^ρ statistics proposed by Harrington and Fleming (1982). The case $\rho = 0$ corresponds to Mantel's (1966) logrank statistic and the case $\rho = 1$ corresponds to the generalization of Wilcoxon's statistic by Peto and Peto (1972) and Prentice (1978).

In typical clinical trials, patients enter the study serially and are then followed until they fail or withdraw from the study, or until the study is terminated. The trial is typically scheduled to end by a certain time t^* and there are also periodic reviews of the data prior to t^*. If significant differences between the treatment groups are found from an interim analysis, a decision might be made to terminate the trial before t^*. Since the response of interest is time to failure and since patients usually do not enter the study at the same time, this means that there are two time scales to be considered, namely, calendar time t as measured from the time the study starts and age time s as measured for each patient from the time he enters the study. Hence the censored rank statistic (1.1) now depends on the calendar time t at which it is evaluated and we shall denote it by $S_n(t)$, whose precise definition will be given in Section 2.

Assuming a Lehmann (proportional hazards) family of the form $1 - G(s) = (1 - F(s))^{1-\theta}$, Jones and Whitehead (1979) considered the use of time-sequential logrank statistics $S_n(t)$ to test sequentially over time the one-sided null hypothesis $H_0' : \theta \leq 0$. They suggested plotting $S_n(t)$ versus $V_n(t)$, where $V_n(t)$ is Mantel's (1966) estimate of the variance of $S_n(t)$ under $F = G$. They argued heuristically that $\{(V_n(t), S_n(t)), t \geq 0\}$ should behave approximately like $\{(v, W(v)), v \geq 0\}$, where $W(v)$ is the standard Wiener process under $\theta = 0$ and is a Wiener process with drift coefficient depending on θ under alternatives θ near 0. Using this Wiener process approximation, they suggested replacing $(v, W(v))$ in a sequential test for the sign of the drift of a Wiener process by $(V_n(t), S_n(t))$ to construct a corresponding sequential logrank test of H_0', and considered in particular the case where the sequential test based on $(v, W(v))$ is a sequential probability ratio test (SPRT). Assuming i.i.d. arrival times and withdrawal times, Tsiatis (1981) established the asymptotic normality of $(S_n(t_1)/\sqrt{n}, \ldots, S_n(t_k)/\sqrt{n})$ under $F = G$ for any $k \geq 1$ and $t_1, \ldots, t_k$, and Sellke and Siegmund (1983) established weak convergence of $\{S_n(t)/\sqrt{n}, t \geq 0\}$ to a zero-mean Gaussian process with independent increments under $F = G$ and general arrival and withdrawal patterns, thus providing a rigorous asymptotic justification of the heuristics of Jones and

Whitehead (1979) under $H_0 : \theta = 0$. Gu and Lai (1991) later showed that $\{(V_n(t)/n, S_n(t)/\sqrt{n}), t \geq 0\}$ converges weakly to $\{(v, W(v)), v \geq 0\}$ under contiguous proportional hazards alternatives, where $W(v)$ is a Wiener process with $EW(v)/v = c$, thus giving a rigorous asymptotic justification of the heuristics of Jones and Whitehead under $H_1 : \theta = c/\sqrt{n}$.

Since interim analyses of the data are performed only at a few calendar times instead of continuously as in the SPRT of the drift of a continuous-time Wiener process, direct use of the type I error probability of the Wiener process SPRT as an approximation to that of the corresponding time-sequential logrank test performed at periodic reviews of the data is overly conservative. There is an extensive literature, commonly referred to as "group sequential methods", that addresses the adjustments needed when the data are analyzed in successive "groups" (and therefore not in a fully sequential manner as in classical sequential analysis). As pointed out above, there are two time scales in the present context of sequential monitoring of censored survival data with staggered entry, namely, calendar time t and age time s. This introduces considerable difficulties in the development of group sequential tesets for such data (cf. Lan and DeMets (1983, 1989)). In Section 2 we give a brief review of the literature and provide a comprehensive methodology for determining stopping boundaries, with good statistical properties, of group sequential censored rank tests.

This methodology is applicable not only to the logrank statistics but also to general score functions ψ in (1.1). It will be shown in Section 3 that the methodology is also applicable to more complex situations where adjustments for prognostic factors have to be made for meaningful comparison of the two treatments. Section 4 presents some simulation results on the performance of these time-sequential censored rank tests, showing that they can achieve both savings in study duration and increase in power over nonsequential censored rank tests (without interim analysis). Making use of the asymptotic theory of time-sequential censored rank statistics developed in Gu and Lai (1991), a theoretical explanation of the savings found in these simulation studies is also given.

2. A Class of Repeated Significance Tests with Censored Rank Statistics

Suppose a clinical trial involves $n = n' + n''$ patients with n' of them assigned to treatment X and n'' assigned to treatment Y. Let $T_i' \geq 0$ denote the entry time and $X_i > 0$ the survival time (or time to failure) after entry of the ith subject in treatment group X and let T_j'' and Y_j denote the entry time and survival time after entry of the jth subject in treatment group Y. The subjects are followed until they fail or withdraw from the study or until the study is terminated. Let $\xi_i'(\xi_j'')$ denote the time to withdrawal, possibly infinite, of the ith (jth) subject in the treatment group $X(Y)$. Thus the data at calendar time t consist of $(X_i(t), \delta_i'(t)), i = 1, \ldots, n'$, and $(Y_j(t), \delta_j''(t)),\ j = 1, \ldots, n''$, where

$$\begin{aligned} &X_i(t) = \min(X_i, \xi_i', (t - T_i')^+),\ Y_j(t) = \min(Y_j, \xi_j'', (t - T_j'')^+),\\ &\delta_i'(t) = I(X_i(t) = X_i),\ \delta_j''(t) = I(Y_j(t) = Y_j), \end{aligned} \tag{2.1}$$

where a^+ is the positive part of number a. At a given calendar time, on the basis of the observed data (2.1) from the two treatment groups, one can compute the rank statistic (1.1) which can be expressed in the present notation as

$$\begin{aligned} S_n(t) = &\sum_{i=1}^{n'} \delta_i'(t)\psi(H_{n,t}(X_i(t))\Big\{1 - \frac{m_{n,t}'(X_i(t))}{m_{n,t}'(X_i(t)) + m_{n,t}''(X_i(t))}\Big\}\\ &- \sum_{j=1}^{n''} \delta_j''(t)\psi(H_{n,t}(Y_j(t)))\frac{m_{n,t}'(Y_j(t))}{m_{n,t}'(Y_j(t)) + m_{n,t}''(Y_j(t))}, \end{aligned} \tag{2.2}$$

where ψ is a nonrandom function on $[0, 1]$ and

$$m_{n,t}'(s) = \sum_{i=1}^{n'} I(X_i(t) \geq s), \quad m_{n,t}''(s) = \sum_{j=1}^{n''} I(Y_j(t) \geq s), \tag{2.3}$$

$$N_{n,t}'(s) = \sum_{i=1}^{n'} I(X_i \leq \xi_i' \wedge (t - T_i')^+ \wedge s),$$

$$N_{n,t}'' = \sum_{j=1}^{n} I(Y_j \leq \xi_j'' \wedge (t - T_j'')^+ \wedge s), \tag{2.4}$$

$$1 - H_{n,t}(s) = \prod_{u<s} \Big\{1 - \frac{\Delta N_{n,t}'(u) + \Delta N_{n,t}''(u)}{m_{n,t}'(u) + m_{n,t}''(u)}\Big\}, \tag{2.5}$$

where we use the convention $0/0 = 0$, $\Delta N_{n,t}'(s) = N_{n,t}'(s) - N_{n,t}'(s-)$, $N_{n,t}'(s-) = \lim_{u \to s-} N_{n,t}'(u)$, and use $\wedge$ to denote minimum.

2.1. Weak convergence of time-sequential censored rank statistics

Suppose that ψ is continuous and has bounded variation on $[0, 1]$ and that the limits

$$b'(t, s) = \lim_{m \to \infty} m^{-1} \sum_{i=1}^{m} P\{\xi_i' \geq s, t - T_i' \geq s\},$$

$$b''(t, s) = \lim_{m \to \infty} m^{-1} \sum_{j=1}^{m} P\{\xi_j'' \geq s, t - T_j'' \geq s\} \tag{2.6}$$

exist and are continuous in $0 \le s \le t$. Moreover, assume that

$$n'/n \to \gamma \text{ as } n(= n' + n'') \to \infty \text{ with } 0 < \gamma < 1. \tag{2.7}$$

Suppose that F and G are continuous and let $\Lambda_F = -\log(1-F)$ and $\Lambda_G = -\log(1-G)$ denote their cumulative hazard functions. Let

$$\mu_n(t) = \int_0^t \psi(H_{n,t}(s)) \frac{m'_{n,t}(s) m''_{n,t}(s)}{m'_{n,t}(s) + m''_{n,t}(s)} (d\Lambda_F(s) - d\Lambda_G(s)).$$

Note that $\mu_n(t) = 0$ if $F = G$. Then Theorem 2 and Example 2 of Gu and Lai (1991) give the following results for every $t^* > 0$:

(i) For fixed F and G, $\{n^{-1/2}(S_n(t) - \mu_n(t)), 0 \le t \le t^*\}$ converges weakly in $D[0, t^*]$ to a zero-mean Gaussian process and $n^{-1}\mu_n(t)$ converges in probability as $n \to \infty$.

(ii) Let $\{Z(t), 0 \le t \le t^*\}$ denote the zero-mean Gaussian process in (i) when $F = G$. This Gaussian process has independent increments and

$$\mathrm{Var}(Z(t)) = \gamma(1-\gamma) \int_0^t \frac{\psi^2(F(s)) b'(t,s) b''(t,s)}{\gamma b'(t,s) + (1-\gamma) b''(t,s)} dF(s). \tag{2.8}$$

(iii) For fixed F (and therefore Λ_F also), suppose that as $n \to \infty$, $G \to F$ such that $\int_0^{t^*} |d\Lambda_G/d\Lambda_F - 1| d\Lambda_F = O(n^{-1/2})$ and $\sqrt{n}\Big(d\Lambda_G/d\Lambda_F(s) - 1\Big) \to g(s)$ as $n \to \infty$, uniformly in $s \in I$ and $\sup_{s \in I} |g(s)| < \infty$ for all closed subintervals I of $\{s \in [0, t^*] : F(s) < 1\}$. Then $\{n^{-1/2} S_n(t), 0 \le t \le t^*\}$ converges weakly in $D[0, t^*]$ to $\{Z(t) + \mu(t), 0 \le t \le t^*\}$, where $Z(t)$ is the same Gaussian process as that in (ii) and

$$\mu(t) = -\gamma(1-\gamma) \int_0^t \frac{\psi(F(u)) g(u) b'(t,u) b''(t,u)}{\gamma b'(t,u) + (1-\gamma) b''(t,u)} dF(u). \tag{2.9}$$

From (ii) and (iii), the limiting Gaussian process of $\{n^{-1/2} S_n(t), t \ge 0\}$ has independent increments under $H_0 : F = G$ and under contiguous alternatives. Two commonly used estimates $V_n(t)$ of the variance of $S_n(t)$ under H_0 are

$$V_n(t) = \int_0^t \frac{\psi^2(H_{n,t}(s)) m'_{n,t}(s) m''_{n,t}(s)}{(m'_{n,t}(s) + m''_{n,t}(s))^2} d(N'_{n,t}(s) + N''_{n,t}(s)), \tag{2.10a}$$

or

$$V_n(t) = \int_0^t \frac{\psi^2(H_{n,t}(s))}{(m'_{n,t}(s) + m''_{n,t}(s))^2} \{(m''_{n,t}(s))^2 dN'_{n,t}(s) + (m'_{n,t}(s))^2 dN''_{n,t}(s)\}. \tag{2.10b}$$

As a compromise between these two choices, Gu and Lai (1991), page 1421, also considered

$$V_n(t) = \{(2.10\text{a}) + (2.10\text{b})\}/2. \tag{2.10c}$$

For all three estimates, $n^{-1}V_n(t)$ converges in probability to (2.8) under H_0 and under contiguous alternatives. Hence, letting $v = n^{-1}V_n(t)$ and $W(v) = n^{-1/2}S_n(t)$, we can regard $W(v),\ v \geq 0$, as the standard Wiener process under H_0. Moreover, if ψ is a scalar multiple of the asymptotically optimal score function, then we can also regard $W(v), v \geq 0$, as a Wiener process with some drift coefficient under contiguous alternatives.

2.2. Choice of stopping boundaries in time-sequential rank tests

Let $0 < t_1 < \cdots < t_k = t^*$ be prespecified times for periodic reviews of the data. To test $H_0 : F = G$ with the time-sequential rank statistics $S_n(t_i)$, Slud and Wei (1982) introduced the following simple approach. First choose positive numbers $\alpha_1, \ldots, \alpha_k$ such that $\sum_1^k \alpha_i = \alpha$ (= the overall significance level). Then use the multivariate normal approximation to the null distribution of $(S_n(t_i)/V_n^{1/2}(t_i))_{1 \leq i \leq k}$ to determine $d_1, \ldots, d_k$ recursively by

$$P_{H_0}\{|S_n(t_j)|/V_n^{1/2}(t_j) \geq d_j \text{ and } |S_n(t_i)|/V_n^{1/2}(t_i) < d_i \text{ for all } i < j\} = \alpha_j. \tag{2.11}$$

With the d_j thus determined, the Slud-Wei repeated significance test rejects H_0 whenever $|S_n(t_j)| \geq d_j V_n^{1/2}(t_j)$ $(1 \leq j \leq k)$ and stops the trial at the first t_j this occurs (or at t^* if this does not occur for $1 \leq j \leq k$).

The Slud-Wei method does not provide practical guidelines concerning how the α_j in (2.11) should be chosen. Lan and DeMets (1983) and Lan et al. (1984) proposed to derive the α_j from the so-called "use function", which specifies how fast we can spend the Type I error α over time. To begin with, let $\{W(v), 0 \leq v \leq 1\}$ be the standard Wiener process and consider the stopping rule $T = \inf\{v \in [0,1] : |W(v)| \geq h(v)\}(\inf \emptyset = \infty)$, where h is a positive function on $[0,1]$ such that $P\{T = 0\} = 0$ and $P\{T \leq 1\} = \alpha$. The use function is $A(v) = P\{T \leq v\}, 0 \leq v \leq 1$. Taking v to represent the proportion of information accumulated at time t, $A(v)$ can be interpreted as the amount of Type I error spent up to time t, with $A(0) = 0$ and $A(1) = \alpha$. In particular, suppose that instead of survival data one has immediate responses from the patients who enter the study serially and are randomized to either treatment, with a target sample size of n at the scheduled end of the trial. Lan and DeMets (1989) call such trials "maximum information trials". Here the proportion of information accumulated at time t_i of interim analysis is $v_i = n_i/n$, where n_i is the total number of patients available at t_i. Hence Lan and DeMets (1983) proposed to choose $\alpha_j = A(v_j) - A(v_{j-1})$

in (2.11). For the time-sequential rank statistics (2.2) in what Lan and DeMets (1989) call "maximum duration trials", the asymptotic null variance of $S_n(t_i)$ is no longer proportional to the sample size n_i at t_i and a natural analogue of n_i/n here is $V_n(t_i)/V_n(t^*)$.

Siegmund (1985), pages 129-131, proposed an alternative approach for logrank statistics, which can be readily extended to more general time-sequential rank statistics (2.2) as follows. Let $\{W(v), v \geq 0\}$ be a Wiener process with drift coefficient θ. Let $0 \leq v_0 < v_1$ and divide $[v_0, v_1]$ into $k-1$ equally spaced subintervals, with endpoints $v_0 = v^{(1)} < \cdots < v^{(k)} = v_1$. Letting $I = \{v^{(1)}, \ldots, v^{(k)}\}$, Siegmund considered Haybittle's (1971) repeated significance test of $H : \theta = 0$, with stopping rule $\tau = \min(v_1, \inf\{v \in I : |W(v)| \geq b\sqrt{v}\})$ and terminal decision rule that rejects H if $\tau < v_1$ and if $|W(v_1)| \geq c\sqrt{v_1}$ in the case $\tau = v_1$, where $0 < c \leq b$ are such that $P_H\{\text{Reject } H\} = \alpha$. Since $(n^{-1}V_n(t), n^{-1/2}S_n(t))$ can be approximated by $(v, W(v))$ under $H_0 : F = G$, he proposed a corresponding repeated significance test that rejects H_0 whenever

$$\begin{aligned}&|S_n(t_i)| \geq bV_n^{1/2}(t_i) \text{ and } v_0 \leq V_n(t_i) < v_1, \text{ for } 1 \leq i \leq k-1, \qquad (2.12)\\ &\text{or } |S_n(t_i)| \geq cV_n^{1/2}(t_i) \text{ and } V_n(t_i) \geq v_1, \text{ for } 1 \leq i \leq k-1, \text{ or } |S_n(t_k)| \geq cV_n^{1/2}(t_k).\end{aligned}$$

Thus, the test may be terminated at time t_N prior to t^*, where

$$N = \min\{i \leq k-1 : V_n(t_i) \geq v_1, \text{ or } v_0 \leq V_n(t_i) < v_1 \text{ and } |S_n(t_i)| \geq bV_n^{1/2}(t_i)\},$$

setting $N = k$ if the above set is empty. In the case of logrank statistics, the null variance of $S_n(t)$ is approximately 1/4 times the expected number of failures up to time t, and Siegmund suggested using this together with the prior information about accrual and failure rates that one uses in the design of the clinical trial to come up with 1/4 times the expected number of failures at t^* as the value of v_1 in (2.12). For the more general rank statistics (2.2), given $F(= G)$ and the distributions of (T_i', ξ_i') and (T_j'', ξ_j''), we can use Monte Carlo simulations to evaluate the null variance of $S_n(t^*)$. In particular, at the design stage we can use prior information about these quantities to find by simulation the value of v_1 in Siegmund's repeated significance test (2.12); see Gu and Lai (1995) for details.

The assumption of a set of evenly spaced "information times" $v^{(1)}, \ldots, v^{(k)}$ at which interim analyses of the data are performed in Siegmund's approach is usually not satisfied in practice. We can circumvent this difficulty by modifying his approach with a "use function" technique. To begin with, consider the continuous-time repeated significance test of $H : \theta = 0$, with stopping rule

$$\tau^* = \min(v_1, \inf\{v \geq v_0 : |W(v)| \geq b\sqrt{v}\}) \qquad (2.13)$$

and terminal decision rule that rejects H if

$$\text{either } \tau^* < v_1, \text{ or } \tau^* = v_1 \text{ and } |W(v_1)| \geq c\sqrt{v_1}, \tag{2.14}$$

where $0 < c \leq b$ are such that $P_H\{\text{Reject } H\} = \alpha$. The use function of this continuous-time Haybittle-type test is defined by

$$A(v_1) = \alpha,\ A(v) = P_H\{\tau^* \leq v\} \text{ for } v_0 \leq v < v_1, \text{ and } A(v) = 0 \text{ if } v < v_0. \tag{2.15}$$

An eigenfunction expansion of $A(v)$ has been given by DeLong (1981), who also tabulates $A(v)$ for a range of values of b and v/v_0. Letting ϕ and Φ denote the standard normal density and distribution functions, Theorem 4.2.1 of Siegmund (1985) gives the approximation

$$A(v) \doteq (b - b^{-1})\phi(b)\log(v/v_0) + 4b^{-1}\phi(b) \text{ for } v_0 < v < v_1. \tag{2.16}$$

Moreover, $A(v_0) = 2(1 - \Phi(b))$ and Eq. (4.18) of Siegmund gives the approximation

$$\begin{aligned} A(v_1) &= P_H\{|W(v_1)| \geq c\sqrt{v_1}\} + P\{|W(v_1)| < c\sqrt{v_1},\ \tau^* < v_1\} \\ &\doteq 2(1 - \Phi(c)) + b\phi(b)\log(v_1c^2/v_0b^2)I(v_1c^2 > v_0b^2). \end{aligned} \tag{2.17}$$

The approximations (2.16) and (2.17) are derived as asymptotic expansions as $b \to \infty$, $c \to \infty$ and $v_0 \to \infty$ such that $b/\sqrt{v_0} \to k_0$, $b/\sqrt{v} \to k(v)$, $b/\sqrt{v_1} \to k_1$ and $c/\sqrt{v_1} \to k_1'$, where k_0, k_1, k_1' and $k(v)$ are constants. Table 4.1 of Siegmund (1985) gives numerical results showing the adequacy of these approximations when $(v - v_0)/v_0$ in (2.16) and b are not too small.

In applying (2.15) as a use function to modify Siegmund's repeated significance test (2.12) for the rank statistics (2.2), v_1 represents an *a priori* estimate of the null variance of $S_n(t^*)$ and we take $v_0 = \epsilon v_1$ with $0 < \epsilon < 1$ to represent some minimal information in the data before interim analysis should be tried. Note that although (2.8) is nondecreasing in t, its estimate $V_n(t)$ may fail to be monotone. To get around this difficulty we redefine $V_n(t_i)$ to be $V_n(t_{i-1})$ if $V_n(t_i) < V(t_{i-1})$. For $1 \leq i \leq k-1$, define $\alpha_i = A(v_1 \wedge V_n(t_i)) - A(V_n(t_{i-1}))$ if $V_n(t_i) \geq v_0$, setting $t_0 = 0 = V_n(0)$, and define $\alpha_i = 0$ if $V_n(t_i) < v_0$. Moreover, define $\alpha_k = \alpha - A(V_n(t_{k-1}))$. Note that $\alpha_i = 0$ is equivalent to skipping interim testing at time t_i. With α_i thus chosen, define the stopping boundary $\{d_j : 1 \leq j \leq k\}$ by (2.11) in which $\alpha_j = 0$ corresponds to $d_j = \infty$.

2.3. A refinement of the Haybittle-Peto repeated significance test

Let $Z_1, Z_2, \ldots$ be i.i.d. normal random variables with unknown mean θ and known variance 1. To test $H : \theta = 0$ at level α, the Neyman-Pearson test rejects

H if $|\sum_1^k Z_i| \geq z_\alpha\sqrt{k}$, where $1-\Phi(z_\alpha)=\alpha$. Sample size calculations in clinical trial applications typically assume an alternative θ of particular interest and find the k that attains some given power $1-\beta$ at θ. The basic idea behind Haybittle's (1971) repeated significance test is to keep k and α as the maximum sample size and significance level but to allow for early stopping when the data are monitored sequentially, at the expense of some minor loss in power at θ. This leads to the stopping rule $\tau_b = \min(k, \inf\{n \geq 1 : |\sum_{i=1}^n Z_i| \geq b\sqrt{n}\})$ and terminal decision rule that rejects H if $\tau_b < k$ or if $\tau_b = k$ and $|\sum_{i=1}^k Z_i| \geq c\sqrt{k}$. Since we require the loss in power at θ to be small relative to the fixed sample size test and also require the maximum sample size to be the same as the fixed sample size k, it is clear that c has to be near z_α, implying that $P_0(\tau_b < k)$ is small in comparison with α. In particular, the Peto-type methods in the field of clinical trials use some relatively large value of b, such as 3, and conventional critical values of c for the final test when the number k of interim analyses is small. Although no precise Tpye I error is guaranteed in these methods, we can use a simple and flexible procedure described below to determine b and c to guarantee a prescribed Type I error.

The fact that $P_0(\tau_b < k)$ is typically small relative to α (or equivalently that most of the Type I error is to be spent at the terminal data t^*) suggests that using an elaborate Lan-DeMets boundary determination procedure would not lead to substantial improvement over the simple procedure that uses a fixed threshold b for $|S_n(t_i)|/V_n^{1/2}(t_i)$ with $t_i < t^*$. We therefore propose using the following simple repeated significance testing procedure involving a maximum of k significance tests. First determine b such that $P_0(\tau_b < k) = \epsilon\alpha$, where $0 < \epsilon < 1$ is small and τ_b is the stopping rule for the Z_i defined above. The approximation formula (4.40) for $P_0(\tau_b < k)$ and Table 4.2 in Siegmund (1985) are useful for this choice of b that ensures $P_0(\tau_b < k)$ to be a small fraction of α.

With b thus chosen, the proposed repeated significance test stops the trial and rejects H_0 at $t_i < t^*$ if $|S_n(t_i)| \geq bV^{1/2}(t_i)$. If the trial proceeds to the terminal date t^*, the test rejects H_0 if $|S_n(t^*)| \geq cV_n^{1/2}(t^*)$, where c is so chosen that

$$\begin{aligned} P\{|W(V_n(t_k))| \geq cV_n^{1/2}(t_k) \text{ or } |W(V_n(t_i))| \geq bV_n^{1/2}(t_i)& \\ \text{for some } i < k | V_n(t_1), \ldots, V_n(t_k)\} = \alpha,& \qquad (2.18) \end{aligned}$$

in which $t_k = t^*$ and $\{W(v), v \geq 0\}$ is a standard Wiener process independent of $\{(X_i, \xi_i', T_i', Y_i, \xi_i'', T_i''), i \geq 1\}$. Letting $a_j = V_n(t_j)$ and $d_k = c, d_j = b$ for $1 \leq j \leq k-1$, the probability (2.18) can be written as a sum of the probabilities $P\{|W(a_j)| \geq d_j\sqrt{a_j}$ and $|W(a_i)| < d_i\sqrt{a_i}$ for all $i < j\}$, which can be computed by the recursive numerical integration algorithm of Armitage, McPherson and Rowe (1969). This procedure is much more convenient than the use function approach at the end of Section 2.2 and has the important advantage of not requiring prior specification of v_0 and v_1, which are needed in Siegmund's procedure and its use-function modification. Note that the choice of c in (2.18) is not predetermined at the beginning of the trial but depends on the actual values of $V_n(t_1), \ldots, V_n(t_k)$, allowing great flexibility in how information accumulates at different times of interim analyses.

3. Adjustments for Concomitant Variables

It is widely recognized that tests of treatment effects based on the rank statistics (1.1) may lose substantial power when the effects of other covariates are strong. A commonly used method to remedy this when logrank statistics are used is to assume the proportional hazards regression model and to use Cox's partial likelihood approach to adjust for other covariates. Tsiatis, Rosner and Tritchler (1985) and Gu and Ying (1995) have proposed group sequential tests using proportional hazards regression to adjust for other covariates in testing whether there are treatment differences on survival.

Instead of the proportional hazards model, we assume the traditional regression model that $h(X_i) - \beta^T \mathbf{U}_i'$ are i.i.d. with distribution function F and that $h(Y_j) - \beta^T \mathbf{U}_j''$ are i.i.d. with distribution function G, where h is a known function, β is an unknown parameter and $\mathbf{U}_i', \mathbf{U}_j''$ represent the covariates. Assuming the null hypothesis $H_0 : F = G$, β can be estimated from the combined sample (2.1) by using rank estimators or M-estimators $\widehat{\beta}_{n,t}$ that are $\sqrt{n}$-consistent under (2.6), (2.7) and certain assumptions on $F(= G)$ and the covariates $\mathbf{U}_i', \mathbf{U}_j''$ (cf. Lai and Ying (1991, 1994)). Letting

$$X_{i,t}(b) = h(X_i(t)) - b^T \mathbf{U}_i', \ Y_{j,t}(b) = h(Y_j(t)) - b^T \mathbf{U}_j'', \tag{3.1}$$

modify (2.2)–(2.5) as follows:

$$\begin{aligned}
m_{n,t,b}'(s) &= \sum_{i=1}^{n'} I\{X_{i,t}(b) \geq s\}, \ m_{n,t,b}''(s) = \sum_{j=1}^{n''} I\{Y_{j,t}(b) \geq s\}, \\
N_{n,t,b}'(s) &= \sum_{i=1}^{n'} I\{X_{i,t}(b) \leq s, \delta_i'(t) = 1\}, \\
N_{n,t,b}''(s) &= \sum_{j=1}^{n''} I\{Y_{j,t}(b) \leq s, \delta_j''(t) = 1\}, \\
1 - H_{n,t,b}(s) &= \prod_{u<s} \Big\{1 - \frac{\Delta N_{n,t,b}'(u) + \Delta N_{n,t,b}''(u)}{m_{n,t,b}'(u) + m_{n,t,b}''(u)}\Big\},
\end{aligned}$$

$$S_n(t;b) = \sum_{i=1}^{n'} \delta_i'(t)\psi(H_{n,t,b}(X_{i,t}(b)))\Big\{1 - \frac{m'_{n,t,b}(X_{i,t}(b))}{m'_{n,t,b}(X_{i,t}(b)) + m''_{n,t,b}(X_{i,t}(b))}\Big\}$$
$$- \sum_{j=1}^{n''} \delta_j''(t)\psi(H_{n,t,b}(Y_{j,t}(b)))\frac{m'_{n,t,b}(Y_{j,t}(b))}{m'_{n,t,b}(Y_{j,t}(b)) + m''_{n,t,b}(Y_{j,t}(b))}.$$

From Theorem 1(ii) of Lai and Ying (1991) and the fact that a patient is randomly assigned to treatment X or Y independently of the patient's covariates, it follows that for every $d > 0$, $S_n(t;b) = S_n(t;\beta) + o_p(\sqrt{n})$ uniformly in $|b-\beta| \leq d/\sqrt{n}$ (cf. Lin (1992), Appendix). Since $\{S_n(t;\beta), 0 \leq t \leq t^*\}$ converges weakly to a Gaussian process with independent increments under H_0, it follows that we can apply the same repeated significance tests as those in Section 2 to the rank statistics $S_n(t) := S_n(t;\widehat{\beta}_{n,t})$. The estimates $V_n(t)$ are the same as those in (2.10) but with $H_{n,t,\widehat{\beta}_n}$ replacing $H_{n,t}$, etc.

For the case $h(x) = \log x$ and using the Slud-Wei (1982) method to construct stopping boundaries, Lin (1992) used rank estimates $\widehat{\beta}_{n,t}$ to estimate β in the construction of repeated significance tests. In particular, he applied the time-sequential logrank test ($\psi = 1$) to interim analysis of data from a clinical trial comparing AZT to placebo. He did not notice, however, the independent increments property of the limiting Gaussian process and therefore had to first estimate the covariances of $S_n(t_i)$ and $S_n(t_j)$ (under H_0) in order to compute the probability in (2.11) by multivariate normal approximation. It is also worth noting here that M-estimators are much easier to compute (cf. Kim and Lai (1998)) than the rank estimators, for which Lin (1992) had to use an annealing algorithm to get around the computational complexity.

4. Simulation Results and Discussion

As pointed out in Gu and Lai (1991), the performance of time-sequential censored rank tests depends on the choice of the stopping boundaries and the score function ψ. The choice of stopping boundaries for repeated significance tests has been addressed in Section 2. The choice of score functions has been discussed in Gill (1980) and Gu, Lai and Lan (1991). Suppose that $F = F_0$ and $G = F_\theta$, where $\{F_\theta, -\epsilon \leq \theta \leq \epsilon\}$ is a family of continuous distribution functions whose cumulative hazard functions $\Phi_\theta = -\log(1 - F_\theta)$ satisfy

$$\int_0^\infty |d\Phi_\theta/d\Phi_0 - 1|d\Phi_0 = O(\theta) \text{ as } \theta \to 0, \text{ and}$$
$$\theta^{-1}\{(d\Phi_\theta/d\Phi_0)(s) - 1\} \to \psi^*(F_0(s)) \tag{4.1}$$

for some continuous function ψ^*, the convergence being uniform on each closed subinterval of $\{s : F_0(s) < 1\}$. Then $\psi = -\psi^*$ is an asymptotically optimal choice of score functions. For this score function, $\{n^{-1/2}S_n(t), 0 \leq t \leq t^*\}$ converges weakly under $H_d : \theta = d/\sqrt{n}$ to $\{Z(t) + \mu(t), 0 \leq t \leq t^*\}$, where $Z(\cdot)$ is a zero-mean Gaussian process with independent increments and $\mu(t) = d\mathrm{Var}(Z(t))$ by (2.8) and (2.9). Hence, in this case, $\{(V_n(t)/n,\ S_n(t)/\sqrt{n}), t \geq 0\}$ converges weakly to $\{(v, W(v)), v \geq 0\}$, where $W(v)$ is a Wiener process with drift coefficient d under H_d (and 0 under H_0).

For other choices of score functions ψ, $\{n^{-1/2}S_n(t), 0 \leq t \leq t^*\}$ still converges weakly under $H_d : \theta = d/\sqrt{n}$ to $\{Z(t) + \mu(t), 0 \leq t \leq t^*\}$, where $\mu(t)$ is given by (2.9) with $g(u) = d\psi^*(F(u))$. However, as shown in Gu and Lai (1991), p. 1419, $\mu(t)$ need not be a monotone function of t even for stochastically ordered alternatives. This has important implications on the power of time-sequential versus fixed-duration tests, as will be illustrated in the following.

Example 1. Logrank statistics are the most commonly used test statistics for censored data, and they are asymptotically optimal under proportional hazards alternatives. We present here a simulation study on the performance of repeated significance tests based on logrank statistics. The simulation study is a continuation, with some modifications, of a previous study by Siegmund (1985), pages 129-131, on the performance of (2.12) in which $S_n(t)$ is the logrank statistic and the estimate $V_n(t)$ is chosen to be 1/4 times the total number of failures up to time t (i.e., $V_n(t) = \{N'_{n,t}(t) + N''_{n,t}(t)\}/4$ with $N'_{n,t}$ and $N''_{n,t}$ defined by (2.4)). We shall use the estimator (2.10c), which is applicable to other score functions, instead of Siegmund's estimator.

In his simulation study, Siegmund considers a total of $n = 350$ patients who arrive independently and uniformly over a 3-year interval, to be assigned independently to either treatment with probability 1/2. There are $k = 10$ periodic reviews at times $t = 1, 1.5, \ldots, 5, 5.5$(years). F is assumed to be exponential with mean 3 (years), so the hazard rate $\lambda_0 = 1/3$. Simulations are conducted under the null hypothesis $F = G$ and under proportional hazards alternatives in which G is exponential with hazard rate $\lambda = \lambda_0/1.8, \lambda_0/1.65, \lambda_0/1.5, \lambda_0/1.4$. Therefore $\theta := \log(\lambda_0/\lambda)$ lies between 0 and 0.6. Noting that there is an accumulation of 50-60 units of information at the end of 5 years or roughly 10-12 units per year, and taking $\alpha = 0.05$ in the approximating group sequential test for the drift of a Wiener process with $k = 10$ evenly spaced groups between v_0 and v_1, Siegmund chooses the following design parameters in the repeated significance test (2.12):

$$v_0 = 11,\ v_1 = 55,\ b = 2.85,\ c = 2.05\ (k = 10). \tag{4.2}$$

Our simulation study uses the same design parameters (4.2) as in Siegmund's study. Instead of strictly random treatment allocation and independent, uniformly distributed patient arrival times within the first 3 years, we assume for simplicity that $n' = n'' = 175$ patients are assigned to each group and that of these 175 patients, 59 enter the study during the first year and 29 enter during the next four 6-month periods, recalling that data reviews are at $1, 1.5, \ldots, 5.5$ years. We also change this patient entry pattern in one example (Case 8 of Table 1), in which 87 of the 175 patients enter during the first year and 22 enter during the next four 6-month periods. While Siegmund assumes no censoring due to patient withdrawal, our simulation study assumes that the censoring variables ξ_i', ξ_j'' are i.i.d. exponential with a median of 12 (years). As in Siegmund's study, we assume that F is exponential with hazard rate $\lambda_0 = 1/3$. In addition to the exponential G with the same range of hazard rates as in Siegmund's study, we also consider three other stochastically ordered alternatives listed in the following table (Cases 6-8). In Cases 6 and 8, the hazard rate of G is less than that of F, corresponding to the stochastically ordered case. For stochastically ordered alternatives, the function ψ^* in (4.1) is ≤ 0 and therefore the limiting drift $\mu(t)$ in (2.9) is nondecreasing in t if the score function ψ is nonnegative. In case 7, the hazard rate of G exceeds that of F in the interval $1 < s < 6$.

Table 1. Power and expected duration of repeated significance test based on logrank statistics, compared with fixed duration tests. Each result is based on 2000 simulations. Hazard rate of $F = \lambda_0 (= 1/3)$.

		Repeated Significance Test		Power of Test with Fixed Duration t^*	
	Hazard Rate of G	Power	Expected Duration	$t^* = 5.5$	$t^* = 3$
(1)	$\lambda_0 (F = G)$	.052	5.4	.049	.049
(2)	$\lambda_0/1.4$	.66	4.7	.70	.43
(3)	$\lambda_0/1.5$	.82	4.3	.84	.57
(4)	$\lambda_0/1.65$	.94	3.7	.95	.73
(5)	$\lambda_0/1.8$	.98	3.3	.98	.84
(6)	$\lambda_0/4$ for $0 \leq s \leq 1$ λ_0 for $s > 1$	.86	3.0	.76	.91
(7)	$\lambda_0/4.5$ for $s \leq 1$ or $s \geq 6$ $\lambda_0/0.9$ for $1 < s < 6$	.79	3.1	.56	.88
(8)	$\lambda_0/5$ for $0 \leq s \leq 1$ λ_0 for $s > 1$; entry non-uniform (see text)	.92	2.5	.81	.93

The table compares the repeated significance test with a fixed duration test that terminates at the scheduled time $t^* = 5.5$ (years). For porportional hazards alternatives (Cases 2-5 of the table), the repeated significance test shows little loss in power despite a substantial reduction in trial duration. For the stochastically ordered alternatives in Cases 6-8 of the table, the repeated significance test even has higher power than the fixed duration test despite an expected duration of only 2.5-3 years as compared to 5.5 years of the fixed duration test. To better understand this phenomenon, we compute by simulation the values of the drift, $ES_n(t)$, and the expected value of the estimated variance, $EV_n(t)$, at $t = 1, 1.5, \ldots, 5.5$, for Cases 5-8 of the table. The results, each of which is based on 2,000 simulations, are plotted in Figure 1, which shows an approximately linear relationship between $ES_n(t)$ and $EV_n(t)$ in Case 5 (proportional hazard alternative), monotone but nonlinear relationships in the ordered hazards alternatives of Cases 6 and 8, and an initially increasing but eventually decreasing $ES_n(t)$ as a function of $EV_n(t)$ in Case 7.

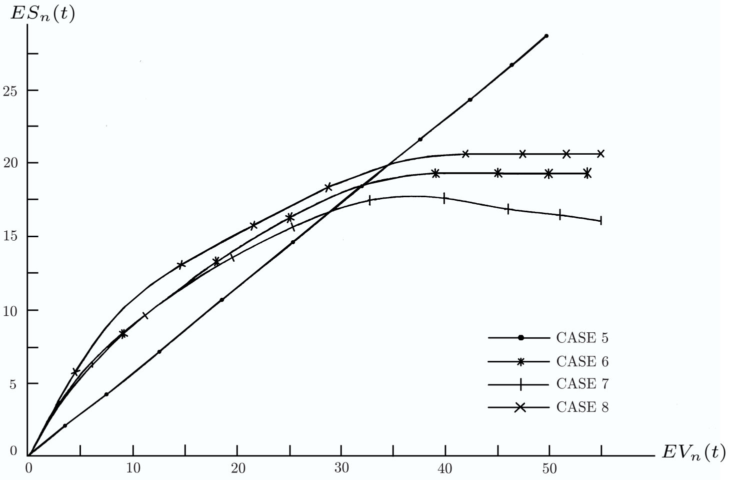

Figure 1. Expected drift vs. expected estimated variance

The figure suggests that for Cases 6-8, a fixed-duration logrank test that terminates earlier may have better power than the one that terminates at $t^* = 5.5$. In particular, we considered the fixed-duration logrank test that terminates at $t^* = 3$ (years), and the results are reported in Table 1. The table shows that, indeed, there is a substantial improvement in power in Cases 6-8 by terminating at $t^* = 3$ than at the later time $t^* = 5.5$. However, this is at the expense of the considerable loss in power for the proportional hazards alternatives of Cases 2-5.

In summary, this simulation study shows that the asymptotic drift of $S_n(t)$ may level off or even decrease with increasing t under stochastically ordered alternatives for which the score function ψ associated with $S_n(t)$ is not asymptotically optimal, such as nonproportional hazards alternatives in the case of logrank statistics. Since the asymptotic variance of $S_n(t)$ continues to increase with t because of the asymptotically uncorrelated increments property under the null hypothesis and contiguous alternatives, the efficacy of $S_n(t)$ may actually decrease with increasing t, therefore allowing repeated significance tests to achieve *both* savings in time and increase in power over fixed-duration tests, as demonstrated in Table 1.

Example 2. The preceding example compares the power and expected duration of Siegmund's repeated significance test (2.12) with a fixed-duration test, based on the commonly used logrank statistics. We now give a comparative study of the performance of the refinement of the Haybittle-Peto test in Section 2.3 and six other stopping rules, again using logrank statistics. We use the program in Gu and Lai (1995) to perform the simulation study which mimics the design specifications in the Beta-Blocker Heart Attack Trial (BHAT). BHAT was a multicenter, double-blind, randomized, placebo-controlled clinical trial designed to test the efficacy of long-term therapy with propranolol given to survivors of an acute myocardial infarction (cf. BHAT (1981)). The trial was scheduled for 4 years, with interim analyses at 11, 16, 21, 28, 34, 40 and 48 months. The trial design assumes an accrucal rate of 149 patients per month for a period of 27 months, so the planned total number of patients is 4123. It is also assumed that each patient is randomized to placebo or treatment upon entering the trial, and is followed for a maximum of 3 years, but has a chance of 7% per year in the study of being lost to follow-up for the placebo group, and 12% (8%, or 6%) during the first (second, or third) year in the study for the treatment group. The actual survival distribution of the placebo group, as reported in BHAT (1982), is a step function with jumps .043, .020, .017, .015, .011 and .018 at 6, 12, 18, 24, 30 and 36 months respectively. The trial design assumes the proportional hazards model with hazard ratio of .699 under the alternative hypothesis H_1 (and 1 under the null hypothesis H_0) of the treatment versus the placebo group. Besides H_1, we also consider H_2 which has time-varying hazard ratios of .599, .708, .615, 1.56, .8 and .323 for each of the six 6-month periods, based on the results reported in BHAT (1982). Table 2 gives the expected duration (in months), and power, under H_0, H_1 and H_2, of the time-sequential logrank test using different stopping rules:

H: Haybittle-type rule in Section 2.3 with $b = 2.9$;

S: Siegmund's rule (2.12) with $b = 2.65, c = 2.15, v_0 = 20, v_1 = 140$;

OB: The O'Brien-Fleming (1979) stopping rule $|S_n(t)| \geq b_i V_n^{1/2}(t_i)$ with $b_1 = 5.46$, $b_2 = 3.87$, $b_3 = 3.16$, $b_4 = 2.74, b_5 = 2.45, b_6 = 2.24$ and $b_7 = 2.07$;

P: Pocock's (1977) stopping rule $|S_n(t_i)| \geq bV_n^{1/2}(t_i)$ with $b = 2.49 (i = 1, \ldots, 7)$;

$\mathrm{U_H}$: Use-function rule defined by (2.16) with $b = 2.9$, $v_0 = 20$, $v_1 = 140$;

$\mathrm{U_{OB}}$ (or $\mathrm{U_P}$): The Lan-DeMets (1983) use-function rule associated with the O'Brien-Fleming (or Pocock) boundary;

F: Fixed-duration rule, stopping at 48 months.

Table 2. Power (p) and expected duration (ET) of different stopping rules in time-sequential logrank tests. Each result is based on 5000 simulations.

		Stopping Rule							
		H	S	OB	P	$\mathrm{U_H}$	$\mathrm{U_{OB}}$	$\mathrm{U_P}$	F
H_0	p	.048	.048	.051	.048	.049	.048	.048	.051
	ET	47.5	47.6	47.6	46.8	47.4	47.1	47.3	48
H_1	p	.970	.972	.972	.940	.969	.957	.962	.981
	ET	31.2	34.3	32.2	28.3	30.7	29.3	29.9	48
H_2	p	.856	.854	.861	.836	.852	.850	.853	.851
	ET	29.8	37.0	32.2	25.7	28.9	28.3	28.6	48

Note that the power of the Haybittle-type rule H, or of Siegmund's rule S, or of the O'Brien-Fleming rule OB is very close to that of the fixed-duration rule F, and that the Haybittle-type rule gives the greatest reduction in trial duration under the alternative hypotheses among the three rules. Although Pocock's rule P seems to have the smallest expected trial duration among all eight rules considered, the power of P is substantially less than that of F. In practice, clinical trialists usually want to have both guaranteed Type I error and guaranteed power at the alternatives that are involved in the determination of the sample size in fixed-duration (fixed-sample-size) trials. They particularly want to avoid getting into the situation where a time-sequential test fails to reject, at the scheduled end of a trial, the null hypothesis which would have been rejected if a fixed-duration test had been used. Table 2 shows that the simple rule H appears to meet these requirements and also to be capable of providing substantial reduction in trial duration where the treatment is indeed efficacious.

Acknowledgements

The research of M. G. Gu was supported by the Natural Science and Engineering Research Council of Canada, and that of T. L. Lai was supported by the National Science Foundation. The authors are grateful to the referees for their constructive comments.

References

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132**, 235-244.

BHAT (β-Blocker Heart Attack Trial Research Group) (1981). β-blocker heart attack trial–design features. *Controlled Clin. Trial* **2**, 275-285.

BHAT (β-Blocker Heart Attack Trial Research Group) (1982). A randomized trial of propranolol in patients with acute myocardial infarction. *J. Amer. Med. Assoc.* **147**, 1707-1714.

DeLong, D. M. (1981). Crossing probabilities for a square root boundary by a Bessel process. *Commun. Statist. - Theor. Meth.* **10**, 2197-2213.

Gill, R. (1980). *Censoring and Stochastic Integrals.* Math. Centre Tract **124**, Mathematische Centrum, Amsterdam.

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Ann. Statist.* **19**, 1403-1433.

Gu, M. G. and Lai, T. L. (1995). Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analysis. Tech. Report, Department of Statistics, Stanford Univ.

Gu, M. G., Lai, T. L. and Lan, K. K. G. (1991). Rank tests based on censored data and their sequential analogues. *Amer. J. Math. and Management Sci.* **11**, 147-176.

Gu, M. G. and Ying, Z. (1995). Group sequential methods for survival data using partial likelihood score processes with covariate adjustment. *Statist. Sinica* **5**, 793-804.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553-566.

Haybittle, J. L. (1971). Repeated assessments of results in clinical trials of cancer treatment. *British J. Radiology* **44**, 793-797.

Jones, D. and Whitehead, J. (1979). Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* **66**, 105-113.

Kim, C. K. and Lai, T. L. (1998). Robust regression with censored and truncated data. To appear in *Multivariate, Design and Sampling* (ed. S. Ghosh). Marcel Dekker, New York.

Lai, T. L. and Ying, Z. (1991). Rank regression methods for left truncated and right censored data. *Ann. Statist.* **19**, 531-556.

Lai, T. L. and Ying, Z. (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data. *Ann. Statist.* **22**, 1222-1255.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

Lan, K. K. G. and DeMets, D. L. (1989). Group sequential procedures: calendar versus information time. *Statistics in Medicine* **8**, 1191-1198.

Lan, K. K. G., DeMets, D. L. and Halperin, M. (1984). More flexible sequential and non-sequential designs in long-term clinical trials. *Comm. Statist. Theory Methods* **13**, 2339-2353.

Lin, D. Y. (1992). Sequential log rank tests adjusting for covariates with the accelerated life model. *Biometrika* **79**, 523-529.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in their consideration. *Cancer Chemotherapy Reports* **50**, 163-170.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. Roy. Statist. Soc. Ser. A* **135**, 185-207.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167-179.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70**, 315-326.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.

Slud, E. 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., Rosner, G. L. and Tritchler, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika* **72**, 365-373.

Department of Mathematics and Statistics, McGill University, Burnside Hall, 805 Sherbrooke Street West, Montreal, QC H3A 2K6, Canada.

E-mail: minggao@math.mcgill.ca

Department of Statistics, Stanford University, Stanford CA 94305-4065, U.S.A.

(Received December 1995; accepted May 1997)