BIOMETRICS 59, 624–631
September 2003

Sequential Conditional Probability Ratio Tests for Normalized Test Statistic on Information Time

Xiaoping Xiong,[1,*] Ming Tan,[2] and James Boyett[1]

[1]Department of Biostatistics, St. Jude Children's Research Hospital, 332 N. Lauderdale St., Memphis, Tennessee 38105, U.S.A.

[2]Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 S. Greene St., Baltimore, Maryland 21201, U.S.A.

*email: xiaoping.xiong@stjude.org

SUMMARY. Sequential conditional probability ratio tests (SCPRTs) provide monitoring procedures with a unique property that a decision reached at early stopping is unlikely to be reversed should the trial continue to the planned end (Xiong, X., 1995, *Journal of the American Statistical Association* **90**, 1463–1473; Tan, M., Xiong, X., and Kutner, M. H., 1998, *Biometrics* **54**, 682–695). It actually provides a probability statement of a conclusion reversal, should the proposed interim analysis plan be used. To broaden its scope of applications, in this article we develop the SCPRT in terms of Brownian motion, which is applicable to most clinical trials with various endpoints. In addition, we utilize the unique structure of the SCPRT to derive a class of adaptive sequential tests that retain the significance level and power in the presence of a nuisance parameter. We illustrate the proposed methods with examples in clinical trials.

KEY WORDS: Clinical trial; SCPRT; Sequential analysis.

1. Introduction

The sequential conditional probability ratio test (SCPRT), proposed by Xiong (1993, 1995), develops the concept that the conclusion of the sequential test based on interim data should hold if the testing continued to the planned end and a nonsequential test was used. This test is derived using the maximum likelihood ratio principle, conditional on the total information at the planned end (e.g., the maximum sample size). Sequential tests are commonly used in clinical trials, and with SCPRT, if a trial is stopped early, then statistically we are confident that the same conclusion would hold had the trial continued to its planned end. The probability of a decision reversal based on the statistical stopping rule is summarized by the discordance probability. In addition, the SCPRT attains the smallest maximum sample size among group sequential tests. Features unique to SCPRT are useful in making early-stopping decisions, as described in more detail for Phase II trials in Tan and Xiong (1996), and for Phase III trials with normal endpoints in Tan, Xiong, and Kutner (1998); they compare SCPRT with other group-sequential tests and the stochastic curtailing procedure. The purpose of this article is to further develop the SCPRT using Brownian motion, because common clinical trials can be generally formulated into this framework, as shown in Lan and Zucker (1993); Whitehead (1999); Jennison and Turnbull (2000). The test statistic in a prospective trial is often indexed as a function of information fraction (time) rather than as a function of sample size, where the information fraction can be roughly understood as the proportion of the variance of currently accumulated samples to the variance of all samples planned. A typical example is the log-rank test statistic, for which the contributions by each sampled individual to the total variance are different. The information fraction is not simply the proportion of sample size at a time point to the total sample size, but is the proportion of the variance that can be approximated by the numbers of events at corresponding time points. Furthermore, utilizing the unique structure of SCPRT by which the sequential test agrees with a nonsequential test, we propose a sequential procedure that is adaptive to updated estimates of nuisance parameters (e.g., unknown variance), so that the overall significance level and power are retained.

2. SCPRTs on Information Time

The sequential test statistic in a planned clinical trial is often indexed as a function of information time (fraction) for which interim information time points can be any set of discrete sequence of $t_1, \ldots, t_k$ in a continuous time interval $[0, 1]$. To develop SCPRT methods for this condition, we first develop SCPRTs for a class of stochastic processes. Sequential test procedures can be embedded in the framework of stochastic processes because a sequential statistic observed at different time points is distributed either exactly or approximately, as a stochastic process is distributed at those time points. For example, let $S^*_{m,n_k} = m^{-1/2}\sigma^{-1}\sum_{i=1}^{n_k} X_i$, where X_i's are samples from normal distribution $\mathcal{N}(\theta, \sigma^2)$, and n_k's and m are given integers. Let S_t for $t > 0$ be a Brownian motion process and let the S_{t_k}'s be values of the process (S_t) at times $t_k = n_k/m$ for the same n_k's and m. Then the joint distribution

of the S^*_{m,n_k}'s is exactly the same as that of the S_{t_k}'s. We may thus regard a sequence of test statistics $(\{S^*_{m,n_k}\}_{k=1}^K)$ as a sequence of observations $(\{S_{t_k}\}_{k=1}^K)$ from a stochastic process $(\{S_t : t>0\})$ at those time points $(t_1, \ldots, t_K)$. We develop SCPRT methods for the stochastic processes and then apply these methods to the statistical tests with samples.

Let S_t, $0 \le t \le t^*$, be a stochastic process, where t is a time index, usually assumed to be a continuous variable in a finite or infinite interval. For example, S_t is a Brownian motion and hence $S_t \sim \mathcal{N}(\theta t, t)$; S_t is a Poisson process and hence $S_t \sim \mathcal{P}(\theta t)$. The properties of these stochastic processes have been well studied and are documented in textbooks (e.g., Karlin and Taylor, 1975). Assume θ is the parameter of underlying distributions for S_t, for all $t>0$, and we want to test $H_0 : \theta \le \theta_0$ vs. $H_a : \theta > \theta_0$. Assume S_t is a sufficient statistic for θ such that the conditional distribution of S_{t_1} given S_{t_2} does not depend on θ for any $t_1 < t_2$. Let t^* be the final information time (e.g., the final stage of the sequential test) and $s^0_{t^*}$ be the cutoff value for S_{t^*}. We define the conditional maximum likelihood ratio as, for $t \le t^*$,

$$L\left(t, S_t \mid t^*, s^0_{t^*}\right) = \frac{\max_{\{s > s^0_{t^*}\}} f(S_t \mid S_{t^*} = s)}{\max_{\{s \le s^0_{t^*}\}} f(S_t \mid S_{t^*} = s)}, \tag{1}$$

where $f(S_t \mid S_{t^*})$, the likelihood of S_t given S_{t^*}, does not depend on θ. Let $(t_1, t_2, \ldots, t^*)$ be a set of discrete and ordered time points (specified times for interim and final looks in clinical trials), and T be the stopping time that equals the first t_k such that $L(t_k, S_{t_k} \mid t^*, s^0_{t^*}) > A$ or $< B$ for some constants A and B. In particular, assume S_t is a Brownian motion process with drift θ on an unit interval, i.e., $S_t \sim \mathcal{N}(\theta t, t)$ for any $t : 0 \le t \le 1$ $(t^* = 1)$. Following the stopping rule above and letting $s^0_{t^*} = z_\alpha$, follow a derivation procedure in Xiong (1993, 1995); we have the upper and lower boundaries for S_{t_k} as

$$\begin{aligned} a_k &= z_\alpha t_k + \{2at_k(1-t_k)\}^{1/2} \quad \text{and} \\ b_k &= z_\alpha t_k - \{2bt_k(1-t_k)\}^{1/2} \quad \text{for} \quad k = 1, \ldots, K, \end{aligned} \tag{2}$$

where K is the total number of looks; $t_1, t_2, \ldots, t_K (= 1)$ are information times of interim looks and the final look. z_α is the upper α-quantile of standard normal distribution; a and b are boundary coefficients that can be determined by a procedure discussed in following section. Then T, the stopping time, equals the first t_k such that $S_{t_k} \notin (b_k, a_k)$. The information time t_k is the ratio of variance at the current stage (kth look) to the variance at the final stage. For examples in applications, $t_k = Var(S^*_{m,n_k})/Var(S^*_{m,m}) = Var(S^*_{n_k})/Var(S^*_m)$, where $S^*_{n_k} = \sum_{i=1}^{n_k} X_i$ and m is the number of samples at the final stage. In particular, $t_k = n_k/m$ if the X_i's are independent and identically distributed. For the log-rank test in survival analysis, $t_k \approx$ (# of events at stage k)/(# of all projected events).

2.1 *Probability of Discordance for Determining SCPRT Boundary*

Appropriate boundaries in (2) should be chosen such that the probability of the conclusion by sequential test being reversed by the test at the planned end is small, but not unnecessarily too small (Xiong, 1995; Tan, et al., 1998). Let D be the event that the conclusion at some interim time t be reversed at the final time $t^* = 1$. Let $\rho(\theta) = P_\theta(D)$, which is the discordance probability given true θ, and let $\rho_{\max} = \max_\theta \rho(\theta)$, which is the maximum discordance probability. Let $\rho_s = P_\theta(D \mid S_{t^*} = s)$, which is the conditional probability of discordance given the last stage observation $S_{t^*} = s$, and let $\rho = \max_s \rho_s$, which is the maximum conditional probability of discordance. Boundary coefficients a and b in (2) are determined by choosing appropriate ρ or $\rho_{\max}$, which provide intuitive benchmarks for designing a sequential test.

Suppose a test statistic S_t is approximated by Brownian motion, i.e., $S_t \sim \mathcal{N}(\eta t, t)$ for $t \in [0, 1]$. We want to test the hypotheses $H_0 : \eta \le 0$ vs. $H_a : \eta > 0$ with significance level α and power $1 - \beta$ for detecting an alternative $\eta_a = z_\alpha + z_\beta$. A nonsequential test at $t^* = 1$ is defined by letting S_1 be the test statistic with critical value z_α. An SCPRT is defined by boundaries given by (2). For simplicity, from now on, we consider only the symmetric boundaries, i.e., those for which $a = b$. For balanced information time (or equal information increments, i.e., $t_k - t_{k-1} = 1/K$), a is calculated for a given ρ in Table 1, for $K = 2, \ldots, 10$, using methods for calculating probabilities of multiple and ordered hittings developed by Xiong (1996), Xiong and Tan (2001), and Xiong, Tan, and Kutner (2002). For different K's, $\rho_{\max}$ is about the same, as shown in the last two columns for $K = 2$ and 10. Table 1 does not depend on α and β.

EXAMPLE 1: Assume $X_1, X_2, \ldots, X_m$ are from $\mathcal{N}(\mu, \sigma^2)$ with known σ^2. We want to test $H_0 : \mu \le \mu_0$ vs. $H_a : \mu > \mu_0$ with significance level α and power $1 - \beta$ for detecting an alternative μ_a. The sample size m for a nonsequential test is calculated by $m = \sigma^2/\lambda_0^2$, where $\lambda_0^2 = \{(\mu_a - \mu_0)/(z_\alpha + z_\beta)\}^2$. Let $S_t = m^{-1/2}\sigma^{-1}\sum_{i=1}^{[mt]}(X_i - \mu_0)$; then $S_t \sim \mathcal{N}(\eta t, t)$ for $t = 0, 1/m, \ldots, 1$, where $[mt]$ denotes the largest integer equal to or smaller than mt. $\eta = m^{1/2}(\mu - \mu_0)/\sigma = (z_\alpha + z_\beta)(\mu - \mu_0)/(\mu_a - \mu_0)$, and $\eta_0 = 0$ and $\eta_a = z_\alpha + z_\beta$ correspond to μ_0 and μ_a, respectively. We then have an SCPRT with a maximum sample size m and interim sample sizes $n_k = t_k m$ for $k = 1, \ldots, K - 1$.

Assume $K = 4$ and $\rho = 0.02$; then the balanced information time is (0.25, 0.5, 0.75, 1) and hence $a = b = 2.953$ by Table 1. At significance level $\alpha = 0.05$, the upper and lower boundaries are (1.463, 2.038, 2.286, 1.645) and (−0.641, −0.393, 0.181, 1.645) by (2), respectively; the maximum discordance probability $\rho_{\max} = 0.0054$ (Table 1). If we choose $\rho = 0.005$, then $a = b = 4.227$ and the upper and lower boundaries are (1.670, 2.276, 2.493, 1.645) and (−0.848, −0.632, −0.025, 1.645), respectively, and $\rho_{\max} = 0.0012$. The sequential boundaries can be represented in terms of $a_k t_k^{-1/2}$ and $b_k t_k^{-1/2}$ for the Z statistic $Z_{t_k} = S_{t_k} t_k^{-1/2}$. For convenience in practice, the sequential boundaries can be represented in terms of a nominal critical significance level by $P_{a_k} = 1 - \Phi(a_k t_k^{-1/2})$ and $P_{b_k} = 1 - \Phi(b_k t_k^{-1/2})$. A smaller ρ results in a larger a (and b) so that upper and lower boundaries are further apart, which leads to a larger expected sample size $(= mE_\mu\{T\})$ and a smaller maximum discordance probability $(\rho_{\max})$. For example, if $\rho = 0.02$, then the nominal critical significance levels for rejection and acceptance of null hypothesis are (0.0017, 0.002, 0.0042, 0.05) and (0.9, 0.7105, 0.417, 0.05), respectively. If $\rho = 0.005$, the nominal critical significance levels are (0.0004, 0.0006, 0.002, 0.05) and (0.9549, 0.8138, 0.5114, 0.05), respectively.

Table 1
Design parameters[1,2] for SCPRTs on information time[3,4]

	Coefficient a: for tests with K looks									$\rho_{\max}$	
ρ (all K)	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	$K=2$	$K=10$
0.001	4.750	5.333	5.675	5.921	6.115	6.275	6.411	6.527	6.627	0.0002	0.0002
0.005	3.315	3.895	4.227	4.459	4.636	4.778	4.895	4.994	5.080	0.0012	0.0012
0.01	2.699	3.271	3.595	3.819	3.987	4.121	4.232	4.325	4.401	0.0025	0.0026
0.02	2.109	2.645	2.953	3.166	3.327	3.456	3.562	3.652	3.729	0.0054	0.0054
0.03	1.769	2.285	2.583	2.789	2.945	3.068	3.170	3.257	3.329	0.0084	0.0085
0.04	1.532	2.031	2.320	2.521	2.672	2.792	2.892	2.975	3.048	0.0116	0.0116
0.05	1.353	1.835	2.118	2.313	2.460	2.577	2.674	2.757	2.828	0.0150	0.0149
0.06	1.209	1.678	1.951	2.142	2.287	2.402	2.597	2.578	2.648	0.0184	0.0183
0.07	1.089	1.545	1.813	2.000	2.141	2.254	2.347	2.426	2.494	0.0220	0.0218
0.08	0.987	1.431	1.693	1.876	2.015	2.125	2.217	2.294	2.361	0.0257	0.0254
0.09	0.898	1.331	1.588	1.767	1.903	2.012	2.101	2.178	2.243	0.0294	0.0290
0.10	0.821	1.243	1.494	1.669	1.803	1.910	2.000	2.072	2.138	0.0333	0.0328
0.15	0.537	0.907	1.133	1.294	1.416	1.515	1.597	1.666	1.726	0.0537	0.0529
0.20	0.354	0.677	0.881	1.027	1.140	1.231	1.307	1.371	1.427	0.0761	0.0750

[1] Boundary coefficient a; number of looks K.
[2] Maximum conditional discordance probability ρ; maximum discordance probability ρ_{max}.
[3] Parameters are calculated for balanced information time.
[4] Table 1 can also be used for design of SCPRTs with unbalanced information time (see Table 2).

For unbalanced information times, we still can use a in Table 1 to calculate SCPRT boundaries. This only results in some slight changes in discordance probability. With random numbers from the uniform distribution on [0, 1], we generate unbalanced information times in Table 2. Assume unbalanced information times (0.236, 0.632, 0.852, 1) and $\rho = 0.02$; we then have $a = b = 2.953$ from Table 1 for $K = 4$. The lower and upper boundaries are (1.420, 2.212, 2.264, 1.645) and (−0.644, −0.133, 0.538, 1.645), respectively. The rejection and acceptance nominal critical significance levels are (0.0017, 0.0027, 0.0071, 0.05) and (0.907, 0.566, 0.280, 0.05), respectively. For this set of boundaries, the nominal ρ and $\rho_{\max}$ are 0.02 and 0.0054, respectively; the actual ρ and $\rho_{\max}$ are 0.021 and 0.0049, respectively. For different K's, as shown in Table 2, the actual ρ is similar to the nominal ρ; the actual $\rho_{\max}$ is smaller than the nominal $\rho_{\max}$, unless t_1 is much closer to 0 than it should be in the balanced information time.

An SCPRT procedure can be summarized as follows. One first calculates the sample size m for a fixed sample size test, according to significance level α, power $1-\beta$, and alternative δ_a. Given the number of looks, K, and the maximum conditional discordance probability, ρ (or the maximum discordance probability $\rho_{\max}$), one determines a and b by Table 1 and uses them to calculate SCPRT boundaries. m is used for calculating the information times and normalizing the test statistic (S_t on $0 \le t \le 1$). To simplify the test procedure in practice; one does not need to normalize S_t; one only needs to sequentially compare the p-values of tests (as fixed sample tests) with the upper and lower critical significance levels at the t_k's. We recommend $\rho = 0.02$ which leads to a reasonable maximum discordance probability ($\rho_{\max} = 0.0054$) and results in an SCPRT boundary that is efficient as well as preserving the accordance (agreement) of conclusions for the test at the early stopping and the test at the planned end.

3. Adaptive SCPRTs for Unknown Variance

Sample size calculation depends on the true variance σ^2 of the underlying distribution, which is usually not well estimated at the planning stage of the study. If σ^2 is underestimated,

Table 2
Examples of SCPRTs with unbalanced information times

K	a	Information times $(t_1, \ldots, t_K)$ $(t_K \equiv 1)$	ρ^1	$\rho^2_{\max}$
2	2.109	(0.135, 1)	0.020	0.0103
3	2.645	(0.391, 0.544, 1)	0.018	0.0050
4	2.953	(0.236, 0.632, 0.852, 1)	0.021	0.0049
5	3.166	(0.287, 0.453, 0.640, 0.934, 1)	0.020	0.0044
6	3.327	(0.299, 0.589, 0.605, 0.660, 0.759, 1)	0.016	0.0036
7	3.456	(0.236, 0.444, 0.610, 0.750, 0.816, 0.939, 1)	0.020	0.0041
8	3.562	(0.031, 0.370, 0.549, 0.801, 0.803, 0.929, 0.951, 1)	0.020	0.0050
9	3.652	(0.272, 0.297, 0.405, 0.492, 0.508, 0.547, 0.652, 0.737, 1)	0.015	0.0038
10	3.729	(0.073, 0.131, 0.247, 0.484, 0.607, 0.711, 0.805, 0.82, 0.853, 1)	0.019	0.0056

[1,2] If balanced information time were used, then $\rho = 0.02$ and $\rho_{\max} \approx 0.0054$ for all K.

the calculated sample size could be substantially smaller than actually required, leading to an underpowered test with potentially serious consequences. For example, a clinical trial on a new drug may be rejected for further investigation, not because of a lack of efficacy, but because of an underestimated sample size. On the contrary, if sample size is overestimated, then time and resources would be wasted. Wittes and Brittain (1990) and Gould and Shih (1992) propose methods for reestimating sample size from internal pilot data where no early stopping is considered. Gould and Shih (1998) include interim analyses that allow early stopping in addition to sample size reestimation. Denne and Jennison (2000) propose updating sample sizes for group sequential tests, based on Stein's two-stage procedure (Stein, 1945) and its generalization. References on flexible monitoring of group sequential tests can be found in Chapter 7 of Jennison and Turnbull (2000). Xiong and Tan presented adaptive group sequential tests based on SCPRTs at the International Chinese Statistical Association (ICSA) 1996 Applied Statistics Symposium (June 8–9, 1996, Johns Hopkins University, Baltimore, Maryland) and at the Biometrics East North American Region 2000 Meeting (March 19–22, 2000, Chicago, Illinois). In this article, we show that the adaptive SCPRT retains the power of sequential tests even when variances are unknown.

To design a test with given significance level and power for some alternative, if the true σ^2 is unknown, then $\hat{\sigma}^2$ (an estimate of σ^2) will be used to (1) calculate the sample size, and (2) calculate the test statistic. To develop procedures that retain the power of the test, we propose compensating for losses of power in the above two calculations separately. To facilitate development, we first propose an adaptive nonsequential test in which the sample size is calculated with $\hat{\sigma}^2$ and the test statistic is calculated with σ^2. Based on the adaptive nonsequential test, we then propose an SCPRT z-test for which the sample size at the final stage of the sequential test is calculated using $\hat{\sigma}^2$, whereas the test statistic is calculated using σ^2. Finally, based on the SCPRT z-test, we propose an adaptive SCPRT t-test for which final sample size and test statistic are both calculated using the updated $\hat{\sigma}^2$ at each stage. Obviously, it is this version of adaptive SCPRTs that are applicable in practice.

3.1 *An Adaptive Nonsequential Test*

Assume σ^2 is the true variance of the underlying distribution (as in Example 1). Let $\hat{\sigma}_p^2$ be an estimate of σ^2 based on previous or preliminary data. Let m be the sample size of a nonsequential test with significance level α and power $1-\beta$ for an alternative μ_a. If σ^2 is unknown, then m is unknown; assume m is estimated by $\hat{m} = \hat{\sigma}_p^2/\lambda_0^2$, where $\lambda_0^2 = \{(\mu_a - \mu_0)/(z_\alpha + z_\beta)\}^2$ (z_α and z_β are upper α- and β-quantiles of standard normal distribution). Assume $S_{\hat{m}} = \hat{m}^{-1/2}\sigma^{-1}\sum_{i=1}^{\hat{m}}(X_i - \mu_0)$ and z_α is the cutoff value for $S_{\hat{m}}$, where σ^2 is the true variance. We rewrite $\hat{m} = \sigma^2/\hat{\lambda}_0^2$, where $\hat{\lambda}_0^2 = \{(\mu_a - \mu_0)/(z_\alpha + z_{\hat{\beta}})\}^2$ and $\hat{\beta} = 1 - \Phi(\hat{m}^{1/2}\sigma^{-1}|\mu_a - \mu_0| - z_\alpha)$; then the conditional power at alternative μ_a is $1-\hat{\beta}$, given sample size $\hat{m}$. The overall power of the test, which includes a step of estimating m by $\hat{m}$ and a step of testing conditional on $\hat{m}$, is $E(1-\hat{\beta})$. Let $R \equiv \hat{\sigma}_p^2/\sigma^2$. Then $R \sim \chi^2(n-1)/(n-1)$, where n is the sample size for $\hat{\sigma}_p^2$. Because $m = \sigma^2\{(\mu_a - \mu_0)/(z_\alpha + z_\beta)\}^{-2}$ and $R = \hat{m}/m$, we have $1 - \hat{\beta} = \Phi\{R^{1/2}(z_\alpha + z_\beta) - z_\alpha\}$. Taking Taylor's expansion of $1-\hat{\beta}$ in terms of R, we have

$$E(1-\hat{\beta}) \approx (1-\beta) - \frac{\phi(z_\beta)}{4(n-1)} \times \left\{(z_\alpha + z_\beta) + (z_\alpha + z_\beta)^2 z_\beta\right\}, \tag{3}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and the density of the standard normal distribution, respectively. The overall power $E(1-\hat{\beta})$ of the above nonsequential test obviously does not equal $1-\beta$. We now derive an adaptive nonsequential test whose overall power approximately equals $1-\beta$ by using equation (3) reversely as follows. Let β^* be the solution of equation

$$(1-\beta^*) - \frac{\phi(z_{\beta^*})}{4(n-1)}\left\{(z_\alpha + z_{\beta^*}) + (z_\alpha + z_{\beta^*})^2 z_{\beta^*}\right\} = 1-\beta. \tag{4}$$

Let $1 - \hat{\beta}^* = \Phi\{(\hat{m}^*)^{1/2}\sigma^{-1}|\mu_a - \mu_0| - z_\alpha\}$, where $\hat{m}^* = \hat{\sigma}_p^2/\lambda_0^{*2}$ and $\lambda_0^{*2} = \{(\mu_a - \mu_0)/(z_\alpha + z_{\beta^*})\}^2$, and $m^* = \sigma_p^2/\lambda_0^{*2}$ is the sample size for achieving power $1-\beta^*$ when σ^2 is known. Then the adaptive nonsequential test with nominal power $1-\beta^*$ and sample size $\hat{m}^*$ has an overall power $E(1-\hat{\beta}^*) \approx 1-\beta$.

EXAMPLE 2. Assuming $\alpha = 0.05$ and $1-\beta = 0.8$, the overall power $E(1-\hat{\beta})$ is calculated in Table 3 for different sample sizes of $\hat{\sigma}^2$. Obviously, $E(1-\hat{\beta})$ is less than $1-\beta$, especially when n is small.

Table 3
The overall power $E(1-\hat{\beta})$ when $1-\beta = 0.8$

Sample size (for $\hat{\sigma}_p^2$) n	10	20	30	40	80	100	120
Overall power $E(1-\hat{\beta})$	0.74	0.772	0.781	0.786	0.793	0.795	0.796

For example, if $n = 10$, then $E(1-\hat{\beta}) = 0.74$, which is substantially smaller than $1-\beta = 0.80$. The nominal power $1-\beta^*$ for achieving an overall power $1-\beta = 0.8$ is calculated by (4) for different n in Table 4. For example, if $n = 10$, then a nominal power $1-\beta^* = 0.867$ is required for achieving an overall power 0.80.

3.2 *Adaptive SCPRT z-Test*

An SCPRT test is in agreement with a nonsequential test, within a probability of discordance that can be chosen to be arbitrarily small (see Section 2). An adaptive nonsequential test maintains power for any unknown σ^2, as proposed above. With the special properties of the SCPRT and the adaptive

Table 4
Nominal power $1-\beta^*$ for achieving $E(1-\hat{\beta}^*) \approx 0.8$

Sample size (for $\hat{\sigma}^2$) n	10	20	30	40	80	100	120
Nominal power $1-\beta^*$	0.867	0.831	0.82	0.815	0.807	0.806	0.805

nonsequential test, we are able to develop a sequential procedure that retains power for an unknown σ^2.

Let $n_1, \ldots, n_{K-1}$ be the partial sample sizes of a sequential test, which can be either specified before the sequential test, or specified according to the estimated final sample size during the sequential test. Let $\hat{\sigma}_k^2$ be the estimate of σ^2 obtained from the n_k observations at the kth look of the sequential test. Given significance level α and power $1-\beta$, let $1-\beta^*(n_k)$ be the nominal power calculated by (4) with $n=n_k$, where n_k is the sample size used for calculating $\hat{\sigma}_k^2$. Let $\hat{m}_k = \hat{\sigma}_k^2 \lambda_k^{-2}$ be an estimate of m at the kth look, where $\lambda_k^2 = \{(\mu_a - \mu_0)/(z_\alpha + z_{\beta^*(n_k)})\}^2$; $\hat{m}_k$ is the final sample size estimated at the kth look. Let $(\hat{t}_1, \ldots, \hat{t}_K)$ be the information time of the sequential test, where $\hat{t}_k = n_k/\hat{m}_k$ for $k \le K-1$. Actually, $\hat{t}_k$ is an estimate of $t_k = n_k/m$, as m is unknown and hence t_k is unknown. Define $n_K \equiv \hat{m}_{K-1}$ and $\hat{m}_K \equiv \hat{m}_{K-1}$; then $\hat{t}_K \equiv 1$. In case of $\hat{t}_k > 1$ for some $k^* < K$, this k^* is taken as the last look K and we define $t_{k^*} = t_K = 1$. Now we have an adaptive SCPRT z-test for which a test statistic $S_{\hat{t}_k} = \hat{m}_k^{-1/2}\sigma^{-1}\sum_{i=1}^{n_k}(X_i - \mu_0)$, and a stopping time T where T is the first $\hat{t}_k$ such that $S_{\hat{t}_k} > a_{\hat{t}_k}$ or $< b_{\hat{t}_k}$ for $k=1,\ldots,K$. If $a_{\hat{t}_k}$ and $b_{\hat{t}_k}$ are calculated by (2) using small ρ (or $\rho_{\max}$) for all k's, then the sequential test has a significance level $\approx \alpha$ and an overall power $\approx 1-\beta$ for any variance σ^2. This is true for several reasons. Assume the sequential test stops at some kth look. The sequential test and the nonsequential test with sample size $\hat{m}_k$ projected at the kth look reach the same conclusion except for a very small probability. In addition, the power of the nonsequential test based on the variance estimate at the kth look is $E\{1-\hat{\beta}^*(n_k)\} \approx 1-\beta$, for any k. Hence, the sequential test has an overall power $\approx 1-\beta$. In above derivation, we implicitly used the assumption that the test statistic (e.g., $\hat{\mu}$) for nonsequential tests is independent of the estimated sample size $\hat{m}$. This assumption holds under the normal distribution because $\hat{\mu}$ is independent of $\hat{\sigma}^2$. Hence, the SCPRT z-test holds when the test statistic is approximately independent of the estimated variance for the corresponding nonsequential tests. If the test statistic is approximated by Brownian motion with independent increments, as in most clinical trials, then the above requirement is sufficiently met.

3.3 *Adaptive SCPRT t-Test*

Replacing σ^2 with $\hat{\sigma}^2$ for calculating the test statistic of the adaptive SCPRT z-test, we obtain an adaptive SCPRT t-test for which the test statistic is $S_{\hat{t}_k} = \hat{m}_k^{-1/2}\hat{\sigma}_k^{-1}\sum_{i=1}^{n_k}(X_i - \mu_0)$, and the sequential boundaries are $a^*_{\hat{t}_k} = t_{\alpha,n_k-1}\hat{t}_k + \{2a\hat{t}_k \times (1-\hat{t}_k)\}^{1/2}$ and $b^*_{\hat{t}_k} = t_{\alpha,n_k-1}\hat{t}_k - \{2a\hat{t}_k(1-\hat{t}_k)\}^{1/2}$ for $k=1,\ldots,K$; t_{α,n_k-1} is the upper α-quantile of t distribution with degree of freedom n_k-1. For the same sample size, a nonsequential z-test and a nonsequential t-test have the same significance level and similar powers. Because an SCPRT agrees with its corresponding nonsequential test, the adaptive SCPRT z-test and the adaptive SCPRT t-test have similar significance level and power, which is verified by simulations in Section 3.4. The actual significance level is slightly inflated from the nominal significance level; we may use a slightly deflated nominal significance level to obtain the targeted actual significance level.

3.4 *Simulation Results*

For significance level $\alpha=0.05$, power $1-\beta=0.8$, and number of looks $K=4$, we simulated 100,000 tests (for each setting) for the two adaptive SCPRT procedures. The information time $(\hat{t}_1, \ldots, \hat{t}_4)$ is determined by $\hat{t}_k = n_k/\hat{m}_k$ where $n_1=10$ and $n_k = k\hat{m}_{k-1}/4$ for $1 < k \le 4$. The operating characteristics for the two tests are calculated in Table 5 along with those for the reference fixed sample size test. The significance level and power for the adaptive SCPRT z-test are very close to the design parameters, especially when ρ is small. The adaptive SCPRT t-test is comparable to the fixed sample size test in power, but has a slightly inflated significance level. For example (Table 5), for a SCPRT t-test with $\sigma^2=64$ and $\rho=0.02$, the nominal significance level $\alpha=0.05$, whereas the actual significance level $\alpha^*=0.067$. However, by starting with $\alpha=0.04$, we have $\alpha^*=0.053$, which is close to the targeted significance level 0.05.

Table 5
Operating characteristics for sequential tests with unknown σ^2

Design parameters		Adaptive SCPRTs[a,b] (H_0:$\mu \le \mu_0$ vs. H_a : $\mu > \mu_0$; $\alpha=0.05$, $1-\beta=0.8$ for $\delta=\mu-\mu_0=0.5$)[c]											
		z-test[d] with $\alpha=0.05$				t-test[e] with $\alpha=0.05$				t-test[e] with $\alpha=0.04$			
σ^2	ρ	α^*	$1-\beta$	$E_{\delta_0}N$	$E_{\delta_a}N$	α^*	$1-\beta$	$E_{\delta_0}N$	$E_{\delta_a}N$	α^*	$1-\beta$	$E_{\delta_0}N$	$E_{\delta_a}N$
1	0.005	0.051	0.778	20.6	22.1	0.063	0.804	20.3	22.3	0.051	0.802	22.2	24.6
	0.02	0.051	0.776	19.4	20.9	0.064	0.804	19.2	21.4	0.053	0.802	20.8	23.5
16	0.005	0.052	0.797	324	357	0.056	0.794	318	350	0.045	0.797	353	391
	0.02	0.056	0.794	300	336	0.063	0.789	292	327	0.050	0.790	323	367
64	0.005	0.052	0.799	1299	1433	0.060	0.793	1277	1405	0.047	0.796	1416	1570
	0.02	0.057	0.795	1201	1346	0.067	0.787	1167	1309	0.053	0.790	1296	1470

[a] For sequential procedures, $K=4$; α is the nominal significance level and α^* is the actual significance level.
[b] $E_{\delta_0}N$ and $E_{\delta_a}N$ are expected sample sizes under null and alternative, respectively.
[c] A fixed sample size test has a sample size of $m=25$, 396, and 1583, respectively, for $\sigma^2=1$, 16, and 64.
[d] SCPRT z-test assumes σ^2 unknown for calculating sample size, but known for calculating the test statistic.
[e] SCPRT t-test assumes σ^2 unknown for both calculations.

4. Sequential Clinical Trials Based on SCPRTs

We illustrate SCPRT methods for clinical trials with several applications, including comparing survival curves and proportions, as well as unplanned interim analyses.

4.1 *Comparing Two Survival Curves*

To compare two survival curves using the log-rank test, assume the accrual calendar times are random and subjects are randomly assigned to two groups with probability π and $1-\pi$, respectively; assume distinct events time $\tau_1 < \tau_2 < \cdots < \tau_r \cdots$ across the two groups. At time τ_j, d_{1j} events (e.g., death) occur in group 1 and d_{2j} events occur in group 2, for $j = 1, 2, \ldots$. Just before time τ_j, there are n_{1j} and n_{2j} subjects at risk in the two groups, respectively; there are $d_j = d_{1j} + d_{2j}$ events among $n_j = n_{1j} + n_{2j}$ individuals at risk. The sequential log-rank test statistic (S_t, t) can be obtained by taking the log-rank score for S_t and taking the Mantel-Haenszel variance for t (Armitage, 1975). At the kth look, the log-rank score $U_k = \sum_{\tau_j \le \tau(k)} (d_{1j} - e_{1j})$ where $e_{1j} = n_{1j} d_j / n_j$ and $\tau(k)$ is the calendar time of the kth look; the variance $V_k = \sum_{\tau_j < \tau(k)}^{k} v_{1j}$, where $v_{1j} = n_{1j} n_{2j} d_j (n_j - d_j) / \{n_j^2 (n_j - 1)\}$. Let $r = \lambda_c / \lambda_t$ be the ratio of hazards λ_c and λ_t for survival functions for the control and treatment groups, respectively. To test $H_0 : r \le 1$ vs. $H_a : r > 1$ for significance level α and power $1-\beta$ to detect an alternative r_a, the planned final variance is calculated by $V_K = (z_\alpha + z_\beta)^2 / [4\pi(1-\pi)\{\ln(r_a)\}^2]$. Denote $\theta = \ln(\lambda)$ and let $S_{t_k} = U_k / {V_K}^{1/2} \sim \mathcal{N}(\theta t_k, t_k)$ where $t_k = V_k / V_K$ for $k = 1, \ldots, K$, the SCPRT boundaries are given by (2). Assume $\pi = 0.5$; then because $V_k \approx \sum_{j \le k} d_j / 4$, the information time can also be approximated simply by the ratio of numbers of events, i.e., $t_k \approx$ (# of events at stage k)/(# of all projected events) for $k = 1, \ldots, K$.

EXAMPLE: *beta-blocker heart attack trial* We reexamine early stopping issues of the well-studied Beta-Blocker Heart Attack Trial (BHAT) as an illustrative example. BHAT was a randomized double-blind trial comparing propranolol ($n = 1916$) with placebo ($n = 1921$) in patients with recent myocardial infarction. The trial was terminated early because of a better-than-expected treatment benefit (DeMets et al., 1984). The study was designed with two-year accrual and two-year follow-up, an anticipated total number of deaths of 628, and seven interim analyses, at the time the Policy and Data Monitoring Board met. The study was to detect the difference of three-year mortality rates of 0.1746 vs. 0.1375, respectively, for the placebo and treatment groups, which leads to hazard ratio $r_a = \lambda_t / \lambda_c = \ln(1 - 0.1746) / \ln(1 - 0.1375) = 1.297$. So the trial was designed to detect a log-hazard ratio of $\theta_a = \ln(r_a) = 0.26$ as the minimum difference of clinical importance with a significance level $\alpha = 0.05$ and power $1 - \beta = 0.9$. In this study, the O'Brien-Fleming boundary with six interim analyses was originally employed, which was crossed at the sixth interim analysis; the trial was then stopped nine months early.

To implement SCPRT, note the numbers of deaths at the first six interim analyses are 56, 77, 126, 177, 247, and 318, respectively. The observed number of deaths at 3.25 years of study was 318, which indicated that number of deaths at the planned end of study (four years) could be substantially smaller than the anticipated 628 deaths. A reasonable guess of the number of deaths at the planned end was 408 (DeMets et al., 1984). Then the information time is (0.137, 0.189, 0.309, 0.434, 0.605, 0.779, 1). Let $\alpha = 0.025$ (the study was designed with a two-sided test) and let $\rho = 0.03$. Then $a = b = 3.068$ for $K = 7$ from Table 1, by which the lower and upper boundaries for the z statistic Z_{t_k} are $(-1.421, -1.231, -0.833, -0.449, 0.073, 0.645, 1.96)$ and $(2.873, 2.934, 3.011, 3.030, 2.977, 2.816, 1.96)$, respectively (see Section 2.1). The SCPRT boundary is still crossed at the sixth interim analysis because the value of the standardized log-rank statistic for the difference in mortality is 2.820 and the boundary value is 2.816. The maximum discordance probability between the sequential test and the nonsequential test at the planned end of the four-year study is $\rho_{\max} = 0.0085$ (see Table 1 for $\rho = 0.03$), which indicates that it is highly unlikely that the conclusion would be reversed had the trial continued to the planned end. The discordance probabilities provide a less conservative assessment of the likelihood of trend reversal than does the stochastic curtailing procedure (Tan et al., 1998). The discordance probability refers to the probability of decision reversal of the whole interim analysis procedure, whereas stochastic curtailing (Lan, Simon, and Halperin, 1982) is local and provides a conditional power of 0.87 (or the conditional probability of reversal is 0.13), assuming the same number of deaths of 408 at the end (DeMets et al., 1984).

4.2 *Comparing Two Response Rates*

Let p_1 and p_2 be the unknown response rates of two treatments and $d = p_2 - p_1$. We want to test $H_0 : d \le 0$ vs. $H_1 : d > 0$ with significance level α and power $1-\beta$ to detect a given d_a. Again n_{1k} and n_{2k} are sample sizes at the kth look of a sequential clinical trial and assume $n_{1k}/n_{2k} = c$ for some c and all $k = 1, \ldots, K$. Let $U_k = \sum_{j=1}^{n_{1k}} X_{1j} - c \sum_{j=1}^{n_{2k}} X_{2j}$. Let $S_{t_k} = U_k V_K^{-1/2}$, then $S_{t_k} \sim \mathcal{N}(\theta t_k, t_k)$ where $t_k = n_{1k}/n_{1K}$ and $V_k = n_{1k}\{p_1(1-p_1) + cp_2(1-p_2)\}$ and $\theta = n_{1K}^{1/2} d / \{p_1(1-p_1) + cp_2(1-p_2)\}^{1/2}$ and $n_{1K} = \{(z_\alpha + z_\beta)/d_a\}^2 \times \{p_1(1-p_1) + c(p_1 + d_a)(1 - p_1 - d_a)\}$. As the final sample size n_{1K} cannot be determined, because nuisance parameter p_1 is unknown, we may use the adaptive SCPRT t-test for which $S_{t_k} = U_k / \hat{V}_K^{1/2}$, where $\hat{t}_k = n_{1k}/\hat{n}_{1K}^{(k)}$ and $\hat{V}_k = n_{1k}\{\hat{p}_{1k}(1-\hat{p}_{1k}) + c\hat{p}_{2k}(1-\hat{p}_{2k})\}$, with $\hat{p}_{ik} = n_{ik}^{-1} \times \sum_{j=1}^{n_{ik}} X_{ij}$ for $i = 1, 2$. The final sample size estimated at the kth look is $\hat{n}_{1K}^{(k)} = \{(z_\alpha + z_\beta)/d_a\}^2 \{\hat{p}_{1k}(1-\hat{p}_{1k}) + c(\hat{p}_{1k} + d_a)(1 - \hat{p}_{1k} - d_a)\}$. If $\hat{p}_{1k} + d_a > 1$, we may use $\hat{n}_{1K}^{(k)} = \{(z_\alpha + z_\beta)/d_a\}^2 \hat{p}_{1k} \times (1 - \hat{p}_{1k})$ or $\hat{n}_{1K}^{(k)} = \{(z_\alpha + z_\beta)/d_a\}^2 \{\hat{p}_{1k}(1-\hat{p}_{1k}) + c\hat{p}_{2k} \times (1 - \hat{p}_{2k})\}$. The sequential boundaries are the same as those in Section 3.3 with replacement of the $t_{\alpha, n_k - 1}$ by $t_{\alpha, n_{1k} - 1}$ or $t_{\alpha, n_{1k} + n_{2k} - 1}$, accordingly.

4.3 *Unplanned Interim Analysis*

In practice, studies without formal planned interim analysis may be stopped early for reasons other than implications from the test statistic. The SCPRT and the discordance probability can be used to evaluate the conclusion made upon the partially collected data. Assume the test statistic S_{t_1} was observed where t_1 is the information time as a ratio of current sample size to the total sample size at the planned end. We may assume an SCPRT with $K = 2$ looks and information time (t_1, t_2), where $t_2 = 1$ corresponds to the planned end. The sequential boundaries are derived from (2) such that S_{t_1} has

just reached one boundary edge at time t_1; hence $S_{t_1} = a_{t_1}$ (if $S_{t_1} > z_\alpha t_1$) or $S_{t_1} = b_{t_1}$ (if $S_{t_1} \le z_\alpha t_1$), by which the coefficient $a = b = (S_{t_1} - z_\alpha t_1)^2 / 2t_1(1 - t_1)$. For an example, 57 subjects in each group were planned for the placebo and drug groups in a study at a local hospital for which no interim analysis was planned. This study was stopped early with 16 patients in each group because of the slower-than-expected accrual and a desire to avoid competition for accrual with a subsequent major study. A nominal p-value $= 0.028$ (for a nonsequential test with 16 in each group) was obtained, which yields $Z_{t_1} = 1.911$ and $S_{t_1} = 1.01245$, because $t_1 = 16/57 = 0.2807$. Assuming $\alpha = 0.05$, then $a = 0.751$, calculated from the equation above. Finally, $\rho_{\max} = 0.0563$, calculated by the ordered-hitting method from Xiong and Tan (2001). This implies that the maximum probability that a decision rejecting H_0 based on partially collected data should be reversed at the planned end is 0.0563. Exact calculation of $\rho_{\max}$ is quite involved and is given in Xiong and Tan (2001). For balanced information times or unbalanced information times that are not too unbalanced (e.g., some examples in Table 2), we may obtain the $\rho_{\max}$ from Table 1 by interpolation according to a. For example, assume $a = 0.751$; this a is between $a = 0.537$ and 0.821 in Table 1 ($K = 2$); the corresponding $\rho_{\max}$ should be between $\rho_{\max} = 0.0537$ and 0.0333, according to Table 1. By interpolation, we have $\rho_{\max} = 0.0383$, which is close to the true $\rho_{\max}$ ($= 0.0373$) for the balanced information time $(t_1, t_2) = (0.5, 1)$, and can be used as a rough estimate of the true $\rho_{\max}$ ($= 0.0563$) for the unbalanced information time $(t_1, t_2) = (0.2807, 1)$.

5. Conclusion

We have proposed an SCPRT procedure that is based on a Brownian motion process, with its sequential boundary modulated by a discordance probability. The proposed procedure is applicable to most clinical trials for which the test statistics herein can be approximated by a Brownian motion. The SCPRT procedure is flexible and easy to design, in addition to having the practical properties mentioned in the Introduction. If the variance σ^2 is known, the maximum sample size of SCPRT is the same as the sample size of the reference fixed sample size test. If σ^2 is unknown and we want to detect a certain difference $(\mu_a - \mu_0)$ with a given power $1 - \beta$ for any true σ^2, then the proposed adaptive SCPRT t-test guarantees a significance level $\approx \alpha$ and power $\approx 1 - \beta$ for any σ^2. Through examples, we have also shown that the SCPRT provides the probability of discordance as a new index to aiding the decision making for stopping a clinical trial early.

Acknowledgements

The research work was supported in part by National Institutes of Health grants R01HL61681 and CA 21765 and by the American Lebanese Syrian Associated Charities, the fundraising arm of St. Jude Children's Research Hospital, both in Memphis, Tennessee. The authors are grateful to the associate editor and two reviewers for their helpful comments and suggestions.

Résumé

Les tests séquentiels conditionnels du rapport de probabilité(SCPRT) fournissent des procédures de monitoring avec la propriété particulière qu'une décision prise en cas d'arrêt précoce a peu de chance d'être inversée si l'essai avait été mené à son terme (Xiong, 1995, JASA 90:1463-1473; Tan, Xiong and Kutner,1998,Biometrics 54:682-695). Le SCPRT fournit en fait la probabilité d'inverser la conclusion si le plan d'analyse intérimaire est utilisé. Pour élargir son champ d'application, nous développons dans cet article le SCPRT en termes de mouvement Brownien applicables à la plupart des essais cliniques. En outre, nous utilisons la structure particulière du SCPRT pour déduire une classe de tests séquentiels adaptatifs qui maintiennent le niveau de signification et la puissance en présence d'un paramètre de nuisance. Nous illustrons les méthodes proposées avec des exemples d'essais cliniques.

References

Armitage, P. (1975). *Sequential Medical Trials,* 2nd edition. Oxford : Blackwell.

DeMets, D. L., Hardy, R., Friedman, L. M., and Lan, K. K. G. (1984). Statistical aspects of early termination in the beta-blocker heart attack trial. *Controlled Clinical Trials* **5,** 362–372.

Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika* **2000,** 125–134.

Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communication in Statistics A* **21,** 2833–2853.

Gould, A. L. and Shih, W. J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* **17,** 89–100.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* London: Chapman and Hall/CRC.

Karlin, S. and Taylor, H. A. (1975). *A First Course in Stochastic Processes,* 2nd edition. New York: Academic Press.

Lan, K. K. G. and Zucker, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine* **12,** 753–765.

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

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16,** 243–258.

Tan, M. and Xiong, X. (1996). Continuous and group sequential conditional probability ratio tests for Phase II clinical trials. *Statistics in Medicine* **15,** 2037–2051.

Tan, M., Xiong, X., and Kutner, M. H. (1998). Clinical trial designs based on sequential conditional probability ratio tests and reverse stochastic curtailing. *Biometrics* **54,** 682–695.

Whitehead, J. (1999). *The Design and Analysis of Sequential Clinical Trials.* New York: Wiley.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials (with comments). *Statistics in Medicine* **9,** 65–72.

Xiong, X. (1993). *Principle of generalized conditional PLR sequential test and its applications.* Technical Report

93-15, Department of Statistics, Purdue University, West Lafayette, Indiana.

Xiong, X. (1995). A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association* **90,** 1463–1473.

Xiong, X. (1996). Absorption probability distributions of random paths from finite populations. *Sequential Analysis* **15**(1), 1–19.

Xiong, X. and Tan, M. (2001). Evaluating sequential tests for a class of stochastic processes. *Computing Science and Statistics* **33,** 30–34.

Xiong, X., Tan, M., and Kutner, M. H. (2002). Computational methods for evaluating sequential tests and post-test estimations via sufficiency principle sufficiency. *Statistica Sinica* **12**(4), 1027–1041.

Received October 2001. *Revised January* 2003.
Accepted January 2003.