

On the inefficiency of the adaptive design for monitoring clinical trials

BY ANASTASIOS A. TSIATIS

*Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695,
U.S.A.*

tsiatis@stat.ncsu.edu

AND CYRUS MEHTA

*Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, Massachusetts 02139,
U.S.A.*

mehta@cytel.com

SUMMARY

Adaptive designs, which allow the sample size to be modified based on sequentially computed observed treatment differences, have been advocated recently for monitoring clinical trials. Although such methods have a great deal of appeal on the surface, we show that such methods are inefficient and that one can improve uniformly on such adaptive designs using standard group-sequential tests based on the sequentially computed likelihood ratio test statistic.

Some key words: Adaptive design; Group-sequential test; Likelihood ratio test.

1. INTRODUCTION

For ethical as well as practical reasons, clinical trials are monitored periodically and, if sufficiently large or small treatment differences are observed at an interim analysis, may be stopped early. Group-sequential tests have been developed that allow for early stopping to reject or accept the null hypothesis while preserving the operating characteristics of the test; that is, maintaining the desired type I error probability under the null hypothesis as well as obtaining the desired power to detect clinically important differences.

Recently, there has been a great deal of interest in what are termed ‘adaptive sequential designs’. Traditionally, in the design of a clinical trial, sample size computations are based on determining the ‘clinically important treatment difference’ that is desired to be detected with some specified power. Often, the criterion for the choice of such a clinically important difference is not straightforward. The appeal of the adaptive design is that it uses observed, estimated treatment differences at interim analyses to modify adaptively the design and sample size. Roughly speaking, the modified sample size is chosen to be that necessary to achieve the desired power for an alternative corresponding to the observed treatment difference at an interim analysis. This process can be repeated several times during the course of the study. Examples of such adaptive designs are given by Shen & Fisher (1999), Cui et al. (1999), Posch & Bauer (1999) and Lehmacher & Wassmer (1999). In order that

the adaptive design attain the desired level of significance, the test statistic used to reject the null hypothesis adaptively weights the increments of the commonly-used test statistic.

In this paper, we prove that such methods are inefficient; that is, for any adaptive test, we can always find a standard group-sequential test which is uniformly better in a manner that will be discussed later. This is proved using a generalisation of the Neyman–Pearson theorem to group-sequential tests. We also illustrate this numerically using an adaptive design that has been recently advocated.

2. NOTATION FOR SEQUENTIAL TESTS

Suppose that decisions to stop a trial, either to reject or accept the null hypothesis, can be made at times $1, \dots, K$ using data represented by the random vectors X_1, \dots, X_K , where X_1 represents the data at the first time, X_2 the data between the first and second time and so forth. We will assume that X_1, \dots, X_K are independent random vectors. This may represent data from K different individuals where, possibly, the decision to stop the study may be made after each individual enters the study, or from groups of individuals if we are considering group-sequential tests.

We begin by considering the problem of testing a simple null hypothesis against a simple alternative hypothesis. The density of X_j under the null hypothesis is denoted by $p_{0j}(x_j)$ and under the alternative hypothesis by $p_{1j}(x_j)$ ($j = 1, \dots, K$). At time j , the decision to stop, and either to reject or to accept the null hypothesis, will be based on the random vector $Y_j = (X_1, \dots, X_j)$, that is all the data available at the j th time point.

Example 1. Consider a clinical trial that will compare a new treatment to placebo. Up to n pairs of patients will be recruited into the trial, where one member of the pair will be randomised to receive treatment and the other placebo. Let Z_1, \dots, Z_n be identically and independently distributed $N(\mu, 1)$ random variables, where Z_j denotes the difference in normally distributed responses between the j th pair of individuals, and μ denotes the mean treatment difference in response. We wish to test the null hypothesis of no treatment difference $H_0: \mu = 0$ against the one-sided alternative $H_A: \mu > 0$. For the time being, we will consider the simple alternative $H_1: \mu = \mu_1 > 0$. Suppose that the decision time points where interim analyses may be conducted are after $n_1 < n_2 < \dots < n_K = n$ observations. Thus, in this example, if we let $n_0 = 0$, $X_j = (Z_{n_{j-1}+1}, \dots, Z_{n_j})$ and $Y_j = (Z_1, \dots, Z_{n_j})$, for $j = 1, \dots, K$, the densities of X_j , under the null and alternative hypotheses, are given by

$$p_{qj}(x_j) = (2\pi)^{-(n_j - n_{j-1})/2} \exp \left\{ - \sum_{l=n_{j-1}+1}^{n_j} (z_l - q\mu_1)^2 / 2 \right\},$$

for $q = 0, 1$ respectively.

A standard K -look group-sequential test will reject H_0 at the first monitoring time point j where $S_j = \sum_{l=1}^{n_j} Z_l$ is sufficiently large, $S_j > u_j$ say, or accept the null hypothesis at the first monitoring time point j where S_j is sufficiently small, $S_j < l_j$ say, where $l_j < u_j$ ($j = 1, \dots, K - 1$) and $u_K = l_K$.

Although this example seems oversimplified, most test statistics used to test treatment differences, whether the outcomes be continuous, discrete, survival or longitudinal, will have, asymptotically, the same distributional structure as above; that is, efficient based tests, properly normalised, which are computed sequentially over time, will have a joint distribution which is asymptotically normal with independent increments and variance proportional to the Fisher information; see Scharfstein et al. (1997). Consequently, for the

general problem, the Fisher information would play the same role as sample size in our example above. In a randomised study comparing two treatments, where the primary outcome is survival, the sequentially computed log-rank test is often used for monitoring. This sequential test would have asymptotically the same distributional structure as the example above with Fisher information roughly proportional to the number of events. We shall return to the simple example above for illustration throughout this paper.

Example 2. As an example of an adaptive design, consider the following adaptation of a standard two-look group sequential test for the hypothesis testing problem described above. In a standard two-look group-sequential test the study may be stopped to reject or accept the null hypothesis at the first look after n_1 observations, according to whether $S_1 > u_1$ or $S_1 < l_1$, and reject or accept the null hypothesis at the second look after n_2 observations, according to whether $l_1 \leq S_1 \leq u_1$ and $S_2 > u_2$ or $l_1 \leq S_1 \leq u_1$ and $S_2 \leq u_2$. The boundaries u_1 , l_1 and u_2 are chosen so that the test has level α under the null hypothesis. We consider an adaptive two-look design which also performs an interim analysis after n_1 observations and rejects or accepts the null hypothesis at the first look according to whether $S_1 > u_1$ or $S_1 < l_1$. If, however, $S_1 \in [l_1, u_1]$ a decision is made possibly to increase the sample size for the second look. To be specific the interval $[l_1, u_1]$ is partitioned into $K - 1$ mutually exclusive sets, Γ_j ($j = 2, \dots, K$) say, where $\cup_{j=2}^K \Gamma_j = [l_1, u_1]$. If $S_1 \in \Gamma_j$, then the second look will be conducted after n_j observations, where without loss of generality $n_1 < n_2 < \dots < n_K$. For $j = 2, \dots, K$, define the test statistic

$$T_j = S_1 + \left(\frac{n_2 - n_1}{n_j - n_1} \right)^{\frac{1}{2}} (S_j - S_1). \tag{1}$$

At the second look the null hypothesis will be rejected if $T_j > u_2$ and accepted if $T_j \leq u_2$. With the statistic T_j constructed as shown above the adaptive test has the same α level as the original standard two-look group-sequential test. Note that, although the adaptive test has at most two looks, it is nevertheless characterised by K distinct monitoring time points. The first look occurs at monitoring time point 1, after n_1 observations. The second look occurs at monitoring time point j , after n_j observations, where j may be any integer between 2 and K . This is inherently different from the standard K -look group sequential test described in Example 1 where the study may be monitored up to K times at monitoring time points corresponding to $n_1 < n_2 < \dots < n_K$ observations. We will return to this adaptive design for illustration in § 4.

Any group-sequential test, using an adaptive design or not, can be represented by a sequence of rejection and acceptance regions $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \dots, K\}$, which are a mutually exclusive and exhaustive partition of the sample space, where $(\mathcal{R}_j, \mathcal{A}_j)$ are Y_j -measurable sets. The events $(\mathcal{R}_j, \mathcal{A}_j)$ correspond to stopping and rejecting or accepting the null hypothesis, respectively, at the j th monitoring time point. In Example 1, $\mathcal{R}_1 = (S_1 > u_1)$, $\mathcal{A}_1 = (S_1 < l_1)$ and, for $j = 2, \dots, K$,

$$\begin{aligned} \mathcal{R}_j &= (l_1 \leq S_1 \leq u_1, \dots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j > u_j), \\ \mathcal{A}_j &= (l_1 \leq S_1 \leq u_1, \dots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j < l_j), \end{aligned}$$

where $u_K = l_K$. In Example 2, $\mathcal{R}_1 = (S_1 > u_1)$, $\mathcal{A}_1 = (S_1 < l_1)$ and, for $j = 2, \dots, K$,

$$\mathcal{R}_j = (S_1 \in \Gamma_j, T_j > u_2), \quad \mathcal{A}_j = (S_1 \in \Gamma_j, T_j \leq u_2).$$

We also define the events $\bar{\mathcal{R}}_j = \mathcal{R}_1 \cup \dots \cup \mathcal{R}_j$ and $\bar{\mathcal{A}}_j = \mathcal{A}_1 \cup \dots \cup \mathcal{A}_j$ to correspond to stopping and rejecting or accepting the null hypothesis, respectively, at or before the j th

monitoring time point for $j = 1, \dots, K$ and the events $\mathcal{C}_j = (\bar{\mathcal{R}}_j \cup \bar{\mathcal{A}}_j)^c$ to correspond to continuing the study beyond the j th interim analysis time point, for $j = 1, \dots, K - 1$. For Example 1, the continuation region $\mathcal{C}_j = (l_1 \leq S_1 \leq u_1, \dots, l_j \leq S_j \leq u_j)$. The events $(\bar{\mathcal{R}}_K, \bar{\mathcal{A}}_K)$ correspond to the overall rejection region and acceptance region for the test and, by construction, $\bar{\mathcal{A}}_K = (\bar{\mathcal{R}}_K)^c$.

A level- α test has the property that $P_0(\bar{\mathcal{R}}_K) = \alpha$, which in turn implies that $P_0(\bar{\mathcal{A}}_K) = 1 - \alpha$, where $P_0(\cdot)$ denotes probability computed under the null hypothesis. The probability of rejecting H_0 at or before the j th time point is given by $P_0(\bar{\mathcal{R}}_j) = \alpha_j$, where the nondecreasing sequence of probabilities $\alpha_1 \leq \dots \leq \alpha_K = \alpha$ denotes the α -spending function of the test as defined originally by Lan & DeMets (1983). Similarly, we can define the probability of accepting H_0 at or before the j th time point by $P_0(\bar{\mathcal{A}}_j) = \theta_j$. We will refer to the nondecreasing sequence $\theta_1 \leq \dots \leq \theta_K = 1 - \alpha$ as the θ -spending function. Note that we allow the possibility that $\alpha_{j-1} = \alpha_j$, which will imply that the null hypothesis cannot be rejected at time point j ; similarly, if $\theta_{j-1} = \theta_j$, then we cannot accept H_0 at the j th time point. For such cases we would accordingly define \mathcal{R}_j or \mathcal{A}_j to be the null set.

From now on, when we refer to group-sequential tests $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \dots, K\}$ with a specified (α, θ) -spending function, we mean tests in which $P_0(\bar{\mathcal{R}}_j) = \alpha_j$ and $P_0(\bar{\mathcal{A}}_j) = \theta_j$, for $j = 1, 2, \dots, K$. We now give a criterion for optimality of group-sequential tests.

DEFINITION. *Among all group-sequential tests $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \dots, K\}$ with a specified (α, θ) -spending function, the test $\{(\mathcal{R}_j^{\text{opt}}, \mathcal{A}_j^{\text{opt}}), j = 1, \dots, K\}$ is optimal within this class if*

$$P_1(\bar{\mathcal{R}}_j^{\text{opt}}) \geq P_1(\bar{\mathcal{R}}_j) \quad (2)$$

for all $j = 1, \dots, K$, where $P_1(\cdot)$ denotes probability computed under the alternative hypothesis.

In making this definition, it is not clear that an ‘optimal’ test will exist. We see later that this is the case when Theorem 1 is proved.

Remark 1. By the definition of optimality (2), the optimal test, if it exists, is not only the most powerful test among all group-sequential tests with a specified (α, θ) -spending function, that is $P_1(\bar{\mathcal{R}}_K^{\text{opt}}) \geq P_1(\bar{\mathcal{R}}_K)$ overall, but also, under the alternative, has a greater chance of rejecting the null hypothesis before or at any decision time point. We emphasise that this does not necessarily imply that the optimal test will always stop the study on average earlier under H_1 , because the study may also be stopped to accept the null hypothesis, but it does imply that the optimal test will stop a study more often earlier to make the correct decision. Consequently if, under the alternative, a non-optimal test stops a study on average earlier than the optimal test, then this could only happen because the null hypothesis is accepted more often earlier on, a situation which, we believe, would be undesirable for investigators trying to demonstrate the superiority of a new treatment.

In § 3, we will demonstrate that, for a specified (α, θ) -spending function, the group-sequential test based on test statistics constructed from the likelihood ratio of the accumulating data is optimal, where we denote the likelihood ratio by $L_j(x_j) = p_{1j}(x_j)/p_{0j}(x_j)$ and $\bar{L}_j(y_j) = \prod_{l=1}^j L_l(x_l)$. For simplicity, from now on, we denote the random variables $L_j(X_j)$ by L_j and $\bar{L}_j(Y_j)$ by \bar{L}_j and will refer to \bar{L}_j as the sequential likelihood ratio test statistic.

To avoid the complications involved with discrete distributions and randomisation tests, we assume that the sequential likelihood ratio test statistic \bar{L}_j is absolutely continuous under the null hypothesis. This is certainly the case for the normal example presented earlier.

The level- α group-sequential test based on the sequential likelihood ratio test statistic, with a prespecified (α, θ) -spending function, is defined as $\{(\mathcal{R}_j^{\text{LR}}, \mathcal{A}_j^{\text{LR}}), j = 1, \dots, K\}$, where $\mathcal{R}_1^{\text{LR}} = (\bar{L}_1 > u_1^{\text{LR}})$, $\mathcal{A}_1^{\text{LR}} = (\bar{L}_1 < l_1^{\text{LR}})$ and, for $j = 2, \dots, K$,

$$\begin{aligned} \mathcal{R}_j^{\text{LR}} &= (l_1^{\text{LR}} \leq \bar{L}_1 \leq u_1^{\text{LR}}, \dots, l_{j-1}^{\text{LR}} \leq \bar{L}_{j-1} \leq u_{j-1}^{\text{LR}}, \bar{L}_j > u_j^{\text{LR}}), \\ \mathcal{A}_j^{\text{LR}} &= (l_1^{\text{LR}} \leq \bar{L}_1 \leq u_1^{\text{LR}}, \dots, l_{j-1}^{\text{LR}} \leq \bar{L}_{j-1} \leq u_{j-1}^{\text{LR}}, \bar{L}_j < l_j^{\text{LR}}) \end{aligned}$$

and $u_K^{\text{LR}} = l_K^{\text{LR}}$.

The upper and lower boundaries $u_j^{\text{LR}}, l_j^{\text{LR}}$, for $j = 1, \dots, K$, are constructed so that $P_0(\mathcal{R}_1^{\text{LR}}) = \alpha_1$, $P_0(\mathcal{A}_1^{\text{LR}}) = \theta_1$ and, for $j = 2, \dots, K$,

$$P_0(\mathcal{R}_j^{\text{LR}}) = \alpha_j - \alpha_{j-1}, \quad P_0(\mathcal{A}_j^{\text{LR}}) = \theta_j - \theta_{j-1}.$$

With these definitions, the above test has level α ; this follows because $P_0(\bar{\mathcal{R}}_K^{\text{LR}}) = \sum_{j=1}^K (\alpha_j - \alpha_{j-1}) = \alpha$, has α -spending function $P_0(\bar{\mathcal{R}}_j^{\text{LR}}) = \sum_{m=1}^j (\alpha_m - \alpha_{m-1}) = \alpha_j$ and θ -spending function $P_0(\bar{\mathcal{A}}_j^{\text{LR}}) = \sum_{m=1}^j (\theta_m - \theta_{m-1}) = \theta_j$.

Note that because the likelihood ratio statistics \bar{L}_j ($j = 1, \dots, K$) are assumed absolutely continuous, the boundaries $(u_j^{\text{LR}}, l_j^{\text{LR}})$ ($j = 1, \dots, K$) are uniquely defined and can be derived recursively using an algorithm such as that in Armitage et al. (1969).

3. OPTIMAL GROUP-SEQUENTIAL TESTS

A relationship that we will use throughout for computing $P_1(\bar{\mathcal{R}}_j)$ is given by

$$P_1(\bar{\mathcal{R}}_j) = E_0(\bar{L}_j I_{\bar{\mathcal{R}}_j}),$$

where $I_{\bar{\mathcal{R}}_j}$ is the indicator function, which equals 1 when $\bar{\mathcal{R}}_j$ is true and equals 0 otherwise, and $E_0(\cdot)$ denotes expectation under the null hypothesis. Since, by construction, $\bar{\mathcal{R}}_j$ is Y_j -measurable, this means that $\bar{\mathcal{R}}_j = (Y_j \in Q_j)$, for some Borel set Q_j in the range space of Y_j . Hence, $E_0(\bar{L}_j I_{\bar{\mathcal{R}}_j})$ equals

$$\int_{Q_j} \frac{p_1(y_j)}{p_0(y_j)} p_0(y_j) dy_j = \int_{Q_j} p_1(y_j) dy_j = P_1(Y_j \in Q_j) = P_1(\bar{\mathcal{R}}_j),$$

where $p_q(y_j) = \prod_{i=1}^j p_{qi}(x_i)$ denotes the density of Y_j under the null, $q = 0$, and alternative, $q = 1$, hypotheses. Consequently, among the class of group-sequential tests $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \dots, K\}$ with a specified (α, θ) -spending function, the optimal group-sequential test $\{(\mathcal{R}_j^{\text{opt}}, \mathcal{A}_j^{\text{opt}}), j = 1, \dots, K\}$, if it exists, must satisfy

$$E_0(\bar{L}_j I_{\bar{\mathcal{R}}_j^{\text{opt}}}) \geq E_0(\bar{L}_j I_{\bar{\mathcal{R}}_j}) \quad (j = 1, \dots, K). \tag{3}$$

To verify (3), it suffices to show that the random variable $\bar{L}_j I_{\bar{\mathcal{R}}_j^{\text{opt}}}$ is stochastically larger than or equal to $\bar{L}_j I_{\bar{\mathcal{R}}_j}$ under the null hypothesis, where, by definition, the random variable U is stochastically larger than or equal to W , under the null hypothesis, if, for all t ,

$$P_0(U > t) \geq P_0(W > t). \tag{4}$$

This will be denoted by $U \succcurlyeq W$.

Therefore, among the class of group-sequential tests $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \dots, K\}$ with a specified (α, θ) -spending function, the test $\{(\mathcal{R}_j^{\text{opt}}, \mathcal{A}_j^{\text{opt}}), j = 1, \dots, K\}$ within this class is optimal if

$$\bar{L}_j I_{\bar{\mathcal{R}}_j^{\text{opt}}} \succcurlyeq \bar{L}_j I_{\bar{\mathcal{R}}_j} \quad (j = 1, \dots, K). \tag{5}$$

The key result is given by the following theorem, which is a generalisation of the Neyman–Pearson theorem to group-sequential tests.

THEOREM 1. *The optimal group-sequential test with a specified (α, θ) -spending function is the group-sequential test $\{(\mathcal{R}_j^{\text{LR}}, \mathcal{A}_j^{\text{LR}}), j = 1, \dots, K\}$, derived using the sequential likelihood ratio test statistics as defined above.*

Proof. The theorem will be proved by verifying (5) using induction. That is, we will show that, for all group-sequential tests $\{(\mathcal{R}_j, \mathcal{A}_j), j = 1, \dots, K\}$ with a specified (α, θ) -spending function, the following conditions hold:

$$\bar{L}_1 I_{\mathcal{R}_1^{\text{LR}}} \geq \bar{L}_1 I_{\mathcal{R}_1}, \quad \bar{L}_1 I_{\mathcal{A}_1^{\text{LR}}} \leq \bar{L}_1 I_{\mathcal{A}_1}, \tag{6}$$

and, for $j = 1, \dots, K - 1$,

$$\bar{L}_j I_{\mathcal{R}_j^{\text{LR}}} \geq \bar{L}_j I_{\mathcal{R}_j}, \quad \bar{L}_j I_{\mathcal{A}_j^{\text{LR}}} \leq \bar{L}_j I_{\mathcal{A}_j} \tag{7}$$

implies that

$$\bar{L}_{j+1} I_{\mathcal{R}_{j+1}^{\text{LR}}} \geq \bar{L}_{j+1} I_{\mathcal{R}_{j+1}}, \quad \bar{L}_{j+1} I_{\mathcal{A}_{j+1}^{\text{LR}}} \leq \bar{L}_{j+1} I_{\mathcal{A}_{j+1}}. \tag{8}$$

We will prove that condition (7) implies condition (8) in three steps.

Step 1. If (7) holds, then this implies that

$$\bar{L}_{j+1} I_{\mathcal{R}_j^{\text{LR}}} \geq \bar{L}_{j+1} I_{\mathcal{R}_j}, \quad \bar{L}_{j+1} I_{\mathcal{A}_j^{\text{LR}}} \leq \bar{L}_{j+1} I_{\mathcal{A}_j}. \tag{9}$$

Step 2. Define

$$\bar{\mathcal{R}}_{j+1}^* = \bar{\mathcal{R}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*), \quad \bar{\mathcal{A}}_{j+1}^* = \bar{\mathcal{A}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} < l_{j+1}^*),$$

where u_{j+1}^* and l_{j+1}^* are chosen so that $P_0(\bar{\mathcal{R}}_{j+1}^*) = \alpha_{j+1}$ and $P_0(\bar{\mathcal{A}}_{j+1}^*) = \theta_{j+1}$. Then, if (9) holds, this implies that

$$\bar{L}_{j+1} I_{\bar{\mathcal{R}}_{j+1}^{\text{LR}}} \geq \bar{L}_{j+1} I_{\bar{\mathcal{R}}_{j+1}^*}, \quad \bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^{\text{LR}}} \leq \bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^*}. \tag{10}$$

Step 3. Finally,

$$\bar{L}_{j+1} I_{\bar{\mathcal{R}}_{j+1}^*} \geq \bar{L}_{j+1} I_{\mathcal{R}_{j+1}}, \quad \bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^*} \leq \bar{L}_{j+1} I_{\mathcal{A}_{j+1}}. \tag{11}$$

Condition (6) of Theorem 1 is equivalent to the Neyman–Pearson theorem. The proofs of Steps 1–3 for condition (7) are given in the Appendix. As a consequence, we conclude that (7) implies (8). This completes the proof of Theorem 1. \square

Remark 2. The assumption that X_1, \dots, X_K are independent is necessary for the proof of Step 1. This assumption is important in showing that the group-sequential test based on the sequential likelihood ratio test statistics, derived using a forward recursion, is optimal. Without independence, counterexamples can be constructed where this is not the case.

Remark 3. If we do not insist on putting the constraints of a θ -spending function on the acceptance regions, then we can consider the problem of finding the optimal group-sequential test among tests with only a specified α -spending function. This would correspond to defining a set of mutually exclusive Y_j -measurable rejection regions \mathcal{R}_j such that $P_0(\mathcal{R}_j) = \alpha_j - \alpha_{j-1}$. The overall acceptance region for such a level- α group-sequential test is $(\mathcal{R}_1 \cup \dots \cup \mathcal{R}_K)^c$. Clearly, the set of group-sequential tests with specified (α, θ) -spending function is a subset of the group-sequential tests with only a specified α -spending function. Using methods similar to those above, we can show that the optimal group-sequential

test, among the class of group-sequential tests with only a specified α -spending function, is the group-sequential test using the likelihood ratio test statistics with upper boundaries only, that is where

$$\bar{\mathcal{R}}_j = \bar{L}_1 \leq \tilde{u}_1^{LR}, \dots, \bar{L}_{j-1} \leq \tilde{u}_{j-1}^{LR}, \bar{L}_j > \tilde{u}_j^{LR} \quad (j = 1, \dots, K),$$

and $\tilde{u}_1, \dots, \tilde{u}_K$ are constructed so that $P_0(\bar{\mathcal{R}}_1) = \alpha_1$, $P_0(\bar{\mathcal{R}}_j) = \alpha_j - \alpha_{j-1}$, for $j = 2, \dots, K$. For such a test we would only accept the null hypothesis at the final K th decision time. Such an optimal test has greater power uniformly through time than the test derived with (α, θ) -spending function restrictions. This test may be desirable if we are interested in stopping early only if the alternative hypothesis is true and are willing to continue collecting data otherwise.

For illustration, in the context of the hypothesis testing problem for Examples 1 and 2 of § 2, Theorem 1 states that, among group-sequential tests with a specified (α, θ) -spending functions, the group-sequential test based on the likelihood ratio test statistics, which has rejection regions

$$\mathcal{R}_j^{LR} = (l_1^{LR} \leq \bar{L}_1 \leq u_1^{LR}, \dots, l_{j-1}^{LR} \leq \bar{L}_{j-1} \leq u_{j-1}^{LR}, \bar{L}_j > u_j^{LR}),$$

and acceptance regions

$$\mathcal{A}_j^{LR} = (l_1^{LR} \leq \bar{L}_1 \leq u_1^{LR}, \dots, l_{j-1}^{LR} \leq \bar{L}_{j-1} \leq u_{j-1}^{LR}, \bar{L}_j < l_j^{LR}),$$

is optimal for testing the null hypothesis $H_0 : \mu = 0$ versus the simple alternative $H_1 : \mu = \mu_1 > 0$. Standard calculations for the normal distribution yield the likelihood ratio test statistics $\bar{L}_j = \exp(\mu_1 S_j - n_j \mu_1^2 / 2)$, where $S_j = Z_1 + \dots + Z_{n_j}$. Since \bar{L}_j is a monotone increasing function of S_j , the optimal group-sequential test is equivalent to the test which rejects H_0 at time j if

$$\mathcal{R}_j^{LR} = (l_1 \leq S_1 \leq u_1, \dots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j > u_j),$$

and accepts H_0 if

$$\mathcal{A}_j^{LR} = (l_1 \leq S_1 \leq u_1, \dots, l_{j-1} \leq S_{j-1} \leq u_{j-1}, S_j < l_j).$$

The constants u_j, l_j ($j = 1, \dots, K$) are derived recursively so that

$$P_0(\mathcal{R}_j^{LR}) = \alpha_j - \alpha_{j-1}, \quad P_0(\mathcal{A}_j^{LR}) = \theta_j - \theta_{j-1}. \tag{12}$$

This is the usual group-sequential one-sided test with upper and lower boundaries with a specified (α, θ) -spending function. As a result of the monotone likelihood ratio property of the normal distribution, this test is independent of the choice $\mu_1 > 0$ and hence it is the uniformly optimal test for the composite hypothesis $H_A : \mu > 0$. Also, by reversing the roles of the α -spending and the θ -spending functions, we can use analogous arguments to show that, among all group-sequential tests with a specified (α, θ) -spending function, the group-sequential test defined above has the property that

$$P_\mu(\bar{\mathcal{A}}_j^{LR}) \geq P_\mu(\bar{\mathcal{R}}_j) \quad (j = 1, \dots, K),$$

when $\mu < 0$; that is, if $\mu < 0$, then the optimal group-sequential likelihood ratio test will accept the null hypothesis with higher probability by time j than any other group-sequential test with the same (α, θ) -spending function uniformly for all $j = 1, \dots, K$.

A consequence of Theorem 1 is that any adaptive test will be less efficient, in the sense described above, or dominated by the corresponding standard group-sequential test based on the likelihood ratio test statistic with the same (α, θ) -spending function. This is illustrated in the following example.

4. NUMERICAL COMPARISONS

We illustrate Theorem 1 by comparing a standard group sequential design, as specified by Example 1, with a two-look adaptive design, as specified by Example 2, where both designs use the same (α, θ) -spending functions. The standard two-look design to be adapted considers up to two equally spaced looks with possible early stopping either to reject or to accept $H_0: \mu = 0$ at the first look using a one-sided level-0.025 test having 90% power to detect $H_1: \mu = \mu_1$ at an effect size of $\mu_1 = 0.466$. We use the Pampallona & Tsiatis (1994) stopping boundaries with shape parameter 0 for both rejection and acceptance of H_0 . The EaSt (2000) software reveals that with these stopping boundaries the above operating characteristics are achieved by setting $n_1 = 25$, $n_2 = 50$, $u_1 = 13.84$, $l_1 = 2.19$ and $u_2 = l_2 = 13.84$. In this design the study is terminated at the first look after accruing n_1 patient pairs if $S_1 > u_1$ or $S_1 < l_1$. If, however, $S_1 \in [l_1, u_1]$ the study proceeds to a second and final look after accruing a total of $2n_1$ patient pairs. Now suppose we intend to adapt this design by increasing the sample size at the second look if $\hat{\mu}$, the estimate of μ at the first look, is smaller than the effect size μ_1 at which 90% power is desired. One adaptive strategy, similar to that proposed by Cui et al. (1999), is to increase the sample size at the second look from n_2 to $N_2 = 2n_1(\mu_1/\hat{\mu})^2$, provided $0 < \hat{\mu} < \mu_1$. This corresponds to the sample size necessary to detect the alternative $\mu_1 = \hat{\mu}$ with the desired 90% power. As a practical matter N_2 will be taken as a multiple of n_1 , say $N_2 = jn_1$ for $j = 2, \dots, K$, where K may be selected by budgetary or administrative criteria in order to set an upper bound on the magnitude of N_2 . To be specific, we define

$$j = \max[2, \min\{\lceil 2(\mu_1/\hat{\mu})^2 \rceil, K\}],$$

where $\lceil \cdot \rceil$ denotes rounding up to the nearest integer.

The above adaptive design corresponds to partitioning the interval $[l_1, u_1]$ into $K - 1$ mutually exclusive, collectively exhaustive subintervals Γ_j ($j = 2, \dots, K$), where $\Gamma_2 = [\mu_1 n_1, u_1]$,

$$\Gamma_j = [\mu_1 n_1 \sqrt{(2/j)}, \mu_1 n_1 \sqrt{\{2/(j-1)\}}] \quad (j = 3, \dots, K-1)$$

and $\Gamma_K = [l_1, \mu_1 n_1 \sqrt{\{2/(K-1)\}}]$. If $S_1 \in \Gamma_j$ the study continues to a second look with sample size $N_2 = jn_1$. The test statistic at the second look is modified from S_2 to T_j as defined by (1) so as to preserve the type I error, and the null hypothesis is rejected if $T_j > u_2$ and is accepted if $T_j \leq u_2$.

We now investigate the properties of this two-look adaptive design for $K = 10$ or a maximum sample size of $n = 250$. Either the study is terminated at the first look after $n_1 = 25$ patient pairs, or it continues to a second and final look, with its sample size augmented to any one of the nine possible values, $n_j = 25j$ ($j = 2, \dots, 10$), depending on the subinterval Γ_j into which S_1 falls. Since the Γ_j are intervals, the probabilities of falling into the rejection region $\mathcal{R}_j = (S_1 \in \Gamma_j, T_j > u_2)$ and the acceptance region $\mathcal{A}_j = (S_1 \in \Gamma_j, T_j \leq u_2)$ can be computed easily using bivariate normal distribution functions which in turn can be used to compute the (α, θ) -spending functions under the null hypothesis as well as the cumulative probabilities of rejection and acceptance by the monitoring time point j for various alternatives. The (α, θ) -spending functions for this design are displayed in Table 1.

We can construct a ten-look group-sequential test based on the likelihood ratio test statistic with $n = 250$, and the j th look at $n_j = 25j$ patient pairs. According to Theorem 1, if we base our early stopping criteria on upper and lower stopping boundaries that are derived from the (α, θ) -spending functions displayed in Table 1, in the manner described

Table 1. The (α, θ) functions for adaptive test, derived from 100 000 simulations

j	$\alpha(j)$	$\theta(j)$	j	$\alpha(j)$	$\theta(j)$	j	$\alpha(j)$	$\theta(j)$
0	0.0000	0.0000	4	0.0136	0.7054	8	0.0197	0.7716
1	0.0028	0.6693	5	0.0160	0.7236	9	0.0205	0.7849
2	0.0056	0.6736	6	0.0176	0.7410	10	0.0250	0.9750
3	0.0103	0.6876	7	0.0188	0.7570			

in the concluding paragraph of § 2, then the group-sequential test based on the likelihood ratio test statistic will be uniformly better than the adaptive test. This is illustrated by Fig. 1 in which the cumulative probabilities of rejection and acceptance by the monitoring time point j are plotted against the information fraction $n_j/250$ for various choices of μ . The solid line represents the test based on the likelihood ratio test statistic while the dotted line represents the adaptive test. When $\mu > 0$ we observe that the test based on the likelihood ratio test statistic has a uniformly higher probability of rejecting H_0 than the corresponding adaptive test. Also the test based on the likelihood ratio test statistic has a uniformly lower probability of accepting H_0 than the corresponding adaptive test. In contrast, when $\mu < 0$ the test based on the likelihood ratio test statistic has a uniformly lower probability of rejecting H_0 and a uniformly higher probability of accepting H_0 than the corresponding adaptive test.

5. CONCLUDING REMARKS

Statisticians are often involved in the design of clinical trials where there is no clear criterion for what constitutes a clinically important treatment difference. Thus, the idea that there exists a design where a trial is started based on some ‘rough’ guess of an anticipated treatment difference but allows the option of adaptively changing the design using the emerging treatment difference has a great deal of appeal. This is why we believe that there has been so much interest lately in such designs. However, some caution and further study is needed before the consideration of such designs. As we demonstrate in this paper, such strategies are inefficient. For any adaptive design, one can always construct a standard group-sequential test based on the sequential likelihood ratio test statistic that, for any parameter value in the space of alternatives, will reject the null hypothesis earlier with higher probability, and, for any parameter value not in the space of alternatives, will accept the null hypothesis earlier with higher probability.

It is important to point out that in our definition of optimality we do not consider the number of interim analyses that actually have to be conducted. As we discussed previously, the dominating group-sequential test must allow the possibility of having the study monitored at any time point where the adaptive design might monitor. Thus, in the example of § 4, the adaptive design would monitor the data at most twice but at 10 possible time points, whereas the dominating group-sequential test might monitor up to 10 times at the same time points. We have found in general that a likelihood ratio based sequential design which monitors often, even after every observation, and a likelihood ratio based group-sequential design which monitors much less frequently, say 5–10 times, with similar (α, θ) -spending functions have similar properties. The minimal number of tests needed before a standard group-sequential design might dominate over an adaptive design might be an interesting issue to pursue further.

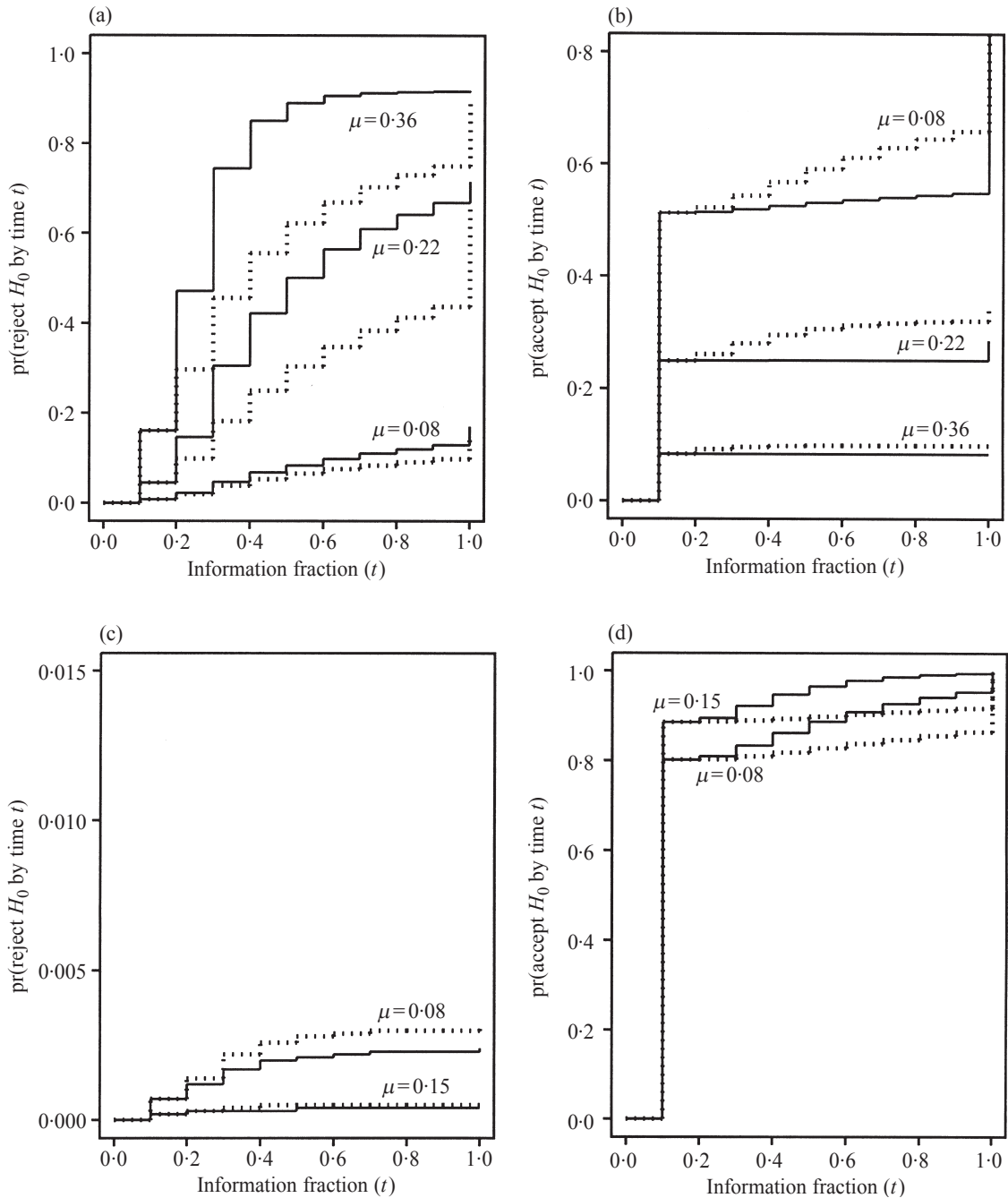


Fig. 1. Early rejection and acceptance probabilities for standard (shown by solid lines) and adaptive (dotted lines) designs. (a) $\text{pr}(\text{reject } H_0 \text{ by time } t)$ for positive μ ; (b) $\text{pr}(\text{accept } H_0 \text{ by time } t)$ for positive μ ; (c) $\text{pr}(\text{reject } H_0 \text{ by time } t)$ for negative μ ; (d) $\text{pr}(\text{accept } H_0 \text{ by time } t)$ for negative μ .

ACKNOWLEDGEMENT

This research was supported by grants from the National Institute of Allergy and Infectious Disease and the National Cancer Institute.

APPENDIX

Proof of Steps 1–3 in Theorem 1

Proof of Step 1. Assume that (7) holds. Note that $\bar{L}_{j+1}I_{\bar{\mathcal{R}}_j^{\text{LR}}} = \bar{L}_jI_{\bar{\mathcal{R}}_j^{\text{LR}}}L_{j+1}$ and $\bar{L}_{j+1}I_{\bar{\mathcal{A}}_j} = \bar{L}_jI_{\bar{\mathcal{A}}_j}L_{j+1}$. Denote the density of the random variable L_{j+1} by $f_{j+1}(x)$. Since L_{j+1} is independent of $\bar{L}_jI_{\bar{\mathcal{R}}_j^{\text{LR}}}$ and $\bar{L}_jI_{\bar{\mathcal{A}}_j}$, this implies that

$$P_0(\bar{L}_{j+1}I_{\bar{\mathcal{R}}_j^{\text{LR}}} > t) = \int P_0(\bar{L}_jI_{\bar{\mathcal{R}}_j^{\text{LR}}} > t|x)f_{j+1}(x) dx. \quad (\text{A1})$$

By (7), $P_0(\bar{L}_jI_{\bar{\mathcal{R}}_j^{\text{LR}}} > t|x) \geq P_0(\bar{L}_jI_{\bar{\mathcal{A}}_j} > t|x)$ for all $t, x > 0$. Therefore, by (A1),

$$P_0(\bar{L}_{j+1}I_{\bar{\mathcal{R}}_j^{\text{LR}}} > t) \geq \int P_0(\bar{L}_jI_{\bar{\mathcal{A}}_j} > t|x)f_{j+1}(x) dx = P_0(\bar{L}_{j+1}I_{\bar{\mathcal{A}}_j} > t).$$

An analogous proof can be used to show that

$$P_0(\bar{L}_{j+1}I_{\bar{\mathcal{A}}_j^{\text{LR}}} > t) \leq P_0(\bar{L}_{j+1}I_{\bar{\mathcal{R}}_j} > t).$$

Proof of Step 2. Assume that (9) holds. We need to prove that, for all $t > 0$,

$$P_0(\bar{L}_{j+1}I_{\bar{\mathcal{R}}_{j+1}^{\text{LR}}} > t) \geq P_0(\bar{L}_{j+1}I_{\bar{\mathcal{R}}_{j+1}^*} > t),$$

or equivalently that

$$P_0(\bar{\mathcal{R}}_{j+1}^{\text{LR}} \cap \bar{L}_{j+1} > t) \geq P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t). \quad (\text{A2})$$

By definition

$$\bar{\mathcal{R}}_{j+1}^{\text{LR}} = \bar{\mathcal{R}}_j^{\text{LR}} \cup (\mathcal{C}_j^{\text{LR}} \cap \bar{L}_{j+1} > u_{j+1}^{\text{LR}}), \quad \bar{\mathcal{R}}_{j+1}^* = \bar{\mathcal{R}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*),$$

which are constructed in such a way that

$$P_0(\bar{\mathcal{R}}_{j+1}^{\text{LR}}) = P_0(\bar{\mathcal{R}}_{j+1}^*) = \alpha_{j+1}.$$

Noting that $\bar{\mathcal{A}}_j = (\bar{\mathcal{R}}_j \cup \mathcal{C}_j)^c$, and using straightforward set operations, we obtain

$$P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) = \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t) \quad (t \leq u_{j+1}^*), \quad (\text{A3})$$

$$P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) \leq \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t) \quad (\text{for all } t), \quad (\text{A4})$$

$$P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) = P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j) \quad (t > u_{j+1}^*), \quad (\text{A5})$$

$$P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) \leq P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j) \quad (\text{for all } t). \quad (\text{A6})$$

Parallel results hold for $P_0(\bar{\mathcal{R}}_{j+1}^{\text{LR}} \cap \bar{L}_{j+1} > t)$ if we replace $\bar{\mathcal{R}}_j$, $\bar{\mathcal{A}}_j$ and u_{j+1}^* by $\bar{\mathcal{R}}_j^{\text{LR}}$, $\bar{\mathcal{A}}_j^{\text{LR}}$ and u_{j+1}^{LR} , respectively, on the right-hand side of (A3)–(A6).

By assumption (9),

$$P_0(\bar{\mathcal{R}}_j^{\text{LR}} \cap \bar{L}_{j+1} < t) \leq P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t), \quad (\text{A7})$$

$$P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j^{\text{LR}}) \leq P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j). \quad (\text{A8})$$

We consider two cases:

(i) for $t \leq u_{j+1}^{\text{LR}}$, we have

$$P_0(\bar{\mathcal{R}}_{j+1}^{\text{LR}} \cap \bar{L}_{j+1} > t) = \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j^{\text{LR}} \cap \bar{L}_{j+1} < t) \quad (\text{by (A3)})$$

$$\geq \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t) \quad (\text{by (A7)})$$

$$\geq P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) \quad (\text{by (A4)});$$

(ii) for $t > u_{j+1}^{\text{LR}}$, we have

$$\begin{aligned}
P_0(\bar{\mathcal{R}}_{j+1}^{\text{LR}} \cap \bar{L}_{j+1} > t) &= P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j^{\text{LR}}) \quad (\text{by (A5)}) \\
&\geq P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j) \quad (\text{by (A8)}) \\
&\geq P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) \quad (\text{by (A6)}).
\end{aligned}$$

Note that the (A3) and (A5) cited here are the parallel versions for $\bar{\mathcal{R}}_{j+1}^{\text{LR}}$.

Similar arguments can be used to show that $P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^{\text{LR}}} > t) \leq P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^*} > t)$.

Proof of Step 3. Note that

$$P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^*} > t) = P_0[\bar{L}_{j+1} > t \cap \{\bar{\mathcal{R}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*)\}]. \quad (\text{A9})$$

By construction,

$$P_0(\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*) = P_0(\mathcal{R}_{j+1}) = \alpha_{j+1} - \alpha_j, \quad (\text{A10})$$

which implies that

$$P_0(\mathcal{R}_{j+1} \cap \bar{L}_{j+1} > t) \leq \alpha_{j+1} - \alpha_j. \quad (\text{A11})$$

By (A9), for $t \leq u_{j+1}^*$ we obtain

$$\begin{aligned}
P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^*} > t) &= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\bar{L}_{j+1} > u_{j+1}^* \cap \mathcal{C}_j) \\
&= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + (\alpha_{j+1} - \alpha_j) \quad (\text{by (A10)}) \\
&\geq P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\mathcal{R}_{j+1} \cap \bar{L}_{j+1} > t) \quad (\text{by (A11)}) \\
&= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_{j+1}) = P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}} > t),
\end{aligned}$$

and, for $t > u_{j+1}^*$,

$$\begin{aligned}
P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^*} > t) &= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\bar{L}_{j+1} > t \cap \mathcal{C}_j) \\
&\geq P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\bar{L}_{j+1} > t \cap \mathcal{R}_{j+1}), \quad (\text{A12}) \\
&= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_{j+1}) = P_0(\bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}} > t).
\end{aligned}$$

Inequality (A12) follows because \mathcal{R}_{j+1} is contained in \mathcal{C}_j . Thus we have proved the first inequality of (11). The second inequality of (11) can be proved in a similar fashion thus completing the proof of Step 3.

REFERENCES

- ARMITAGE, P., MCPHERSON, C. K. & ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. R. Statist. Soc. A* **132**, 235–44.
- CUI, L., HUNG, H. M. J. & WANG, S-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–7.
- EAST. (2000). Software for Group Sequential Inference. Cambridge, MA: Cytel Software Corporation.
- LAN, K. K. G. & DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–63.
- LEHMACHER, W. & WASSMER, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–90.
- PAMPALLONA, S. & TSIATIS, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statist. Plan. Inference* **42**, 19–35.
- POSCH, M. & BAUER, P. (1999). Adaptive two stage designs and the conditional error function. *Biomet. J.* **41**, 689–96.
- SCHARFSTEIN, D. O., TSIATIS, A. A. & ROBINS, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *J. Am. Statist. Assoc.* **92**, 1342–50.
- SHEN, Y. & FISHER, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190–7.

[Received February 2002. Revised July 2002]