

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, Cambridge, MA

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 likelihood ratio 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

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, 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, Hung & Wang (1999), Posch & Bauer (1999), Lehman & 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 likelihood ratio 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 generalization of the Neyman-Pearson theorem to group-sequential tests. We also illustrate this numerically using some adaptive designs that have been recently advocated.

2. NOTATION FOR SEQUENTIAL TESTS

Suppose that decisions to stop a trial, either to reject or accept the null hypoth-

esis, 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 to either reject or 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: 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 randomized 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 points where interim analyses may be conducted are after $n_1 < n_2 < \dots < n_K = n$ observations. Thus, in this example, letting $n_0 = 0$, $X_j = (Z_{n_{j-1}+1}, \dots, Z_{n_j})$ and $Y_j = (Z_1, \dots, Z_{n_j})$, $j = 1, \dots, K$, the density of X_j ,

under the null and alternative hypotheses, is 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 group-sequential test will reject H_0 at the first time point j where $S_j = \sum_{l=1}^j Z_l$ is sufficiently large, say, $S_j > u_j$, or accept the null hypothesis at the first time point j where S_j is sufficiently small, say, $S_j < l_j$.

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, see Scharfstein, Tsiatis & Robins (1997). We shall return to this example for illustration throughout this paper.

Any level- α group-sequential test, using an adaptive design or not, can be represented by the sequence of rejection and acceptance rules $\{(R_j, A_j), j = 1, \dots, K\}$, where (R_j, A_j) are Y_j -measurable sets, and R_j and A_j are disjoint and have the interpretation that if a study was not stopped to reject or accept the null hypothesis by the j th time point, then the null hypothesis will be rejected at the j th time point if R_j occurs and accepted if A_j occurs. At the final time point, $A_K = (R_K)^C$, where $(\cdot)^C$ denotes the complement of a set. In the example, $R_j = (S_j > u_j)$ and $A_j = (S_j < l_j)$. This sequence of rejection and acceptance rules is used to define a group-sequential test with rejection, acceptance and continuation regions defined recursively as follows:

$$\mathcal{R}_1 = R_1,$$

$$\mathcal{A}_1 = A_1,$$

$$\mathcal{C}_1 = (\mathcal{R}_1 \cup \mathcal{A}_1)^C.$$

For $j = 2, \dots, K - 1$

$$\begin{aligned}
\mathcal{R}_j &= \mathcal{C}_{j-1} \cap R_j, \\
\mathcal{A}_j &= \mathcal{C}_{j-1} \cap A_j. \\
\bar{\mathcal{R}}_j &= \bar{\mathcal{R}}_{j-1} \cup \mathcal{R}_j, \\
\bar{\mathcal{A}}_j &= \bar{\mathcal{A}}_{j-1} \cup \mathcal{A}_j, \\
\mathcal{C}_j &= (\bar{\mathcal{R}}_j \cup \bar{\mathcal{A}}_j)^C.
\end{aligned} \tag{1}$$

For $j = K$

$$\begin{aligned}
\mathcal{R}_K &= \mathcal{C}_{K-1} \cap R_K, \\
\mathcal{A}_K &= \mathcal{C}_{K-1} \cap A_K, \\
\bar{\mathcal{R}}_K &= \bar{\mathcal{R}}_{K-1} \cup \mathcal{R}_K, \\
\bar{\mathcal{A}}_K &= \bar{\mathcal{A}}_{K-1} \cup \mathcal{A}_K.
\end{aligned}$$

The events $(\mathcal{R}_j, \mathcal{A}_j)$ correspond to stopping and rejecting or accepting the null hypothesis, respectively, at the j th time point. The events $(\bar{\mathcal{R}}_j, \bar{\mathcal{A}}_j)$ correspond to stopping and rejecting or accepting the null hypothesis, respectively, at or before the j th time point for $j = 2, \dots, K$. The events $\mathcal{C}_j, j = 1, \dots, K - 1$ correspond to continuing the study beyond the j th time point. 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 non-decreasing 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 $P_0(\bar{A}_j) = \theta_j$. We will refer to the non-decreasing 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.

From now on, when we refer to the group-sequential test $\{(R_j, A_j), j = 1, \dots, K\}$ with a $\alpha(\theta)$ -spending function, we mean the test with rejection regions and acceptance regions induced by (1) with the corresponding α -spending and θ -spending function defined above. We now give a criterion for optimality of group-sequential tests.

Definition: Among all group-sequential tests $\{(R_j, A_j), j = 1, \dots, K\}$ with a specified $\alpha(\theta)$ -spending function, the test $\{(R_j^{opt}, A_j^{opt}), j = 1, \dots, K\}$ is optimal within this class if

$$P_1(\bar{\mathcal{R}}_j^{opt}) \geq P_1(\bar{\mathcal{R}}_j) \text{ for all } j = 1, \dots, K, \quad (2)$$

where $P_1(\cdot)$ denotes probability computed under the alternative hypothesis.

Remark: By the definition of optimality (2), the optimal test is not only the most powerful test among all group-sequential tests with a specified $\alpha(\theta)$ -spending function; that is, $P_1(\bar{\mathcal{R}}_K^{opt}) \geq P_1(\bar{\mathcal{R}}_K)$ overall, but also, under the alternative, has a greater chance of rejecting the null hypothesis before or at any decision time point. We emphasize 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 $\alpha(\theta)$ -spending function, the group-sequential test based on the likelihood ratio statistic 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 .

To avoid the complications involved with discrete distributions and randomization tests, we assume that the likelihood ratio 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 likelihood ratio statistic, with a prespecified $\alpha(\theta)$ -spending function, is defined as $\{(R_j^{LR}, A_j^{LR}), j = 1, \dots, K\}$, where

$$R_j^{LR} = (\bar{L}_j > u_j^{LR}) \text{ and} \quad (3)$$

$$A_j^{LR} = (\bar{L}_j < l_j^{LR}). \quad (4)$$

The upper and lower boundaries u_j^{LR}, l_j^{LR} , $j = 1, \dots, K$ are constructed so that

$$P_0(\mathcal{R}_1^{LR}) = P_0(\bar{L}_1 > u_1^{LR}) = \alpha_1, \quad P_0(\mathcal{A}_1^{LR}) = P_0(\bar{L}_1 < l_1^{LR}) = \theta_1, \quad (5)$$

and for $j = 2, \dots, K$

$$P_0(\mathcal{R}_j^{LR}) = P_0(l_1^{LR} \leq \bar{L}_1 \leq u_1^{LR}, \dots, l_{j-1}^{LR} \leq \bar{L}_{j-1} \leq u_{j-1}^{LR}, \bar{L}_j > u_j^{LR}) = \alpha_j - \alpha_{j-1},$$

$$P_0(\mathcal{A}_j^{LR}) = P_0(l_1^{LR} \leq \bar{L}_1 \leq u_1^{LR}, \dots, l_{j-1}^{LR} \leq \bar{L}_{j-1} \leq u_{j-1}^{LR}, \bar{L}_j < l_j^{LR}) = \theta_j - \theta_{j-1}.$$

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

Note: Because the likelihood ratio statistics \bar{L}_j , $j = 1, \dots, K$ are assumed absolutely continuous, the boundaries (u_j^{LR}, l_j^{LR}) , $j = 1, \dots, K$ are uniquely defined and can be derived recursively using algorithms such as that in Armitage, McPherson & Rowe (1969).

3. OPTIMALITY OF THE GROUP-SEQUENTIAL LIKELIHOOD RATIO TEST

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 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

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

Consequently, the optimal group-sequential test with a specified $\alpha(\theta)$ -spending function must satisfy

$$E_0(\bar{L}_j I_{\bar{\mathcal{R}}_j^{opt}}) \geq E_0(\bar{L}_j I_{\bar{\mathcal{R}}_j}), \quad j = 1, \dots, K. \quad (6)$$

To verify (6), it suffices to show that the random variable $\bar{L}_j I_{\bar{\mathcal{R}}_j^{opt}}$ 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 ,

$$P_0(U > t) \geq P_0(W > t). \quad (7)$$

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

Therefore, among all group-sequential tests $\{(R_j, A_j), j = 1, \dots, K\}$ with a specified $\alpha(\theta)$ -spending function, the test $\{(R_j^{opt}, A_j^{opt}), j = 1, \dots, K\}$ is optimal if

$$\bar{L}_j I_{\bar{\mathcal{R}}_j^{opt}} \succeq \bar{L}_j I_{\bar{\mathcal{R}}_j}, \quad j = 1, \dots, K. \quad (8)$$

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

Theorem 1.: The optimal group-sequential test with a specified $\alpha(\theta)$ -spending function is the group-sequential likelihood ratio test $\{(R_j^{LR}, A_j^{LR}), j = 1, \dots, K\}$ defined by (4) and (5).

Proof. The theorem will be proved by verifying (8) using induction. That is, we will show that for all group-sequential tests $\{(R_j, A_j), j = 1, \dots, K\}$ with a specified $\alpha(\theta)$ -spending function, the following two conditions hold:

$$(i) \quad \bar{L}_1 I_{\bar{\mathcal{R}}_1^{LR}} \succeq \bar{L}_1 I_{\bar{\mathcal{R}}_1}, \quad \bar{L}_1 I_{\bar{\mathcal{A}}_1^{LR}} \preceq \bar{L}_1 I_{\bar{\mathcal{A}}_1} \quad (9)$$

and

$$(ii) \quad \bar{L}_j I_{\bar{\mathcal{R}}_j^{LR}} \succeq \bar{L}_j I_{\bar{\mathcal{R}}_j}, \quad \bar{L}_j I_{\bar{\mathcal{A}}_j^{LR}} \preceq \bar{L}_j I_{\bar{\mathcal{A}}_j} \quad (10)$$

implies that

$$\bar{L}_{j+1} I_{\bar{\mathcal{R}}_{j+1}^{LR}} \succeq \bar{L}_{j+1} I_{\bar{\mathcal{R}}_{j+1}}, \quad \bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}^{LR}} \preceq \bar{L}_{j+1} I_{\bar{\mathcal{A}}_{j+1}}. \quad (11)$$

We will prove condition (ii) in three steps, (a), (b) and (c).

(a) If (10) holds, then this implies

$$\bar{L}_{j+1} I_{\bar{\mathcal{R}}_j^{LR}} \succeq \bar{L}_{j+1} I_{\bar{\mathcal{R}}_j} \quad \text{and} \quad \bar{L}_{j+1} I_{\bar{\mathcal{A}}_j^{LR}} \preceq \bar{L}_{j+1} I_{\bar{\mathcal{A}}_j}. \quad (12)$$

(b) Define $\bar{\mathcal{R}}_{j+1}^* = \bar{\mathcal{R}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*)$ and $\bar{\mathcal{A}}_{j+1}^* = \bar{\mathcal{A}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} < l_{j+1}^*)$, 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 (12)

holds, this implies that

$$\bar{L}_{j+1}I_{\bar{\mathcal{R}}_{j+1}^{LR}} \succeq \bar{L}_{j+1}I_{\bar{\mathcal{R}}_{j+1}^*} \quad \text{and} \quad \bar{L}_{j+1}I_{\bar{\mathcal{A}}_{j+1}^{LR}} \preceq \bar{L}_{j+1}I_{\bar{\mathcal{A}}_{j+1}^*}. \quad (13)$$

(c) If (13) holds, then

$$\bar{L}_{j+1}I_{\bar{\mathcal{R}}_{j+1}^*} \succeq \bar{L}_{j+1}I_{\bar{\mathcal{R}}_{j+1}} \quad \text{and} \quad \bar{L}_{j+1}I_{\bar{\mathcal{A}}_{j+1}^*} \preceq \bar{L}_{j+1}I_{\bar{\mathcal{A}}_{j+1}}. \quad (14)$$

The proofs of (ii)a, (ii)b and (ii)c are given in the Appendix. As a consequence, we conclude that (10) implies (11). Also condition (i) of the theorem is a simple consequence of the proof of (ii)c. This completes the proof of Theorem 1.

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

Remark 2: Another way to view group-sequential tests with a specified $\alpha(\theta)$ -spending functions is as a partitioning of the sample space into Y_j -measurable rejection regions \mathcal{R}_j and Y_j -measurable acceptance regions \mathcal{A}_j , where $P_0(\mathcal{R}_j) = \alpha_j - \alpha_{j-1}$ and $P_0(\mathcal{A}_j) = \theta_j - \theta_{j-1}$, $j = 1, \dots, K$. 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 $\alpha(\theta)$ -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 likelihood ratio test with upper boundaries only, that is, where

$$\mathcal{R}_j = \bar{L}_1 \leq u_1^{LR}, \dots, \bar{L}_{j-1} \leq u_{j-1}^{LR}, \bar{L}_j > u_j^{LR}, \quad j = 1, \dots, K.$$

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 $\alpha(\theta)$ -spending function restrictions. This test may be desirable if we are only interested in stopping early if treatment differences occur and are willing to continue collecting data otherwise.

Example: Returning to the example of §2, the theorem states that, among group-sequential tests with a specified $\alpha(\theta)$ -spending functions, the group-sequential likelihood ratio test which has continuation regions

$$\mathcal{C}_j^{LR} = (l_1^{LR} \leq \bar{L}_1 \leq u_1^{LR}, \dots, l_j^{LR} \leq \bar{L}_j \leq u_j^{LR}),$$

rejection regions

$$\mathcal{R}_j^{LR} = \mathcal{C}_{j-1}^{LR} \cap \bar{L}_j > u_j^{LR},$$

and acceptance regions

$$\mathcal{A}_j^{LR} = \mathcal{C}_{j-1}^{LR} \cap \bar{L}_j < l_j^{LR},$$

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 yields the likelihood ratio 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 likelihood ratio test is equivalent to the test where the j th continuation region is defined by

$$\mathcal{C}_j^{LR} = (l_1 \leq S_1 \leq u_1, \dots, l_j \leq S_j \leq u_j),$$

rejects H_0 at time j if

$$\mathcal{R}_j^{LR} = \mathcal{C}_{j-1}^{LR} \cap S_j > u_j,$$

and accepts H_0 if

$$\mathcal{A}_j^{LR} = \mathcal{C}_{j-1}^{LR} \cap S_j < l_j.$$

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

$$P_0(\mathcal{R}_j^{LR}) = \alpha_j - \alpha_{j-1} \text{ and } P_0(\mathcal{A}_j^{LR}) = \theta_j - \theta_{j-1}. \quad (15)$$

This is the usual group-sequential one-sided test with upper and lower boundaries with a specified $\alpha(\theta)$ -spending function. Because 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 $\alpha(\theta)$ -spending function, the group-sequential likelihood ratio test defined above has the property that

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

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 $\alpha(\theta)$ -spending function uniformly for all $j = 1, \dots, K$.

4. NUMERICAL COMPARISONS

Any level- α adaptive sequential test has the properties that there is some maximum sample size N_{\max} , a random variable T , corresponding to the patient accrual when the study was stopped, expressed as a proportion of N_{\max} , and a random variable D assuming a value 1 if H_0 is rejected; 0 otherwise. So, for example, if $T = t, D = 1$,

then the study is stopped after tN_{\max} observations and H_0 is rejected. A level- α test has the property $P_0(T \leq 1, D = 1) = \alpha$. If the random variable TN_{\max} has discrete mass at values $n_1 < n_2 < \dots < n_K = N_{\max}$, then these correspond to the K decision time points described previously. The corresponding α -spending function for this test is $\alpha_j = P_0(T \leq n_j/N_{\max}, D = 1)$ and the θ -spending function is $\theta_j = P_0(T \leq n_j/N_{\max}, D = 0)$ for $j = 1, \dots, K$. According to Theorem 1., the group-sequential likelihood ratio test, which monitors the data at the time points $n_1 < n_2 < \dots < n_K$ using the likelihood ratio statistic, with upper and lower boundaries constructed using (5) to have the same $\alpha(\theta)$ -spending as the adaptive test, is uniformly better than the adaptive test.

We illustrate with an adaptive group sequential test of the type proposed by Cui, Hung and Wang (1999). Consider again the example, first introduced in section 2, for comparing a new treatment to a placebo. Suppose the study is designed for 5% significance and 90% power with up to 4 looks and possible early stopping to either reject or accept H_0 . We select the rejection and acceptance rules $\{(R_j, A_j), j = 1, \dots, 4\}$ from the Pampallona and Tsiatis (1994) family with shape parameter 0. With this choice of boundaries one can show that N , the maximum number of pairs of patients one must commit to the study in order for a four-look one-sided group sequential test conducted at the 5% significance level to attain 90% power to detect an effect size of μ_1 , satisfies the relationship $N = 9.37/\mu_1^2$.

In a classical group sequential design the value of N would remain fixed for the duration of the study. We would plan to monitor the data four times, with the j th look being taken after $n_j = jN/4$ pairs of patients had been accrued and would terminate the study the first time that a stopping boundary was crossed. In contrast,

one of the adaptive designs proposed by Cui, Hung and Wang (1999) permits a one-time increase in N at a pre-specified interim look $L < 4$. In this design the maximum sample size could, in principle, be increased from N to $N(\mu_1/\hat{\mu})^2$ if, at the L th look, $0 < \hat{\mu} < \mu_1$. As a practical matter, however, budgetary and administrative constraints would automatically set an upper limit to the the maximum sample size regardless of the magnitude of the multiplicative factor $(\mu_1/\hat{\mu})^2$. Denote this practical upper limit by N_{\max} . Then the maximum sample size of the adaptive design is increased at look L from N to

$$M = \min \left\{ N_{\max}, \max \left\{ N, \left(\frac{\mu_1}{\hat{\mu}} \right)^2 N I_{[0 < \hat{\mu} < \mu_1]} \right\} \right\} \quad (16)$$

where $I_{[\cdot]}$ is the indicator function. The remaining $4 - L$ looks are spaced equally with $(M - LN/4)/(4 - L)$ additional patient pairs being accrued between successive looks, and the stopping boundaries initially proposed continue to be used at looks $L + 1, \dots, 4$. Thus, in this 4-look adaptive design, if $j \leq L$ the j th look is taken after accruing $n_j = jN/4$ patient pairs, while if $j > L$ the j th look is taken after accruing

$$\tilde{n}_j = n_L + \frac{(j - L)(M - LN/4)}{4 - L} \quad (17)$$

patient pairs. Cui, Hung and Wang (1999) prove that, despite the possibility of an increase in maximum sample size from N to M at look L , the overall type-1 error will be preserved without any change in the stopping boundaries provided the test statistic adopted for the remaining $4 - L$ looks is adjusted from $Q_j = (\tilde{n}_j)^{-1/2}(Z_1 + Z_2 + \dots + Z_{\tilde{n}_j})$ to the weighted sum

$$\tilde{Q}_j = \sqrt{\frac{n_L}{n_j}} \sum_{l=1}^{n_L} \frac{Z_l}{\sqrt{n_L}} + \sqrt{\frac{n_j - n_L}{n_j}} \sum_{l=N_L+1}^{\tilde{n}_j} \frac{Z_l}{\sqrt{\tilde{n}_j - n_L}} \quad (18)$$

for $j = L + 1, \dots, 4$.

We will compare the performance of the above four-look adaptive group sequential test in which $L = 1$, $N = 50$ and $N_{\max} = 200$ with that of a non-adaptive group sequential likelihood ratio test having the same value for N_{\max} . (Note that $N = 50$ corresponds to an effect size $\mu_1 = 0.43$.) The first step is to obtain the $\alpha(\theta)$ function for this adaptive test. We do this by simulating the test 100,000 times under the null hypothesis $\mu = 0$. We thereby generate 100,000 ordered pairs (T, D) . Using these ordered pairs, we construct two empirical cumulative density function; $\alpha(t) = \Pr(T \leq t, D = 1)$ and $\theta(t) = \Pr(T \leq t, D = 0)$. These two empirical CDF's are displayed in Table 1. For convenience we have constructed these functions by computing cumulative probabilities at only ten distinct values of t . Our criterion for selecting each support point t is that either $\alpha(t)$ increases by 0.005, relative to the previous support point, or else $\theta(t)$ increases by 0.1 relative to the previous support point.

Table 1: Alpha-Theta Functions for Adaptive Test (derived from 100,000 simulations)

t	$\alpha(t)$	$\theta(t)$	t	$\alpha(t)$	$\theta(t)$	t	$\alpha(t)$	$\theta(t)$
0.000	0.0000	0.0000	0.257	0.0206	0.6001	0.562	0.0394	0.8867
0.062	0.0014	0.1174	0.320	0.0259	0.6123	0.957	0.0442	0.9056
0.125	0.0109	0.5921	0.417	0.0307	0.6344	1.000	0.0529	0.0000
0.207	0.0156	0.5936	0.525	0.0352	0.6596			

We can construct a ten-look group sequential likelihood ratio test with $N_{\max} = 200$ and interim monitoring at the information fractions displayed in column 1 of Table 1. According to Theorem 1., if we base our early stopping criteria on upper

and lower stopping boundaries that are derived, as shown in equation (5), from the $\alpha(\theta)$ spending function displayed in Table 1, then the likelihood ratio test will be uniformly better than the adaptive test. This is illustrated by Figure 1 in which rejection and acceptance probabilities are plotted against the information fraction, t , for various choices of μ . The solid line represents the likelihood ratio test while the dotted line represents the adaptive test. When $\mu > 0$ we observe that the likelihood ratio test has a uniformly higher probability of rejecting H_0 than the corresponding adaptive test. Also the likelihood ratio test has a uniformly lower probability of accepting H_0 than the corresponding adaptive test. In contrast, when $\mu < 0$ the likelihood ratio test 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, 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 likelihood ratio 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.

Therefore, in our opinion, the issue should not be whether to use an adaptive

design or a standard (optimal) group-sequential test, but rather, which standard group-sequential test to use. Specifically, since a standard group-sequential test is determined by the choice of N_{\max} and the $\alpha(\theta)$ -spending function, the question should be how to choose these. One strategy, which we believe is consistent with the goals of an adaptive design, is for the statistician and collaborators to choose a range of possible treatment differences that includes a minimally acceptable treatment difference, as well as larger treatment differences that are believed to be plausible with the new treatment. One can then search among standard group-sequential tests, with specified power to detect the minimally acceptable difference, to find designs which will have high probability of rejecting H_0 as early as possible for the plausible range of treatment differences. Examples of such designs and searching strategies are given by Jennison (1987) and Eales & Jennison (1992).

Such a strategy is superior to one where a trial is started based on some anticipated plausible treatment difference and then the sample size is adaptively increased if the data suggest a smaller difference.

ACKNOWLEDGMENT

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

APPENDIX

Proof of (ii)a

Assume (10) holds. Note that $\bar{L}_{j+1}I_{\bar{\mathcal{R}}_j^{LR}} = \bar{L}_jI_{\bar{\mathcal{R}}_j^{LR}}L_{j+1}$ and $\bar{L}_{j+1}I_{\bar{\mathcal{R}}_j} = \bar{L}_jI_{\bar{\mathcal{R}}_j}L_{j+1}$. Denote the density of the random variable L_{j+1} by $f_{j+1}(x)$. Because L_{j+1} is inde-

pendent of $\bar{L}_j I_{\bar{\mathcal{R}}_j^{LR}}$ and $\bar{L}_j I_{\bar{\mathcal{R}}_j}$, this implies that

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

By (10), $P_0(\bar{L}_j I_{\bar{\mathcal{R}}_j^{LR}} > t/x) \geq P_0(\bar{L}_j I_{\bar{\mathcal{R}}_j} > t/x)$ for all $t, x > 0$. Therefore

$$(\text{A.1}) \geq \int P_0(\bar{L}_j I_{\bar{\mathcal{R}}_j} > t/x) f_{j+1}(x) dx \geq P_0(\bar{L}_j I_{\bar{\mathcal{R}}_j} > t).$$

An analogous proof can be used to show that

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

Proof of (ii)b

Assume (12) holds. We need to prove that for all $t > 0$,

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

or equivalently that

$$P_0(\bar{\mathcal{R}}_{j+1}^{LR} \cap \bar{L}_{j+1} > t) \geq P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t). \quad (\text{A.2})$$

By definition

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

and

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

which are constructed in such a way that

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

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

$$P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t) = \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t), \quad t < u_j^*, \quad (\text{A.3})$$

$$\leq \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t) \text{ for all } t, \quad (\text{A.4})$$

$$= P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j), \quad t > u_j^*, \quad (\text{A.5})$$

$$\leq P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j), \text{ for all } t. \quad (\text{A.6})$$

Similar results hold for $P_0(\bar{\mathcal{R}}_{j+1}^{LR} \cap \bar{L}_{j+1} > t)$.

By assumption (12),

$$P_0(\bar{\mathcal{R}}_j^{LR} \cap \bar{L}_{j+1} < t) \leq P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t), \quad (\text{A.7})$$

and

$$P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j^{LR}) \leq P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j). \quad (\text{A.8})$$

We consider two cases:

(i) $u_j^* \leq u_j^{LR}$:

For (i) and $t \leq u_j^{LR}$, we have

$$\begin{aligned} P_0(\bar{\mathcal{R}}_{j+1}^{LR} \cap \bar{L}_{j+1} > t) &= \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j^{LR} \cap \bar{L}_{j+1} < t) \text{ by (A.3)} \\ &\geq \alpha_{j+1} - P_0(\bar{\mathcal{R}}_j \cap \bar{L}_{j+1} < t) \text{ by (A.7)} \\ &\geq P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t). \text{ by (A.4)} \end{aligned}$$

(ii) $u_j^{LR} < u_j^*$:

For (ii) and $t > u_j^{LR}$, we have

$$\begin{aligned} P_0(\bar{\mathcal{R}}_{j+1}^{LR} \cap \bar{L}_{j+1} > t) &= P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j^{LR}), \text{ by (A.5)} \\ &\geq P_0(\bar{L}_{j+1} > t) - P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{A}}_j), \text{ by (A.8)} \\ &\geq P_0(\bar{\mathcal{R}}_{j+1}^* \cap \bar{L}_{j+1} > t). \text{ by (A.6)} \end{aligned}$$

To complete the proof of (A.2), we need to consider case (i) for $t > u_j^{LR}$ and case (ii) for $t \leq u_j^{LR}$. These, however, are proved exactly as for case (ii), $t > u_j^{LR}$ and case (i), $t \leq u_j^{LR}$.

Finally, a symmetric argument can be used to show that

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

Proof of (ii)c

Assume (13) holds; that is,

$$P_0[\bar{L}_{j+1} > t \cap \{\bar{\mathcal{R}}_j \cup (\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*)\}] \quad (\text{A.9})$$

$$\geq P_0\{\bar{L}_{j+1} > t \cap (\bar{\mathcal{R}}_j \cup \mathcal{R}_{j+1})\}. \quad (\text{A.10})$$

By construction

$$P_0(\mathcal{C}_j \cap \bar{L}_{j+1} > u_{j+1}^*) = P_0(\mathcal{R}_{j+1}) = \alpha_{j+1} - \alpha_j, \quad (\text{A.11})$$

which implies

$$P_0(\mathcal{R}_{j+1} \cap \bar{L}_{j+1} > t) \leq \alpha_{j+1} - \alpha_j. \quad (\text{A.12})$$

For $t \leq u_{j+1}^*$

$$\begin{aligned} (\text{A.9}) &= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\bar{L}_{j+1} > u_{j+1}^* \cap \mathcal{C}_j) \\ &= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + (\alpha_{j+1} - \alpha_j) \text{ by (A.11)} \\ &\geq P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\mathcal{R}_{j+1} \cap \bar{L}_{j+1} > t) \text{ by (A.12)} \\ &= (14). \end{aligned}$$

For $t > u_{j+1}^*$

$$\begin{aligned} (\text{A.9}) &= P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\bar{L}_{j+1} > t \cap \mathcal{C}_j) \\ &\geq P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_j) + P_0(\bar{L}_{j+1} > t \cap \bar{\mathcal{R}}_{j+1}), \quad (\text{A.13}) \\ &= (14). \end{aligned}$$

(A.13) follows because $\bar{\mathcal{R}}_{j+1}$ is contained in \mathcal{C}_j .

REFERENCES

- ARMITAGE, P., MCPHERSON, C.K., & ROWE, B.C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc., Series A* **132**, 235–244.
- CUI, L., HUNG, H.M.J. & WANG, S-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–857.
- EALLES, J.D., & JENNISON, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13–24.
- JENNISON, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika* **74**, 155–165.
- LAN, G.K.K., & DEMETS, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.
- LEHMACHER, W. & WASSMER, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.
- PAMPALLONA, S. & TSIATIS, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provisions for early stopping in favor of the null hypothesis. *J. Statist. Planning and Inference* **42**, 19–35.
- POSCH, M., & BAUER, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal* **41**, 689–696.
- SCHARFSTEIN, D.O., TSIATIS, A.A., & ROBINS, J.M. (1997). Semiparametric efficiency and its implications on the design and analysis of group-sequential studies. *J. Amer. Statist. Assoc.* **92**, 1342–1350.

SHEN, Y. & FISHER, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190–197.

Figure 1: Comparison of Likelihood Ratio (___) and Adaptive (---) Designs

