Published in final edited form as:

Seq Anal. 2012 ; 31(4): 441–457. doi:10.1080/07474946.2012.719433.

Efficient Adaptive Randomization and Stopping Rules in Multi-arm Clinical Trials for Testing a New Treatment

Tze Leung Lai and Olivia Yueh-Wen Liao

Department of Statistics, Stanford University, Stanford, California, USA

Abstract

Motivated by applications to confirmatory clinical trials for testing a new treatment against a placebo or active control when the new treatment has k possible treatment strategies (arms)—for example, k possible doses for a new drug—we develop an asymptotic theory for efficient outcome-adaptive randomization schemes and optimal stopping rules. Our approach consists of developing asymptotic lower bounds for the expected sample sizes from the k treatment arms and the control arm and using generalized sequential likelihood ratio procedures to achieve these bounds. Implementation details of our design and analysis and comparative simulation studies are also provided.

Keywords

Adaptive allocation; Information bound; Interim analysis; Internal pilot; Multi-armed bandit theory; Multiple treatment

1. INTRODUCTION

Although it is common practice to determine the dose or dosing regimen for a new drug, or the treatment strategy for a new treatment, in a phase II trial, Shun et al. (2008) have pointed out that there are difficulties in designing the phase II trial when there is “long-term follow-up, large sample size but lack of patient resources, and expensive study supplies.” They pointed out that “it may be necessary to evaluate a few doses in a phase III confirmatory study,” leading to a confirmatory trial with multiple treatment arms of the new treatment to compare against the control. In the case of $k = 2$ treatment arms, they proposed a ‘two-stage winner design’ to test the null hypothesis $H : \mu_1 = \mu_2 = 0$ versus the alternative hypothesis $\max(\mu_1, \mu_2) > 0$ where $\mu_i = \mu_i - \mu_0$, assuming the outcomes to be normally distributed with a common known variance and unknown means μ_0, μ_1 , and μ_2 for the control and the two treatment arms, respectively. The first stage of the design randomizes patients, with equal probabilities, to the three arms. The second stage chooses the treatment arm $\hat{i}$ with the larger sample mean at the end of the first stage and uses the sample mean difference between arm $\hat{i}$ and the control arm, divided by the standard error, as the final test statistic of H_0 .

In this article we consider the more general framework of k treatment arms and an exponential family with density function $f_{\theta}(x) = e^{\theta x - \psi(\theta)}$, with respect to a measure on the real line, for the outcomes of a treatment (or control) that has parameter θ . We consider the more general null hypothesis $H_0 : \max(\theta_1, \dots, \theta_k) \leq \theta_0$, where θ_0 is the parameter value of the control (arm 0) and θ_i is that of treatment arm $i = 1, \dots, k$. Unlike the approach of Shun et al. (2008), which does not consider the uncertainties in using the sample means to estimate the θ_i to choose the winning arm at the end of the first stage, we use insights from multi-armed bandit theory in Lai (1987) and optimal sequential selection of the best of k populations in Chan and Lai (2006) to come up with uncertainty adjustments via an adaptive randomization scheme. Moreover, unlike their fixed sample size procedure, we use efficient stopping rules that allow early stopping to either reject or accept H_0 . As in Lai and Shih (2004), efficiency is measured by both expected sample size and power, subject to the type I error constraint of falsely rejecting H_0 when it is true.

We develop the basic theory for asymptotically efficient stopping and adaptive randomization rules in the setting of fully sequential procedures, which require unblinding the data at every stage. In practice, for double-blind randomized clinical trials, unblinding the results is only carried out periodically prior to the meetings of the Data and Safety Monitoring Committee, resulting in group sequential trials. In Section 3 we show how the approach in Section 2 can be modified for the group sequential setting and provide implementation details, in particular on computational methods to maintain the prescribed type I error probability. A simulation study is given in Section 4 to show the advantage of the approach over other methods. Section 5 gives concluding remarks and discussion of related adaptive designs, reviewed recently by Dragalin (2010), of multi-arm phase II trials to identify the dose to be used in phase III testing of the new treatment.

2. ASYMPTOTICALLY EFFICIENT ALLOCATION AND STOPPING RULES

Randomized clinical trials are typically double-blind in the sense that patients and their physicians do not know the treatment to which the patient is randomly assigned; this is needed for the feasibility of patient accrual under informed consent. Therefore, patients are randomized to the $k + 1$ arms when there are k treatment strategies for the new treatment to be tested against the control arm. For efficiency, adaptive randomization schemes can be used, but the data-dependent probability, described in the protocol, of assigning to a treatment arm has to be bounded away from 0 and 1 to achieve informed consent in practice unless it has been eliminated from consideration. Specifically, the randomization probability p_{it} of assigning the patient to treatment arm i at time t is assumed to satisfy

$$p_{it} \geq \epsilon \quad \text{for some } 0 < \epsilon \leq 1/(k+1), \quad (2.1)$$

unless treatment arm i has been eliminated prior to t . Let T_i be the total sample size from arm i in the trial, $i = 0, \dots, k$. In this section we develop a test of $H_0 : \max(\theta_1, \dots, \theta_k) \leq \theta_0$ that has asymptotically optimal total expected sample size, subject to constraints on the type I error probability and type II error probabilities at certain alternatives. The test consists of a stopping rule, a treatment allocation rule via adaptive randomization, and sequentially updated test statistics. This structure was also described by Dragalin (2010, p. 446).

First consider the case $k = 1$ and known θ_0 , for which we can derive information bounds on the expected sample size. In this case, there is no need to sample from arm 0. Let $\tilde{\theta} > \theta_0$ be a given alternative and let $\theta_0 < \theta^* < \tilde{\theta}$ be such that $I(\theta^*, \theta_0) = I(\theta^*, \tilde{\theta})$, where

$$I(\theta, \lambda) = E_{\theta} \log [f_{\theta}(X) / f_{\lambda}(X)] = (\theta - \lambda) \psi'(\theta) - \{\psi(\theta) - \psi(\lambda)\} \quad (2.2)$$

is the Kullback-Leibler information number. Subject to the error probability constraints $P_{\theta_0}(\text{Reject } H_0) \leq \alpha$ and $P_{\tilde{\theta}}(\text{Accept } H_0) \leq \tilde{\alpha}$, it follows from Hoeffding (1960) lower bound that as $\alpha \rightarrow 0$ and $\tilde{\alpha} \rightarrow 0$,

$$E_{\theta} T_1 \geq \begin{cases} (1+o(1)) |\log \alpha| / I(\theta, \theta_0) & \text{if } \theta > \theta^* \\ (1+o(1)) |\log \tilde{\alpha}| / I(\theta, \tilde{\theta}) & \text{if } \theta < \theta^*. \end{cases} \quad (2.3)$$

Moreover, if $|\log \alpha| \sim |\log \tilde{\alpha}|$, then (2.3) can also be extended to the case $\theta = \theta^*$. In this case, the asymptotic lower bound (2.3) can be attained by a 2-SPRT (Lorden, 1976) when θ is specified or by a sequential generalized likelihood ratio (GLR) test that replaces θ (when it is unknown) in the 2-SPRT by sequentially updated maximum likelihood estimates; see Lai (2001, pp. 306–307).

We next consider the case $k = 1$ with unknown θ_0 . In this case, we need to sample also from arm 0 to learn θ_0 . Suppose we randomize equally between arms 0 and 1, as in typical clinical trials. Instead of using the natural parameter θ , it will be more convenient to reparameterize the exponential family by $\mu = \psi'(\theta)$, noting that ψ' is a smooth increasing function and that μ is the mean of f_{θ} . The Kullback-Leibler information number (2.2) will be denoted by $I(\mu_{\theta}, \mu_{\lambda})$ instead of $I(\theta, \lambda)$, and θ_0 and $\tilde{\theta}$ will be relabeled as μ_0 and $\tilde{\mu}$, in which $\tilde{\mu} = \mu_0 + \delta$ with $\delta > 0$ specified but μ_0 unknown. Analogous to the θ^* with $I(\theta^*, \theta_0) = I(\theta^*, \tilde{\theta})$ in the case of known θ_0 , we now define $\tilde{\mu}^* \in (\mu_0, \tilde{\mu})$ by

$$\mathcal{J}(\tilde{\mu}^*, \mu_0) = I(\tilde{\mu}^*, \tilde{\mu}), \quad \text{where } \mathcal{J}(\mu, v) = I(\mu, (\mu+v)/2) + I(v, (\mu+v)/2). \quad (2.4)$$

In Section 2.4, it will be shown that subject to the probability constraints

$$P_{\mu_1=\mu_0}(\text{Reject } H_0) \leq \alpha, \quad P_{\mu_1=\tilde{\mu}}(\text{Accept } H_0) \leq \tilde{\alpha}, \quad (2.5)$$

the following analog of (2.3) still holds as $\alpha \rightarrow 0$ and $\tilde{\alpha} \rightarrow 0$:

$$E_{\mu_0, \mu_1} T_1 \geq \begin{cases} (1+o(1)) |\log \alpha| / \mathcal{J}(\mu_1, \mu_0) & \text{if } \mu_1 > \tilde{\mu}^* \\ (1+o(1)) |\log \tilde{\alpha}| / I(\mu_1, \tilde{\mu}) & \text{if } \mu_1 < \tilde{\mu}^*. \end{cases} \quad (2.6)$$

Moreover, $T_0 \sim T_1$ since subjects are randomized to either arm with equal probability.

Assuming that $|\log \alpha| \sim |\log \tilde{\alpha}|$, the asymptotic lower bound (2.6) is attained by a sequential GLR test; see Section 2.4. In (2.5) and the sequel, a parameter constraint in the subscript of P is used to denote P_{μ_0, μ_1} such that (μ_0, μ_1) satisfies the constraint; this simplifies the

notation substantially. In particular, the second inequality in (2.5) means that $P_{\mu_0, \mu_1}(\text{Accept } H_0) \geq \tilde{\alpha}$ for all (μ_0, μ_1) satisfying $\mu_1 = \tilde{\mu} (= \mu_0 + \delta)$.

For $k > 1$, we extend the probability constraints (2.5) to

$$\begin{aligned} P_{\max(\mu_1, \dots, \mu_k) = \mu_0}(\text{Reject } H_0) &\leq \alpha, \quad P_{\max(\mu_1, \dots, \mu_k) = \tilde{\mu}}(\text{Accept } H_0) \leq \tilde{\alpha}, \\ P_{\mu_i = \tilde{\mu}}(\text{Eliminate arm } i \text{ before trial ends}) &\leq \tilde{\alpha} \quad \text{for } i=1, \dots, k. \end{aligned} \quad (2.7)$$

The third constraint in (2.7) is added because we allow elimination of a treatment arm during the trial but do not want to eliminate arms whose means exceed μ_0 by at least δ since these arms carry substantial information against the null hypothesis. One can regard the second and third constraints in (2.7) as safeguards on the power of the test at certain prescribed alternatives. We first assume $k = 2$ in Sections 2.1 and 2.2 because the key ideas can be described by considering relatively few exhaustive cases. We then consider general k in Section 2.3 and provide proofs in Section 2.4.

2.1. Asymptotic Lower Bounds When $k = 2$

If $\mu_1 = \mu_2$, arms 1 and 2 are indistinguishable and can therefore be combined into a single arm. In this case, (2.6) still holds with T_1 replaced by $T_1 + T_2$, and $T_0 \sim T_1 + T_2$ when subjects are randomized to arm 0 with probability $1/2$. It remains to consider $\mu_1 > \mu_2$, and we assume $\mu_1 > \mu_2$ for definiteness and divide into two cases.

Case 1— $\mu_2 = \mu_0$. In this case, only arm 1 carries information about H_0 . Therefore, the lower bound (2.6) still holds for T_1 . The constraint (2.1) suggests the optimal allocation proportion $1 : 1 : \epsilon$ for arms 0, 1, 2 before arm 2 is eliminated. Moreover, it will be shown in Section 2.4 that the third constraint in (2.7) implies that as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$,

$$\#_2 \geq (1 + o_p(1)) |\log \alpha| / I(\mu_2, \tilde{\mu}) \quad \text{under } P_{\mu_0, \mu_1, \mu_2}, \quad (2.8)$$

where $\#_2$ denotes the number of observations from arm 2 before it is eliminated. Since $T_2 \sim \min\{\#_2, (\epsilon + o_p(1))T_1\}$, it then follows from (2.6) and (2.8) that in this case,

$$\frac{E_{\mu_0, \mu_1, \mu_2}(T_2)}{|\log \alpha|} \geq (1 + o(1)) \min \left\{ \frac{1}{I(\mu_2, \tilde{\mu})}, \epsilon \min \left(\frac{1}{\mathcal{J}(\mu_1, \mu_0)}, \frac{1}{\mathbf{I}(\mu_1, \tilde{\mu})} \right) \right\}. \quad (2.9)$$

Moreover, $T_0 = 1 + o_p(1))T_1$.

Case 2— $\mu_2 > \mu_0$. In this case, both arms 1 and 2 carry information about the null hypothesis $H_0 : \max(\mu_1, \mu_2) = \mu_0$. Moreover, one does not expect to eliminate arm 2 before rejecting H_0 since $\mu_2 > \mu_0$. Let $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and define

$$\mathcal{J}_\epsilon(\boldsymbol{\mu}, \mathbf{v}) = I\left(\mu_1, \frac{\mu_1 + \epsilon\mu_2 + \mathbf{v}}{2 + \epsilon}\right) + \epsilon I\left(\mu_2, \frac{\mu_1 + \epsilon\mu_2 + \mathbf{v}}{2 + \epsilon}\right) + \mathbf{I}\left(\mathbf{v}, \frac{\mu_1 + \epsilon\mu_2 + \mathbf{v}}{2 + \epsilon}\right), \quad (2.10)$$

which reflects the sampling proportion $1 : \varepsilon : 1$ from arms 1, 2, and 0. It is shown in Section 2.4 that in this case,

$$\frac{E_{\mu_0, \mu_1, \mu_2}(T_1)}{|\log \alpha|} \geq (1+o(1)) \min \left\{ \frac{1}{\mathcal{J}_\varepsilon(\boldsymbol{\mu}, \mu_0)}, \frac{1}{I(\mu_1, \tilde{\mu})} \right\} \quad (2.11)$$

as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$. Moreover, $T_0 \sim T_1$ and $T_2 \sim \varepsilon T_1$.

2.2. Asymptotically Efficient Test When $k = 2$

The lower bounds in the preceding section depend on whether $\mu_2 = \mu_0$ or not. To attain these bounds when the μ_i are actually unknown, we use an adaptive approach that replaces the unknown μ_i by the sample mean $\hat{\mu}_{i,t(i)}$, $i = 0, 1, 2$, at every stage t , where $t(i)$ denotes the total number of observations from arm i up to stage t . As pointed out by Lai (1987), uncertainty adjustments must be made in replacing the unknown parameters by their sequential estimates for an adaptive method to be asymptotically efficient as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$. In the present case, such adjustment is already built in via the requirement (2.1) of “forced sampling” from apparently inferior arms. The idea of forced sampling was first introduced by Robbins (1952) for two-armed bandits and ensures the consistency of the estimates.

Let $\hat{\mu}_t^0 = \hat{\mu}_{0,t(0)}$, $\hat{\mu}_t^{\mathbf{I}} = \max(\hat{\mu}_{1,t(1)}, \hat{\mu}_{2,t(2)})$ and $\hat{\mu}_t^{\mathbf{II}} = \min(\hat{\mu}_{1,t(1)}, \hat{\mu}_{2,t(2)})$, in which we let $\mathbf{I}$ or $\mathbf{II}$ denote the arm that attains the maximum or minimum; if there are ties, just label one as $\mathbf{I}$ and the other as $\mathbf{II}$. Note that $\mathbf{I}$ and $\mathbf{II}$ in fact depend on t , which we drop from the subscript for notational simplicity. The statistics we use to test H_0 sequentially are the GLR statistics e^{ℓ_t} , where

$$\begin{aligned} \ell_t &= \sum_{i=0}^2 t(i) \sup_{\theta_i} \left\{ \theta_i \hat{\mu}_{i,t(i)} - \psi(\theta_i) \right\} - \sup_{\theta_1 \leq \theta_0, \theta_2 \leq \theta_0} \sum_{i=0}^2 t(i) \left\{ \theta_i \hat{\mu}_{i,t(i)} - \psi(\theta_i) \right\} \\ &= \begin{cases} 0 & \text{if } \hat{\mu}_t^{\mathbf{I}} \leq \hat{\mu}_t^0, \\ \sum_{i=0, \mathbf{I}} t(i) I\left(\hat{\mu}_t^i, \left[t(0) \hat{\mu}_t^0 + t(\mathbf{I}) \hat{\mu}_t^{\mathbf{I}} \right] / [t(0) + t(\mathbf{I})]\right) & \text{if } \hat{\mu}_t^{\mathbf{II}} \leq \hat{\mu}_t^0 < \hat{\mu}_t^{\mathbf{I}}, \\ \sum_{i=0, \mathbf{I}, \mathbf{II}} t(i) I\left(\hat{\mu}_t^i, \left[t(0) \hat{\mu}_t^0 + t(\mathbf{I}) \hat{\mu}_t^{\mathbf{I}} + t(\mathbf{II}) \hat{\mu}_t^{\mathbf{II}} \right] / t\right) & \text{if } \hat{\mu}_t^{\mathbf{II}} > \hat{\mu}_t^0. \end{cases} \quad (2.12) \end{aligned}$$

We have assumed in (2.12) that no arm has been eliminated by time t . In the case an arm has been eliminated, it is not involved in the subsequent GLR test, which then only involves the remaining treatment arm and arm 0, consistent with the case $\hat{\mu}_t^{\mathbf{II}} \leq \hat{\mu}_t^0 < \hat{\mu}_t^{\mathbf{I}}$ in (2.12).

Elimination of treatment arm i occurs at time t if its sample mean does not exceed $\hat{\mu}_t^0$ and its elimination does not violate the third constraint in (2.7). Specifically, for $i = 1, 2$, let

$$I_{i,t} = \sup_{(\mu_0, \mu_i) \in H_i} \left\{ t(i) I\left(\hat{\mu}_{i,t(i)}, \mu_i\right) + t(0) I\left(\hat{\mu}_{0,t(0)}, \mu_0\right) \right\} \quad (2.13)$$

be the GLR statistic for testing $H_i: \mu_1 = \mu_0 + \delta (= \tilde{\mu})$ based on sequential observations from treatment arm i and the control arm 0. Arm i is eliminated at time t if

$$\hat{\mu}_{i,t(i)} \leq \hat{\mu}_t^0 \quad \text{and} \quad I_{i,t} \geq c_{\tilde{\alpha}}, \quad (2.14)$$

where $c_{\tilde{\alpha}} \sim |\log \tilde{\alpha}|$ is chosen to satisfy the third constraint in (2.7). The trial stops for futility whenever both arms 1 and 2 are eliminated.

We start the trial with equal randomization for arms 0, 1, 2 up to stage $t^* = o(|\log \alpha|)$ to collect information about the three means. At any stage $t > t^*$, if one treatment arm has been eliminated, use equal randomization for the remaining treatment arm and arm 0. Otherwise, for $t > t^*$, randomize to 0, I, II according to the sampling proportion $1 : 1 : \varepsilon$ if $\hat{\mu}_t^{\text{I}} - \hat{\mu}_t^{\text{II}} > \delta_t$ and $2 : 1 : 1$ otherwise, where $\delta_t \rightarrow 0$ but $t\delta_t \rightarrow \infty$. With this adaptive sampling rule, the sequential GLR test of H_0 stops sampling and rejects H_0 as soon as

$$\hat{\mu}_t^{\text{I}} > \hat{\mu}_t^0 \quad \text{and} \quad \ell_t \geq a_{\alpha}. \quad (2.15)$$

It also stops for futility, accepting H_0 , as soon as

$$\hat{\mu}_t^{\text{I}} < \hat{\mu}_t^0 + \delta \quad \text{and} \quad I_{\text{I},t} \geq b_{\tilde{\alpha}}. \quad (2.16)$$

The thresholds $a_{\alpha} \sim |\log \alpha|$ and $b_{\tilde{\alpha}} \sim |\log \tilde{\alpha}|$ are chosen such that the first two constraints in (2.7) are satisfied. By Theorem 2.1 below, this procedure attains the asymptotic lower bounds in Section 2.1 as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$.

2.3. Asymptotic Theory for General k

We first extend the procedure in the preceding section to $k - 2$ treatment arms by describing the extension for each of the four components of the procedure.

(a) Elimination of inferior arms—The elimination criterion (2.14) can still be applied to arm $i = 1, \dots, k$. Let $\mathcal{K}_t$ denote the set of all surviving arms (i.e., which have not been eliminated) at the end of stage t .

(b) Randomization proportions—Let $s = t - 1$ and let $\mathbf{I}_s = \arg \max_{i \in \mathcal{K}_s} \hat{\mu}_{i,s(i)}$, picking any maximizer if there are ties. Let

$$\hat{\mu}_s^{\text{I}} = \hat{\mu}_{i,s(i)} \quad \text{with} \quad i = \mathbf{I}_s, \quad \mathcal{H}_s = \left\{ i \in \mathcal{K}_s : |\hat{\mu}_{i,s(i)} - \hat{\mu}_s^{\text{I}}| \leq \delta_s \right\}, \quad (2.17)$$

and let $|\mathcal{H}_s|$ denote its cardinality. Apply equal randomization to the arms of $\mathcal{K}_{t-1}$ if $t = t^* = o(|\log \alpha|)$. For $t > t^*$, randomize to arm 0 with probability proportional to 1, to each arm $i \in \mathcal{K}_{t-1} - \mathcal{H}_{t-1}$ with probability proportional to ε , and to each arm $i \in \mathcal{H}_{t-1}$ with probability proportional to $1/|\mathcal{H}_s|$. The randomization generates an observation at stage t and thereby updates $\mathcal{K}_t$.

(c) Early stopping for futility—Stop sampling and accept H_0 as soon as $\mathcal{K}_t = \emptyset$ or (2.16) occurs.

(d) GLR criterion for rejecting H_0 —Early stopping can also occur if (2.15) occurs, rejecting H_0 upon stopping, where

$$\ell_t = \sum_{i \in \{0\} \cup \mathcal{K}_t} t(i) \sup_{\theta_i} \left\{ \theta_i \hat{\mu}_{i,t(i)} - \psi(\theta_i) \right\} - \sup_{\max\{\theta_i : i \in \mathcal{K}_t\} \leq \theta_0} \sum_{i \in \{0\} \cup \mathcal{K}_t} t(i) \left\{ \theta_i \hat{\mu}_{i,t(i)} - \psi(\theta_i) \right\}. \quad (2.18)$$

An explicit formula for ℓ_t similar to the second equality in (2.12) also holds for general k .

The asymptotic lower bounds in Section 2.1 can also be extended to general k and the above procedure can be shown to attain these bounds. This is the content of the following theorem, which is proved in Section 2.4.

Theorem 2.1—Let $\mu^* = \max_{i \neq k} \mu_i$, $A = \{i \neq k : \mu_0 < \mu_i < \mu^*\}$, $s_A = \sum_{i \in A} \mu_i$,

$n_A = \sum_{i \in A} 1$, with the sum over an empty set being 0. Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$ and define

$$\mathcal{J}_\epsilon(\boldsymbol{\mu}, \mu_0) = \begin{cases} I\left(\mu^*, \frac{\mu^* + \mu_0 + \epsilon s_A}{2 + \epsilon n_A}\right) + \epsilon \sum_{i \in A} I\left(\mu_i, \frac{\mu^* + \mu_0 + \epsilon s_A}{2 + \epsilon n_A}\right) + I\left(\mu_0, \frac{\mu^* + \mu_0 + \epsilon s_A}{2 + \epsilon n_A}\right) & \text{if } \mu^* > \mu_0, \\ 0 & \text{if } \mu^* \leq \mu_0. \end{cases} \quad (2.19)$$

Let $H = \{i \neq k : \mu_i = \mu^*\}$. Then for any sequential decision rule satisfying (2.1) and (2.7), as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$,

$$E_{\mu_0, \dots, \mu_k} \left(\sum_{i \in H} T_i \right) / |\log \alpha| \geq (1 + o(1)) / \max \{ \mathcal{J}_\epsilon(\boldsymbol{\mu}, \mu_0), I(\mu^*, \tilde{\mu}) \}. \quad (2.20)$$

Moreover, $T_0 = (1 + o_p(1)) \sum_{i \in H} T_i$ and for any arm j with $\mu_j = \mu_0$ and $\mu_j < \mu^*$,

$$E_{\mu_0, \dots, \mu_k} (T_j) / |\log \alpha| \geq (1 + o(1)) \min \left\{ \frac{1}{I(\mu_j, \tilde{\mu})}, \frac{\epsilon}{\max \{ \mathcal{J}_\epsilon(\boldsymbol{\mu}, \mu_0), I(\mu^*, \tilde{\mu}) \}} \right\}. \quad (2.21)$$

The asymptotic lower bounds (2.20) and (2.21) are attained by the procedure described above for which δ_s in (2.17) satisfies $\delta_s \rightarrow 0$ but $\sqrt{s} \delta_s \rightarrow \infty$ as $s \rightarrow \infty$.

2.4. Proofs

To prove Theorem 2.1, we first focus on the case $k = 2$, which is analyzed in Section 2.1, concerning lower bounds for the expected sample size for the $k + 1$ arms, and in Section 2.2 on how to attain these bounds. Details of the proof of these results in Sections 2.1 and 2.2 are given here, before moving on to extend them to the case $k > 2$ for the proof of Theorem 2.1. To relate the lower bounds to a generalization of the Hoeffding (1960) lower bound, we have also considered the case $k = 1$, for which we have given the corresponding lower bound (2.6) and stated that it can be achieved asymptotically by a sequential GLR test. We begin by proving (2.6) and showing how it can be attained.

Proof and Attainment of (2.6). Note that the case $k = 1$ involves two arms, labeled 0 and 1, to which we assign subjects with equal probability $1/2$. Given (μ_1, μ_0) , let $\bar{\mu} = (\mu_1 + \mu_0)/2$ and the test has to satisfy $P_{\bar{\mu}, \bar{\mu}}^-(\text{Reject } H_0) \leq \alpha$, $P_{\mu_0, \bar{\mu}}(\text{Accept } H_0) \leq \tilde{\alpha}$ in view of (2.5).

A change of measures from $P_{\bar{\mu}, \bar{\mu}}^-$ to P_{μ_0, μ_1} leads to the likelihood ratio statistic (Radon-Nikodym derivative) $e^{-\ell_t(0)}$, where

$$\ell_t(0) = t(0) \left\{ \left(\theta_0 - \bar{\theta} \right) \hat{\mu}_{0,t(0)} - \psi(\theta_0) + \psi(\bar{\theta}) \right\} + t(1) \left\{ \left(\theta_1 - \bar{\theta} \right) \hat{\mu}_{1,t(1)} - \psi(\theta_1) + \psi(\bar{\theta}) \right\} \quad (2.22)$$

and $\psi'(\bar{\theta}) = \bar{\mu}$. Similarly, changing $P_{\mu_0, \bar{\mu}}$ to P_{μ_0, μ_1} leads to $e^{-\ell_t(1)}$, where

$$\ell_t(1) = t(1) \left\{ \left(\theta_1 - \tilde{\theta} \right) \hat{\mu}_{1,t(1)} - \psi(\theta_1) + \psi(\tilde{\theta}) \right\}. \quad (2.23)$$

Therefore, taking $0 < \rho < 1$ and $\tilde{\rho} > 1$ such that $\rho\tilde{\rho} < 1$ and letting m be the largest integer $\leq \rho \min \{ |\log \alpha| / \mathcal{J}(\mu_1, \mu_0), |\log \tilde{\alpha}| / I(\mu_1, \tilde{\mu}) \}$, we can use the same argument as that in Lai (1981, pp. 325-326) to show that $P_{\mu_0, \mu_1}(T \leq m)$ is bounded above by

$$\alpha^{1-\rho\tilde{\rho}} + \tilde{\alpha}^{1-\rho\tilde{\rho}} + P \left\{ \max_{t \leq m} \ell_t(0) > \tilde{\rho} m \mathcal{J}(\mu_1, \mu_0) \right\} + P \left\{ \max_{t \leq m} \ell_t(1) > \tilde{\rho} m \mathcal{J}(\mu_1, \mu_0) \right\}$$

since $t(0) = (1 + o_p(1)) t(1)$, $(\theta_1 - \tilde{\theta}) \mu_1 - \psi(\theta_1) + \psi(\tilde{\theta}) = I(\mu_1, \tilde{\mu})$ and

$$\left\{ \left(\theta_0 - \bar{\theta} \right) \mu_0 - \psi(\theta_0) + \psi(\bar{\theta}) \right\} + \left\{ \left(\theta_1 - \bar{\theta} \right) \mu_1 - \psi(\theta_1) + \psi(\bar{\theta}) \right\} = I(\mu_0, \bar{\mu}) + I(\mu_1, \bar{\mu}),$$

which is equal to $\mathcal{J}(\mu_1, \mu_0)$. This shows that $P_{\mu_0, \mu_1}(T \leq m) \rightarrow 0$ as $\alpha \rightarrow 0$ and $\tilde{\alpha} \rightarrow 0$. Since ρ can be arbitrarily close to 1, it then follows that

$$E_{\mu_0, \mu_1} T_1 \geq (1 + o(1)) \min \{ |\log \alpha| / \mathcal{J}(\mu_1, \mu_0), |\log \tilde{\alpha}| / I(\mu_1, \tilde{\mu}) \}. \quad (2.24)$$

From (2.4) and (2.24), (2.6) follows.

Noting that the actual mean vector (μ_0, μ_1) is unknown and that the null hypothesis $H_0: \mu_1 = \mu_0$ is composite, we use the sequential GLR test that replaces the unknown parameters in $\ell_t(0)$ and $\ell_t(1)$ by their maximum likelihood estimates at stage t , leading to

$\hat{\ell}_t^{(1)} = t(1) I(\hat{\mu}_{1,t(1)}, \hat{\mu}_{0,t(0)} + \delta)$ and

$$\hat{\ell}_t^{(0)} = \sum_{i=0}^1 t(i) I(\hat{\mu}_{i,t(i)}, [t(0) \hat{\mu}_{0,t(0)} + t(1) \hat{\mu}_{1,t(1)}] / t). \quad (2.25)$$

The sequential GLR test that attains the asymptotic lower bound (2.6), as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$, rejects H_0 at stage t if $\hat{\mu}_{1,t(1)} > \hat{\mu}_{0,t(0)}$ and $\hat{\ell}_t^{(0)} \geq a_\alpha$, and stops and accepts H_0 if $\hat{\mu}_{1,t(1)} < \hat{\mu}_{0,t(0)} + \delta$ and $\hat{\ell}_t^{(1)} \geq b_{\tilde{\alpha}}$, where $a_\alpha \sim b_{\tilde{\alpha}} \sim |\log \alpha|$ are chosen so that (2.5) holds; see Lai (2001, pp. 306–307).

Proof of (2.8). Note that here we have $k = 2$ arms and $\mu_2 > \mu_0, \mu_1 > \mu_2$. By (2.1), $P_{\mu_0, \mu_1, \mu_2}\{t(i) \rightarrow \infty \text{ for every } i = 1, \dots, k\} \rightarrow 1$ as $t \rightarrow 0$. Applying a change of measures from $P_{\mu_0, \mu_1, \bar{\mu}}$ to P_{μ_0, μ_1, μ_2} , an argument similar to that used to prove the lower bound (2.6) can be used to prove (2.8).

Proof of (2.11). Note that here we have $\mu_1 > \mu_2 > \mu_0$. Since the optimal allocation rule samples minimally from arm 2 under the constraints (2.1), $t(2) \sim \infty(1)$ and $t(0) \sim t(1)$ for all large t , a.s. $[P_{\mu_0, \mu_1, \mu_2}]$. We modify the proof of (2.6) by letting $\bar{\mu} = (\mu_1 + \epsilon \mu_2 + \mu_0) / (2 + \epsilon)$ and considering the probability constraints $P_{\bar{\mu}, \bar{\mu}, \bar{\mu}}(\text{Reject } H_0) \leq \alpha$ and $P_{\mu_0, \bar{\mu}, \mu_2}(\text{Accept } H_0) \leq \tilde{\alpha}$ implied by (2.7). A change of measures from $P_{\bar{\mu}, \bar{\mu}, \bar{\mu}}$ to P_{μ_0, μ_1, μ_2} leads to the likelihood ratio statistics $e^{-\ell_t(0)}$, where $\ell_t(0)$ is given by the right-hand side of (2.22) together with an additional summand $t(2) \left\{ \left(\theta_2 - \bar{\theta} \right) \hat{\mu}_{2,t(2)} - \psi(\theta_2) + \psi(\bar{\theta}) \right\}$. This additional summand leads to the summand $\epsilon \mathbf{I}(\mu_2, \bar{\mu})$ in (2.10) with $v = \mu_0$. With this modification, we can use the same argument in the proof of (2.6) to prove the lower bound (2.11), since changing $P_{\mu_0, \bar{\mu}, \mu_2}$ to P_{μ_0, μ_1, μ_2} leads to $e^{-\ell_t(1)}$, where $\ell_t(1)$ is given by the right-hand side of (2.23) in which $\bar{\theta}$ is redefined here for $k = 2$.

Proof of Theorem 2.1. We first consider the case $k = 2$ and show that the sequential GLR test in Section 2.2 indeed attains the asymptotic lower bounds in Section 2.1. The argument is an extension of that used earlier in the paragraph containing (2.25) to establish asymptotic efficiency of the sequential GLR tests in the case $k = 1$. In fact, ℓ_t defined by (2.12) is an extension to the case $k = 2$ of $\hat{\ell}_t(0) I_{\{\hat{\mu}_{1,t(1)} > \hat{\mu}_{0,t(0)}\}}$ in (2.25). The major complication in this extension lies in the adaptive randomization rule and the possibility of an arm being eliminated before the trial stops; these features are needed for the rule to be asymptotically efficient when $k > 1$. As we have already pointed out in the first paragraph of Section 2.2, the requirement (2.1) of forced sampling from the apparently inferior arm ensures consistency, at rate $O_p(1/\sqrt{t})$, of the estimates $\hat{\mu}_t^{\mathbf{H}}, \hat{\mu}_t^{\mathbf{I}}$, and $\hat{\mu}_t^0$. This consistency implies that $T_i/|\log \alpha|$ converges in probability, for $i = 1, 2$, to the lower bound given in Section 2.1, where it is pointed out that in the case $\mu_1 = \mu_2$, arms 1 and 2 are indistinguishable and can be combined into a single arm. Uniform integrability arguments involving exponential bounds for tail probabilities of sample means from the underlying exponential family of distributions can then be used to show that the sequential GLR test indeed attains the asymptotic lower bounds in Section 2.1, similar to the case $k = 1$.

The case $k > 2$ in Theorem 2.3 can be proved similarly. In particular, the asymptotic lower bounds (2.20) and (2.21) can be proved by using similar changes of measures. Note that the

set A in Theorem 2.1 excludes the arm i that attains μ^* and therefore is empty in Case 1 of Section 2.1 that considers $k = 2$ with $\mu_1 > \mu_2$ and $\mu_2 = \mu_0$.

3. EXTENSION TO GROUP SEQUENTIAL DESIGN

As pointed out in Section 1, group sequential designs are commonly used in double-blind randomized clinical trials. Sequential decisions concerning randomization proportions, futility stopping, and rejection of H_0 are made only at the times of interim analysis. Let n_j denote the total number of observations up to the time of the j th interim analysis, $j = 1, \dots, J$, and $n_j(i)$ denote that from arm i , so that n_J is the total sample size by the scheduled end of the trial. Thus, n_J can be regarded as the maximum sample size of the trial, which can terminate earlier at an interim analysis prior to its scheduled end; see Lai and Shih (2004). By restricting t to the set $\{n_1, \dots, n_J\}$ for the four components of the procedure in Section 2.3, we have a group sequential test of H_0 . In particular, we can take $t^* = n_1$ in (b) of Section 2.3 so that equal randomization is used up to the first interim analysis. The randomization proportion determined at the j th interim analysis remains the same for observations $n_j + 1, \dots, n_{j+1}$. Moreover, arm elimination and early stopping can only be determined at interim analyses. With $n_1 = o(|\log \alpha|)$, the procedure is still asymptotically efficient among all group sequential procedures satisfying (2.1) and (2.7) and having $n_1, \dots, n_J$ observations at successive interim analyses satisfying

$$\liminf (n_j - n_{j-1}) / |\log \alpha| > 0, \quad j=2, \dots, J, \quad (3.1)$$

as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$. This can be proved by modifying the asymptotic lower bounds in Theorem 2.1 and the arguments used by Lai and Shih (2004) to prove the asymptotic efficiency of group sequential tests.

The maximum sample size n_J of the group sequential design reflects the maximum resource in terms of patient accrual costs and duration that can be allocated to the trial. Clearly a trial that involves $k = 2$ treatment arms to test the treatment's effectiveness over a control needs a larger maximum sample size than the conventional setting of $k = 1$ treatment arm to detect the same effect size with prescribed power. On the other hand, if one actually does not know which treatment arm to choose at the design stage and does not want to commit to a single treatment arm without generating sufficient clinical data to guide the choice, then the above group sequential design can include the k treatment arms in the trial, for which it uses an adaptive randomization scheme to assign the patients to the $k + 1$ treatment and control arms. At $t = n_j$ for $j = 1, \dots, J - 1$, the group sequential GLR test rejects H_0 if

$$\hat{\mu}_t^{\mathbf{I}} > \hat{\mu}_{0,t(0)} \quad \text{and} \quad \ell_t \geq a_\alpha, \quad (3.2)$$

where $\hat{\mu}_t^{\mathbf{I}}$ is defined in (2.17). It also accepts H_0 and stops for futility if (2.16) holds and eliminates arm i from further consideration if (2.14) holds. When early stopping for efficacy or futility has not occurred and the trial proceeds to its scheduled end at $t = n_J$, the group sequential GLR test rejects H_0 if

$$\hat{\mu}_t^I > \hat{\mu}_{0,t(0)} \quad \text{and} \quad \ell_t \geq d_\alpha. \quad (3.3)$$

To maintain the type I error probability at a prescribed level α , we choose the “efficacy boundaries” a_α and d_α such that

$$P_{\mu_1=\dots=\mu_k=\mu_0} \{ (3.2) \text{ occurs at some } t=n_j \text{ with } j < J \} = \epsilon_1 \alpha, \\ P_{\mu_1=\dots=\mu_k=\mu_0} \{ (3.2) \text{ does not occur before } n_j \text{ and } (3.3) \text{ holds at } t=n_j \} = (1 - \epsilon_1) \alpha, \quad (3.4)$$

where $0 < \epsilon_1 < 1$. The type I error $P_{\max(\mu_1, \dots, \mu_k)=\mu_0}(\text{Reject } H_0)$ is maintained at level α by controlling it at the least favorable parameter configurations $\mu_1 = \dots = \mu_k = \mu_0$. We describe below how the probabilities in (3.4) can be computed.

3.1. Evaluation of Boundary Crossing Probabilities in (3.4)

As noted in the proof of Theorem 2.1, $T_i/|\log \alpha|$ converges in probability, as $\alpha + \tilde{\alpha} \rightarrow 0$ such that $|\log \alpha| \sim |\log \tilde{\alpha}|$, to the asymptotic lower bound in (2.21) or (2.20) in which we use

$\sum_{i \in H} T_i$ to replace the individual sample sizes T_i from the indistinguishable arms in the set H of best arms. Because of property (b) in Section 2.3 for the proposed allocation rule, $T_i/|\log \alpha|$ still converges in probability, for $i \in H$, to $|H|^{-1}$ times the right-hand side of (2.20), where $|H|$ denotes the cardinality of H . Hence, the adaptive allocation rules are asymptotically nonrandom, with sampling proportion $p_0 : p_1 : \dots : p_k$ that depends on $(\mu_0, \dots, \mu_k)$. In fact, for $1 \leq i \leq k$, the p_i can be taken to be the right-hand side of (2.21), or $|H|^{-1}$ times the right-hand side of (2.20), while p_0 can be taken to be the right-hand side of

(2.20). Using this asymptotic approximation and normalizing the p_i so that $\sum_{i=0}^k p_i = 1$, we can approximate $t(i)$ by $p_i t$ at $t = n_j$ for $1 \leq j \leq J$. In particular, under $\mu_0 = \dots = \mu_k$, $(t(0), t(1), \dots, t(k))$ can be approximated by $(t/2, t/2k, \dots, t/2k)$ at $t = n_j$ for $1 \leq j \leq J$.

To evaluate the probabilities in (3.4), we use this approximation and apply the multivariate central limit theorem to $\left\{ \left(S_{n_j/2}^0, S_{n_j/(2k)}^1, \dots, S_{n_j/(2k)}^k \right), \quad 1 \leq j \leq J \right\}$, where $S_m^i = m \hat{\mu}_{i,m}$ is the sample sum based on m observations from arm i . Under $\mu_0 = \dots = \mu_k$,

$$\left\{ \Delta_j^i := \left(\frac{n_j}{2(k+1)\psi''(\mu)} \right)^{1/2} \left(\hat{\mu}_{i,n_j/2k} - \hat{\mu}_{0,n_j/2} \right) : 1 \leq i \leq k \right\} \Rightarrow N(0, \Sigma) \quad (3.5)$$

and $\sqrt{n_j}(\Delta_j^1, \dots, \Delta_j^k)$ has asymptotically independent increments in $1 \leq j \leq J$, where Σ is the correlation matrix due to the common term $\hat{\mu}_{0,n_j/2}$ in Δ_j^i . Moreover, at $t = n_j$, ℓ_t defined in (2.18) can be approximated under $\mu_0 = \mu_k$ by a quadratic form of $(\Delta_j^1, \dots, \Delta_j^k)$ with an $o_p(1)$ error, as can be shown by standard asymptotic arguments for GLR statistics showing their limiting χ^2 distribution under the null hypothesis.

3.1.1. Recursive Numerical Integration Using (3.5) to Evaluate (3.4) for $k = 2$ —Recursive numerical integration, introduced by Armitage et al. (1969) for computing type I

error probabilities of repeated significance tests of a normal mean, has been extended to asymptotically normal test statistics, with independent increments over time, in group sequential trials; see chapter 19 of Jennison and Turnbull (2000). Although the stopping criterion (2.15) and the GLR statistics (2.18) are considerably more complicated than those in traditional group sequential tests, the approximation in the preceding paragraph simplifies the boundary crossing probabilities in (3.4) to those for a k -dimensional normal random walk. In particular, for $k = 2$, the GLR statistic ℓ_t at $t = n_j$ is given by (2.12) and can be expressed as

$$\ell_t = \begin{cases} o_p(1) & \text{if } \Delta_t^I < 0, \\ \left(\Delta_j^I\right)^2 / 2 + o_p(1) & \text{if } \Delta_t^I \leq 0 < \Delta_t^I, \\ \frac{3}{16} \left[3\left(\Delta_j^I\right)^2 + 3\left(\Delta_j^{\text{II}}\right)^2 - 2\Delta_t^I \Delta_t^{\text{II}} \right] + o_p(1) & \text{if } \Delta_t^{\text{II}} > 0. \end{cases} \quad (3.6)$$

Therefore, the early stopping criterion (3.2) for rejecting H_0 at $t = n_j$ can be expressed as the occurrence of one of the following disjoint events, using the approximations (3.6) and $t(0) \approx n_j/2$, $t(1) \approx n_j/4$:

$$\begin{aligned} \Omega_{1,j} &= \left\{ \max_{1 \leq i \leq 2} \Delta_j^i \geq \sqrt{2a_\alpha} \right\} \\ \Omega_{2,j} &= \left\{ \frac{4}{3} \sqrt{a_\alpha} \leq \Delta_j^1 \leq \sqrt{2a_\alpha} \text{ and } 0 \leq \Delta_j^2 \leq \frac{1}{3} \left(\Delta_j^1 - 2 \left[2 \left(2a_\alpha - \left(\Delta_j^1 \right)^2 \right) \right]^{1/2} \right) \right\} \\ \Omega_{3,j} &= \left\{ \frac{4}{3} \sqrt{a_\alpha} \leq \Delta_j^2 \leq \sqrt{2a_\alpha} \text{ and } 0 \leq \Delta_j^1 \leq \frac{1}{3} \left(\Delta_j^2 - 2 \left[2 \left(2a_\alpha - \left(\Delta_j^2 \right)^2 \right) \right]^{1/2} \right) \right\} \\ \Omega_{4,j} &= \left\{ \sqrt{2a_\alpha}/3 \leq \Delta_j^1 \leq \sqrt{2a_\alpha} \text{ and } \frac{1}{3} \left(\Delta_j^1 + 2 \left[2 \left(2a_\alpha - \left(\Delta_j^1 \right)^2 \right) \right]^{1/2} \right) \leq \Delta_j^2 \leq \sqrt{2a_\alpha} \right\}. \end{aligned}$$

In addition, since

$$\Delta_j^i = \sqrt{\frac{n_{j-1}}{n_j}} \Delta_{j-1}^i + \sqrt{\frac{2}{3n_j \psi''(\mu)}} \left\{ 2 \left(S_{n_j/4}^i - S_{n_{j-1}/4}^i \right) - \left(S_{n_j/2}^0 - S_{n_{j-1}/2}^0 \right) \right\}, \quad (3.7)$$

the conditional density $p_j \left(\Delta_j^1, \Delta_j^2 \mid \Delta_{j-1}^1, \Delta_{j-1}^2 \right)$ of $\left(\Delta_j^1, \Delta_j^2 \right)$ given $\left(\Delta_{j-1}^1, \Delta_{j-1}^2 \right)$ is

$$\varphi \left(\sqrt{n_j} \Delta_j^1 - \sqrt{n_{j-1}} \Delta_{j-1}^1, \sqrt{n_j} \Delta_j^2 - \sqrt{n_{j-1}} \Delta_{j-1}^2; \frac{(n_j - n_{j-1})}{3} \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \right),$$

where $\varphi(\cdot, \cdot; \Sigma)$ is the density function of a bivariate normal random vector with zero mean vector and covariance matrix Σ . The first probability in (3.4) can be expressed as a sum of bivariate integrals

$$\sum_{j=1}^{J-1} \int_{\Omega_{1,j} \cup \dots \cup \Omega_{4,j}} f_j(x, y) dx dy,$$

in which the integrands can be computed recursively by

$$f_j \left(\Delta_j^1, \Delta_j^2 \right) = \int_{(\Omega_{1,j-1} \cup \dots \cup \Omega_{4,j-1})^c} f_{j-1}(x, y) p_j \left(\Delta_j^1, \Delta_j^2 | x, y \right) dx \quad dy,$$

with $f_0 \equiv 1$. Similarly, replacing a_α by d_α in $\Omega_{m,J}$ for $1 \leq m \leq 4$, we can evaluate

$$\int_{\Omega_{1,J} \cup \dots \cup \Omega_{4,J}} f_J(x, y) dx \quad dy$$

for the second probability in (3.4) to determine d_α .

3.1.2. Recursive Integration to Evaluate an Upper Bound of (3.4) for $k > 2$ —The above recursive numerical algorithm in the case $k = 2$ involves bivariate integration.

Although it can be extended to general k , each recursive step to evaluate $f_j \left(\Delta_j^1, \dots, \Delta_j^k \right)$ would involve k -fold integration. Instead of evaluating (3.4) directly, we develop a sharp upper bound that can be computed recursively by univariate integration. First note that $\ell_t \leq \ell'_t$ for all t , where

$$\ell'_t = \sum_{i=0}^k t(i) \sup_{\theta_i} \left\{ \theta_i \hat{\mu}_{i,t(i)} - \psi(\theta_i) \right\} - \sup_{\theta_1 = \dots = \theta_k = \theta_0} \sum_{i=0}^k t(i) \left\{ \theta_i \hat{\mu}_{i,t(i)} - \psi(\theta_i) \right\} \quad (3.8)$$

is the GLR statistic for testing $H'_0: \mu_1 = \dots = \mu_k = \mu_0$ (or, equivalently, $\theta_1 = \dots = \theta_k = \theta_0$).

Therefore, we can bound the type I error probability by

$$\begin{aligned} P_{H_0}(\text{reject } H_0) &\leq P_{\mu_1 = \dots = \mu_k = \mu_0} \left(\left\{ \ell_t \geq a_\alpha \text{ at some } t < J \right\} \cup \left\{ \hat{\mu}_t^I > \hat{\mu}_{0,t(0)} \text{ and } \ell_t \geq d_\alpha \text{ at } t = n_J \right\} \right) \\ &\leq P_{\mu_1 = \dots = \mu_k = \mu_0} \left(\left\{ \ell'_t \geq a_\alpha \text{ at some } t < J \right\} \cup \left\{ \hat{\mu}_t^I > \hat{\mu}_{0,t(0)} \text{ and } \ell'_t \geq d_\alpha \text{ at } t = n_J \right\} \right). \end{aligned} \quad (3.9)$$

The first inequality in (3.9) is due to dropping the constraint $\hat{\mu}_t^I > \hat{\mu}_{0,t(0)}$ in (3.4) and the second one uses the bound $\ell_t \leq \ell'_t$.

The upper bound in the third line of (3.9) can also be decomposed into two summands that are similar to the probabilities in (3.4) so that a_α can be determined by the equation

$$P_{\mu_1 = \dots = \mu_k = \mu_0} \left(\ell'_t \geq a_\alpha \text{ at } t = n_j \text{ for some } j < J \right) = \epsilon_1 \alpha \quad (3.10)$$

for the first summand. Since $t(0) \approx n_j/2$ and $t(i) \approx n_j/(2k)$ for $1 \leq i \leq k$ and $t = n_j$, standard likelihood theory shows that $2\ell'_{n_j}$ has a limiting χ_k^2 distribution under the null hypothesis H'_0 . Moreover, in view of the central limit theorem for $\left(S_{n_j/2}^0, \dots, S_{n_j/(2k)}^k \right)$, we can apply the result of Jennison and Turnbull (1991) for the normal case to conclude that $2\ell'_{n_j}$ conditioned on $2\ell'_{n_{j-1}}$ has the approximate distribution

$$[(n_j - n_{j-1})/n_j] \chi_k^2 \left(2\ell'_{n_{j-1}} n_{j-1} / (n_j - n_{j-1}) \right), \quad (3.11)$$

where $\chi_k^2(u)$ denotes a chi-square distribution with k degrees of freedom and noncentrality parameter u . Using this Markov property, we can evaluate (3.10) and the probability

$P_{\mu_1=\dots=\mu_k=\mu_0} \left(\ell'_{n_j} < a_\alpha \text{ for } 1 \leq j \leq J-1 \text{ and } \ell'_{n_J} \geq d_\alpha \right)$, which is given below (3.12), by recursive numerical integration as in Jennison and Turnbull (1991).

Let $p = P_{\mu_1=\dots=\mu_k=\mu_0} \left(\max_{1 \leq i \leq k} \Delta_J^1 \leq 0 \right)$, which can be approximated by the probability that the maximum of k independent normal random variables does not exceed another

independent normal random variable to which $S_{n_j/2}^0 / \sqrt{n_j/2}$ converges weakly. By conditioning on the latter normal random variable, p can be easily evaluated by univariate numerical integration. For a multivariate normal random vector $\mathbf{Z}$ with mean $\mathbf{0}$, $\mathbf{Z}/\|\mathbf{Z}\|$ is independent of $\|\mathbf{Z}\|$. Moreover, under H_0' , $2\ell'_{n_j} = \sum_{i=1}^k \left(\Delta_j^i \right)^2 + o_p(1)$ for $1 \leq j \leq J$ and Δ_j^i is approximately a zero-mean normal random walk divided by its standard deviation. Making use of this, it can be shown that

$$\begin{aligned} P_{\mu_1=\dots=\mu_k=\mu_0} \left(\ell'_{n_j} < a_\alpha \text{ for } 1 \leq j \leq J-1, \hat{\mu}_t^{\mathbf{I}} > \hat{\mu}_{0,t(0)} \text{ and } \ell'_t \geq d_\alpha \text{ at } t=n_J \right) \\ \leq (1-p+o(1)) P_{\mu_1=\dots=\mu_k=\mu_0} \left(\ell'_{n_j} < a_\alpha \text{ for } 1 \leq j \leq J-1 \text{ and } \ell'_{n_J} \geq d_\alpha \right). \end{aligned}$$

Therefore, to maintain the type I error probability of the test at level α , we can determine d_α by the equation

$$(1-p) P_{\mu_1=\dots=\mu_k=\mu_0} \left(\ell'_{n_j} < a_\alpha \text{ for } 1 \leq j \leq J-1, \text{ and } \ell'_{n_J} \geq d_\alpha \right) = (1-\epsilon_1) \alpha. \quad (3.12)$$

3.1.3. Monte Carlo Evaluation of (3.4)—An alternative to normal approximation and recursive numerical integration is to use Monte Carlo simulations to evaluate the probabilities in (3.4). We can use a bisection search within a range of possible values for a_α , using Monte Carlo to compute the left-hand side of the first equation in (3.4). After determining a_α in this way, we carry out another bisection search for d_α , again using Monte Carlo to compute the probability in the second equation in (3.4).

3.2. Computation of Thresholds for Futility Stopping and Arm Elimination

Although futility stopping does not inflate the type I error, it may substantially decrease power if it is not carried out carefully. In addition, premature elimination of an apparently inferior, but actually the best, treatment arm can also substantially decrease power. The second and third probability constraints in (2.7) are related to power considerations at the alternatives $\max_{1 \leq i \leq k} \mu_i = \mu_0 + \delta$ and $\mu_i = \mu_0 + \delta$ for $i = 1, \dots, k$. The futility stopping criterion (2.16) for the proposed test involves comparing the mean of the leading treatment arm (which maximizes $\hat{\mu}_{i,t(i)}$ over the surviving arms i) to the alternative $\mu_0 + \delta$. The GLR

statistic for testing $H_i: \mu_i = \mu_0 + \delta$ (with unknown μ_0) is $I_{i,t}$ given in (2.13). To satisfy the second probability constraint in (2.7), we can choose $b_{\tilde{\alpha}}$ in (2.16) so that it satisfies

$$P_{\mu_i=\mu_0+\delta} \left\{ \hat{\mu}_{i,t(i)} < \hat{\mu}_{0,t(0)} + \delta \quad \text{and} \quad I_{i,t} \geq b_{\tilde{\alpha}} \quad \text{at} \quad t=n_j \quad \text{for some} \quad 1 \leq j < J \right\} = \epsilon_2 \tilde{\alpha} \quad (3.13)$$

for some $0 < \epsilon_2 < 1$. Since the above probability is the same for $i = 1, \dots, k$, we can consider $i = 1$ for definiteness. Using the normal approximation to signed root likelihood ratio statistics of H_i (see section 3.4 of Lai and Shih, 2004), we can use recursive numerical integration to evaluate the probability in (3.13). Note that the event in (3.13) compares a given treatment arm i with the control arm 0, whereas (2.16) involves the leading arm $\mathbf{I}_t$ that may change with t . On the other hand, the set of alternatives $\max_{1 \leq i \leq k} \mu_i = \mu_0 + \delta$ contains the least favorable parameter configurations with $\mu_0 + \delta = \mu_i > \max_{1 \leq i' \leq k, i' \neq i} \mu_{i'}$. Under such parameter configuration, $\mathbf{I}_t = i$ for all t with probability approaching 1 and therefore the probability of early stopping using the futility criterion (2.16) approaches the left-hand side of (3.13).

Similarly, to satisfy the third probability constraint in (2.7), we choose $c_{\tilde{\alpha}}$ in the arm elimination criterion (2.14) such that

$$P_{\mu_i=\mu_0+\delta} \left\{ \hat{\mu}_{i,t(i)} < \hat{\mu}_{0,t(0)} \quad \text{and} \quad I_{i,t} \geq c_{\tilde{\alpha}} \quad \text{at} \quad t=n_j \quad \text{for some} \quad 1 \leq j < J \right\} = \epsilon_3 \tilde{\alpha}. \quad (3.14)$$

Note that the event in (3.13) requires that $\hat{\mu}_{i,t(i)} < \hat{\mu}_{0,t(0)} + \delta$ and that in (3.14) requires $\hat{\mu}_{i,t(i)} < \hat{\mu}_{0,t(0)}$ instead. The proposed test involves tuning parameters ϵ , ϵ_1 , ϵ_2 , and ϵ_3 given by (2.1), (3.4), (3.13), and (3.14), respectively. The choice of ϵ depends on the number of treatment arms and n_J , while the choice of ϵ_1 , ϵ_2 , and ϵ_3 is related to error spending and is discussed in Lai and Shih (2004). The simulation study in the next section illustrates these choices.

4. A SIMULATION STUDY

Consider the case of $k = 2$ treatment arms and a control arm, with normally distributed responses that have unknown means μ_i from arm i and common known variance 1. The simulation study compares a two-stage winner design (2SW) proposed by Shun et al. (2008) with the group sequential (GS) design developed in Section 3. Since the performance of the two-stage design depends heavily on the choice of the maximum and the first-stage sample sizes, we use the same first-stage and maximum sample sizes for both designs. The maximum sample size n_J is 500 and the GS design has $J = 5$ groups with common group size 100. With $n_J = 500$, the study is designed to have 80% power for detecting alternatives with $\delta = 0.27$. The GS design uses $a_{\alpha} = 7.12$, $d_{\alpha} = 4.28$, $b_{\tilde{\alpha}} = -1.93$, and $c_{\tilde{\alpha}} = -1.87$ corresponding to $\alpha = 0.05$, $\tilde{\alpha} = 0.2$, $\epsilon_1 = \frac{1}{2}$, $\epsilon_2 = \frac{1}{3}$, $\epsilon_3 = \frac{1}{6}$. The choice of ϵ_1 and ϵ_2 follows those recommended by Lai and Shih (2004) for error spending in modified Haybittle-Peto tests. The choice $\epsilon_3 = \epsilon_2/2$ is to ensure a stricter criterion for arm elimination than futility stopping. The trial starts with equal randomization to all three arms until the first interim analysis and then updates the adaptive allocation ratio using $\delta_t = t^{-1/4}$ and $\epsilon = 1/6$ in Section 2.2. Note that $k + 1 = 3$ and ϵ is half of its upper bound $1/(k + 1)$ prescribed in (2.1).

Moreover, δ_t satisfies $\delta_t \rightarrow 0$ and $\sqrt{t}\delta_t \rightarrow \infty$. We have tried several other choices of (ε, δ_t) and the results are similar, showing the robustness of the performance of the GS design to these tuning parameters. With the same first-stage and maximum sample sizes, the 2SW design conducts the first interim analysis at the time when 100 subjects have been observed. It drops the treatment arm with the smaller sample mean and performs the final analysis using a t -test to compare the remaining treatment and control arms when each of them has accrued a total of 234 subjects. The elimination arm has an expected sample size of $100/3 = 33$. The distribution of the final t -statistic is no longer normal. Shun et al. (2008) proposed normal approximations to calculate the type I error and power. Using this approximation, $H_0 : \mu_1 = \mu_2 = \mu_0$ is rejected when the final one-sided t -statistic is greater than 1.8. Since 2SW does not involve early stopping, a fixed sample size (FSS) test using the same adaptive allocation rule as GS is also included for comparison.

Table 1 summarizes the operating characteristics of these three tests. Each result is based on 100,000 simulations. With slight loss in power, the GS test has substantially smaller expected sample sizes than FSS. After the first stage, 2SW eliminates the treatment arm with the smaller sample mean and therefore uses an allocation proportion 1 : 0 : 1 for arms **I**, **II**, and 0, which is much more extreme than 1 : ε : 1 used by GS or FSS. The ratio of each arm's sample size to the total sample size, averaged over the 100,000 simulations, is shown in parentheses in Table 1, which also gives the probability of correctly selecting the best arm (PCS) at the end of the trial. In 2SW, the treatment arm that remains after the first stage is the selected arm, while in GS and FSS, the leading treatment arm with the larger sample mean at the end of the trial is selected. In Table 1, PCS (given in square brackets) is the probability of correctly choosing the actual best arm, which is arm 2 in all cases, at the end of the trial (regardless of rejecting H_0 or not). In cases where the two treatment arms are in fact equally good, the probability of choosing arm 2 is indicated. In particular, when $\mu_1 = 0.15$ and $\mu_2 = 0.27$, although the three tests have similar power ranging from 77% to 81%, 2SW has only 69% probability of choosing arm 2, compared to that of 80% and 83% for GS and FSS, respectively.

5. DISCUSSION

Asymptotically efficient adaptive allocation schemes have been developed by Lai (1987) for the multi-armed bandit problem and by Chan and Lai (2006) for optimal selection of the population with the largest mean in an exponential family subject to a lower bound constraint on the probability of correct selection. Although the adaptive randomization scheme in the present article is in the same spirit, there are important differences. First, the optimization problem considered herein is related to testing the null hypothesis of no improvement in mean response of a new treatment, with k possible treatment strategies (arms), over a control treatment, whereas those in Lai (1987) and Chan and Lai (2006) are concerned with selecting the arm with the largest mean. Secondly, the setting considered herein is that of a randomized clinical trial in which patients are randomized to treatment arms, rather than being assigned to the treatments by some prescribed function of the past observations. As pointed out in Section 2, the randomization probabilities have to satisfy a

lower bound of the type (2.1) to achieve informed consent of patients in the clinical trial setting.

Although we have followed Shun et al. (2008) to demonstrate the efficacy of a new drug via rejection of the null hypothesis $H_0 : \max(\theta_1, \dots, \theta_k) = \theta_0$, this demonstration may still be inadequate for winning approval of the drug because rejection of H_0 only implies that some dosing regimen of the drug is better than the control. Although the dose $\hat{i}$ with the largest sample mean would be recommended in labeling the drug if it is approved by the regulatory agency, $\hat{i}$ is a random variable that has an unknown sampling distribution. In Table 1, PCS gives the probability that $\hat{i}$ is in fact the best arm. The probability is at least 80% for GS in all cases that have indeed a unique best arm with a larger mean than μ_0 .

The design proposed herein for confirmatory Phase III trials is a group sequential design with adaptive treatment allocation. In contrast, the review of Dragalin (2010) is about adaptive designs for exploratory Phase II trials to study drug efficacy and dose-response relationship and to determine the dose for Phase III testing. The adaptive designs “explicitly address multiple trial goals, adaptively allocate subjects according to ongoing information needs, and allow termination for both early success and futility considerations” (Dragalin, 2010, p. 445). In Dragalin's review, statistical analysis of the data from these phase II trials consists of fitting pharmacokinetic and pharmacodynamic models on dose-toxicity and dose-efficacy relationships and multiple testing to compare different treatment strategies with an active control or placebo. The adaptive designs reviewed involve sample size re-estimation, multistage learning of dose-response relationships, or multiple treatment comparison. The group sequential design for confirmatory phase III testing developed herein has a much more focused objective of demonstrating that the new treatment is better than the active control or placebo. It addresses the situation when the phase II trial identifies a few doses (or treatment strategies), rather than a single one, to be used for phase III testing. Asymptotic theory and simulation results show that it has nearly optimal expected sample size subject to prespecified power at certain alternatives and type I error probability constraint.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant DMS-1106535 and the National Cancer Institute grant 1 P30 CA 124435. The authors are grateful to the referee for helpful comments and suggestions.

REFERENCES

- Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. *Journal of Royal Statistical Society, Series A*. 1969; 132:235–244.
- Chan HP, Lai TL. Sequential Generalized Likelihood Ratios and Adaptive Treatment Allocation for Optimal Sequential Selection. *Sequential Analysis*. 2006; 25:179–201.
- Dragalin V. Sequential Methods in Multi-Arm Clinical Trials. *Sequential Analysis*. 2010; 29:444–462.
- Hoeffding W. Lower Bounds for the Expected Sample Size and the Average Risk of a Sequential Procedure. *Annals of Mathematical Statistics*. 1960; 31:352–368.
- Jennison C, Turnbull BW. Exact Calculations for Sequential t , χ^2 and F Tests. *Biometrika*. 1991; 78:133–141.

- Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall; New York: 2000.
- Lai TL. Asymptotic Optimality of Invariant Sequential Probability Ratio Tests. *Annals of Statistics*. 1981; 9:318–333.
- Lai TL. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *Annals of Statistics*. 1987; 15:1091–1114.
- Lai TL. Sequential Analysis: Some Classical Problems and New Challenges (with Discussions). *Statistica Sinica*. 2001; 11:303–408.
- Lai TL, Shih MC. Power, Sample Size and Adaptation Considerations in the Design of Group Sequential Clinical Trials. *Biometrika*. 2004; 91:507–528.
- Lorden G. 2-SPRT'S and the Modified Kiefer-Weiss Problem of Minimizing an Expected Sample Size. *Annals of Statistics*. 1976; 4:281–291.
- Robbins H. Some Aspects of the Sequential Design of Experiments. *Bulletin of American Mathematical Society*. 1952; 58:527–535.
- Shun Z, Lan KKG, Soo Y. Interim Treatment Selection Using the Normal Approximation Approach in Clinical Trials. *Statistics in Medicine*. 2008; 27:597–618. [PubMed: 17619239]

Table 1

Power, expected total sample size (SS), and the expected ratio of arms 1, 2, and 0 in the total sample size for the GS, FSS, and 2SW designs. Also given is the probability [PCS] of correctly selecting the best arm at the end of the trial

(μ_1, μ_2, μ_0)	GS		FSS		2SW	
	Power [PCS]	SS (ratio)	Power [PCS]	SS (ratio)	Power [PCS]	SS (ratio)
(0, 0, 0)	0.05 [0.50]	383 (0.28, 0.28, 0.44)	0.048 [0.50]	500 (0.27, 0.28, 0.45)	0.05 [0.50]	500 (0.27, 0.27, 0.46)
(0.15, 0.15, 0)	0.46 [0.50]	426 (0.28, 0.28, 0.44)	0.50 [0.50]	500 (0.28, 0.28, 0.45)	0.49 [0.50]	500 (0.27, 0.27, 0.46)
(0, 0.27, 0)	0.77 [0.96]	375 (0.17, 0.40, 0.43)	0.80 [0.98]	500 (0.15, 0.41, 0.44)	0.77 [0.86]	500 (0.12, 0.41, 0.47)
(0.1, 0.27, 0)	0.76 [0.88]	377 (0.21, 0.36, 0.43)	0.79 [0.91]	500 (0.19, 0.37, 0.44)	0.74 [0.75]	500 (0.17, 0.37, 0.47)
(0.15, 0.27, 0)	0.78 [0.80]	375 (0.23, 0.34, 0.43)	0.81 [0.83]	500 (0.21, 0.35, 0.44)	0.77 [0.69]	500 (0.19, 0.34, 0.47)
(−0.1, 0.27, 0)	0.79 [0.99]	374 (0.15, 0.42, 0.43)	0.81 [0.99]	500 (0.14, 0.42, 0.44)	0.82 [0.94]	500 (0.09, 0.44, 0.47)
(0.27, 0.27, 0)	0.88 [0.50]	340 (0.29, 0.29, 0.42)	0.89 [0.50]	500 (0.28, 0.28, 0.45)	0.90 [0.50]	500 (0.27, 0.27, 0.46)