Group Sequential Designs for Developing and Testing Biomarker-guided Personalized Therapies in Comparative Effectiveness Research

Tze Leung Lai[a], Olivia Yueh-Wen Liao[b], Dong Woo Kim[c,*]

[a] *Department of Statistics, Stanford University, Stanford, CA, USA*
[b] *Onyx Pharmaceuticals, South San Francisco, CA, USA*
[c] *Department of Electrical Engineering, Stanford University, Stanford, CA, USA*

Abstract

Biomarker-guided personalized therapies offer great promise to improve drug development and improve patient care, but also pose difficult challenges in designing clinical trials for the development and validation of these therapies. We first give a review of the existing approaches, briefly for clinical trials in new drug development and in more detail for comparative effectiveness trials involving approved treatments. We then introduce new group sequential designs to develop and test personalized treatment strategies involving approved treatments.

Keywords: adaptive randomization, biomarker classifiers, generalized likelihood ratio statistics, group sequential design, multiple testing, targeted therapies.

1. Introduction

The development of imatinib (Gleevec), the first drug to target the genetic effects of chronic myeloid leukemia (CML) while leaving healthy cells unharmed, has revolutionized the treatment of cancer, leading to hundreds of kinase inhibitors and other targeted drugs that are in various stages of development in the anticancer drug pipeline. However, most new targeted treatments have resulted in only modest clinical benefit, with less than 50% remission rates and less than one year of progression-free survival. While the targeted treatments are devised to attack specific targets, the "one size fits all" treatment regimens commonly used may have diminished their effectiveness. In contrast, trastuzumab (Herceptin), which treats only patients with HER-2 positive metastatic breast cancer, has better remission rate and longer progression-free survival because it targets the "right" patient population. Genome-guided and risk-adapted personalized therapies of this kind are expected to substantially improve the effectiveness of these treatments.

Although personalized therapies that are tailored for individual patients have great promise to improve drug development and patient care, there are challenges in designing clinical trials for the development and

[*] Corresponding author at: Department of Electrical Engineering, Stanford University, 350 Serra Mall, Stanford, CA 94305-9505, USA, Tel: +1 650 796 9538

Email address: dwkim88@stanford.edu (Dong Woo Kim)

validation of these therapies because traditional trial designs often require large sample sizes that far exceed practical constraints on funding and study duration. Adaptive designs have been proposed to overcome these challenges in new drug development for regulatory approval. There are two important preliminaries in designing a phase III clinical trial for such drugs. One is to identify the biomarkers that are predictive of response, and the other is to develop a biomarker classifier that identifies patients who are sensitive to the treatment, denoted Dx+. An example is Herceptin, for which strong evidence of the relationship between the biomarker, HER2, and the drug effect was found early and led to narrowing the patient recruitment to HER2-positive patients in the phase III trial. In the ideal setting that the biomarker classifier can partition the patient population into drug-sensitive (Dx+) and drug-resistant (Dx-) subgroups, it is clear that Dx- patients should be excluded from the clinical trial. In practice, however, the cut-point for the Dx+ group is often based on data from early phase trials with relatively small sample sizes and has substantial statistical uncertainty (variability). Thus, a dilemma arises at the design stage of the phase III trial. Should the trial only recruit Dx+ patients who tend to have larger effect size, or should it have broad eligibility from the entire intended-to-treat (ITT) patient population but a diluted overall treatment effect size? The former has the disadvantage of an overly stringent exclusion criterion that misses a large fraction of patients who can benefit from the treatment if the classifier imposes relatively low false positive rate for Dx+ patients, while the latter has the disadvantage of ending up with an insignificant treatment effect by including patients that do not benefit from the treatment. To address this dilemma in the context of a phase III trial with a time-to-event endpoint, Brannath et al. [1] propose a two-stage trial design, in which the selection of the ITT or Dx+ population is performed based on conditional power at the first interim analysis. For the final analysis, a weighted combination of the second-stage p-value (based on the second-stage data) and the first-stage p-value, together with Simes' step-up procedure [2] to adjust for multiple testing, are used to ensure that the adaptive test maintains the prescribed type I error of the phase III trial. Jenkins et al. [3] extend the design of Brannath et al. to a phase II-III trial in which the phase II trial has a short-term survival endpoint that is used to select the ITT or Dx+ population for the phase III trial with a long-term survival endpoint. Earlier Wang et al. [4] have introduced a similar design for normally distributed outcomes. The basic idea underlying these adaptive designs is to use a weighting scheme of the form $S_1 + \gamma S_2$ that combines the first-stage and second-stage test statistics S_1 and S_2 or to choose the critical value of the Studentized second-stage statistic as some function of that of the first-stage to preserve the type I error probability; see [5, Section 8.1.2].

The main focus of this paper is on designing clinical trials for the development and validation of personalized therapies based on approved cancer treatments, which usually have well-understood molecular targets, mechanisms of action, and mechanisms of resistance. It is natural to try to use this information in conjunction with the patient's biomarkers that can predict sensitivity or resistance to the treatments, thereby developing a biomarker-guided strategy (BGS) to personalize treatment selection for the individual patients. After a review of previous methods in the literature, we introduce new group sequential designs in Section 2. Statistical inference in these designs is also discussed, and Section 3 demonstrates their advantages in simulation studies after providing implementation details. Section 4 gives further discussion and concluding remarks.

2. Development and Validation Trials for Biomarker-Guided Personalized Therapies

2.1. Review of existing approaches

Simon [6] has considered the development of biomarker classifiers for treatment selection and the design of validation trials for comparing a BGS to "standard of care" (SOC) that does not use the biomarkers to select treatments. For the validation trial, which he regards as an analog of a phase III trial, he shows that the *biomarker-strategy design* which randomizes patients to BGS and SOC is inefficient and proposes an *enrichment design* as an alternative. He also points out that development studies of the BGS "are often based on a convenience sample of patients for whom tissue is available but who are heterogeneous with regard to treatment and stage," and have the goal of developing a genomic classifier and evaluating its predictive accuracy by split-sample methods or cross-validation. The estimated predictive accuracy can be used to determine whether the classifier "is promising and worthy of phase III evaluation," analogous to phase II clinical trials. A difficulty with this approach is that the convenience sample comes from observational studies which have "no specific eligibility criteria, no primary endpoint or hypotheses and no defined analysis plan," but which often involve "multiple biomarkers to evaluate, multiple ways of measuring and combining the candidate biomarkers." Although it would be desirable to base the development of BGS on data from well designed clinical trials, it is difficult to obtain funding for such trials in practice. On the other hand, if the estimated predictive accuracy for the BGS developed from the convenience sample shows promise, then it may be possible to obtain funding for the validation trial. This is similar to phase I and II cancer trials that are single-arm and limited to relatively small sample sizes. Only after the phase II trial provides significant results showing that the new treatment has better response rate than some historical control rate can a randomized phase III trial with a survival endpoint be conducted. The limitations of these designs are discussed by Lai et al. [7] who point out in particular that the data that suggest the BGS "are preliminary and do not provide a uniform level of confidence in the recommendations made in each stratum."

Recognizing these limitations of the BGS developed, Lai et al. [7] propose to test in the validation trial not only the strategy null hypothesis defined by the BGS but also an intersection null hypothesis

$H_0 : p_{j1} = \cdots = p_{jK}$ for $1 \leq j \leq J$ in the case of J biomarker-classified patient subgroups and K treatments to choose from, where p_{jk} denotes the response rate of the jth subgroup to the kth treatment. Rejection of H_0 implies that there is some biomarker strategy, not necessarily the BGS set up for validation, that has better response rate than random assignment of the K treatments. If the biomarker strategy coincides with the BGS, this already validates the BGS. Even if it is not the case and the strategy null hypothesis is not rejected, the biomarker strategy that rejects H_0 would guide further development. In this way, the validation trial can be used not only to test the BGS but also to continue learning biomarker strategies from the clinical trial data. The strategy null hypothesis is $H_0^* : \sum_{j=1}^{J} \pi_j \left(P_j - \bar{q}_j\right) \leq 0$, where π_j is the prevalence of subgroup j, P_j is the average response of patients in subgroup j to the treatment recommended by the BGS and $\bar{q}_j$ is that to the treatments not recommended by the BGS, which is what an enrichment design attempts to test. As pointed out by Lai et al. [7], H_0^* represents a "hypothetical version" of SOC that assumes equal probabilities of choosing the K treatments in a biomarker subgroup, "lacking a true representation of a physician's choice condition."

Mandrekar and Sargent [8] give a review of designs of clinical trials for predictive biomarker validation in the context of real trials, and discuss their merits and limitations. In particular, they consider the "biomarker-stratified design" that randomizes patients to treatments within each biomarker class and focuses on the treatment-marker interaction in the analysis plan, with the MARVEL (marker validation of erlotinib in lung cancer) study as an example for which the sample size is prospectively specified separately for each biomarker class. They also describe prospectively specified analysis of data from a previously conducted RCT comparing treatments, but point out that "while a well conducted retrospective validation study may be accepted as a marker validation strategy in certain instances, the gold standard for predictive marker validation continues (appropriately) to be a prospective RCT."

A Bayesian alternative to frequentist testing of BGS is described by Zhou et al. [9] and Lee et al. [10] for the BATTLE (Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination) trial of personalized therapies for non-small cell lung cancer (NSCLC). As pointed out by [11, pp. 45-46] concerning the biomarker classifiers, "the signaling pathways and targeted agents were selected on the basis of the highest scientific and clinical interest at the time (2005)" and included EGFR mutation/copy number amplification, KRAS/BRAF mutation, VEGF/VEGFR expression and RXR/CyclinD1 expression, together with the recommended targeted agent for each; see Fig. 1 and refs. 9-12 of [11]. Although this provides a BGS similar to Simon's framework, the BATTLE trial uses an adaptive randomization scheme to select $K = 4$ treatments for $n = 255$ NSCLC patients belonging to $J = 5$ biomarker classes, one of which contains patients whose biomarker scores are all negative. Let y_{mjk} denote the indicator variable of disease control, which is defined by progression-free survival at 8 weeks after treatment, of the mth patient in class j receiving treatment k. The adaptive randomization scheme is based on a Bayesian probit model for $p_{jk} = P\left(y_{mjk} = 1\right) = P\left(\xi_{mjk} > 0\right)$, where ξ_{mjk} is assumed to be a latent normal random variable with variance 1 and mean $\mu_{jk} \sim \mathcal{N}\left(\phi_k, \sigma^2\right)$ such that $\phi_k \sim \mathcal{N}\left(0, \tau^2\right)$. Large values of τ^2 in the hierarchical Bayesian model can be used to approximate a vague prior. The posterior mean $\gamma_{jk}^{(t)}$ of p_{jk} given all the observed indicator variables up to time t can be computed by Gibbs sampling. Letting $\hat{\gamma}_{jk}^{(t)} = \max\left(\gamma_{jk}^{(t)}, 0.1\right)$, the randomization proportion for a patient in the jth class to receive treatment j at time $t+1$ is proportional to $\hat{\gamma}_{jk}^{(t)}$. Moreover, a refinement of this scheme allows suspension of treatment k from randomization to a biomarker subgroup.

The results of the BATTLE trial are reported by Kim et al. [11, pp. 46-48, 52]. Despite applying the Bayesian approach to adaptive randomization, "standard statistical methods (used in the Results section) included the Fisher's exact test for contingency tables and log-rank test for survival data" together with standard confidence intervals based on normal approximations, without adjustments for Bayesian adaptive randomization (AR) and possible treatment suspension, even though Zhou et al. [9] have noted that "one known ramification of the AR design is that it results in biased estimates due to dependent samples." The overall 8-week disease control rate (DCR) using the biomarker-guided AR scheme was 46%, compared to "the historical 30% DCR estimate in similar patients (ref. 14)", showing that the "learn-as-we-go" approach in Bayesian AR can indeed "leverage accumulating patient data to improve the treatment outcome" by "allowing more patients to be assigned to more effective therapies and fewer patients to be assigned to less effective therapies."; see [11, pp. 46, 52].

Note that unlike Simon's enrichment design that randomizes patients to SOC and the BGS to be validated, the BATTLE design aims at showing that the AR treatment assignment has higher DCR than some historical estimate of the DCR of SOC. In their discussion, Kim et al. [11, pp. 49-50] describe what they have learned from the BATTLE trial for a future BATTLE-2 trial, which will use EGFR mutations rather than EGFR mutation/copy number to narrow the biomarker subgroup because "EGFR mutations were far more predictive" and which will not use RXR that "had little, if any, predictive value in optimizing treatments." In their framework, AR provides a design for simultaneously treating patients with a given set of approved targeted agents based on the patients' biomarker profiles, and learning the treatment allocation rule from accumulating data.

The preceding paragraph shows that the BATTLE and BATTLE-2 trials share the philosophy of the classical *multi-arm bandit problem*. Suppose there are K treatments of unknown efficacy to be chosen sequentially to treat a large class of n patients. How should we allocate the treatments to maximize the mean treatment effect? Lai and Robbins [12] and Lai [13] consider the problem in the setting where the treatment effect has a density function $f\left(x; \theta_k\right)$ for the kth treatment, where the θ_k are unknown parameters. There is an apparent dilemma between the need to learn the unknown parameters and the objective of allocating patients to the best treatment to maximize the total treatment effect $S_n = X_1 + \cdots + X_n$ for the n patients. If the θ_k were known, then the optimal rule would use the treatment with parameter $\theta^* = \arg\max_{1 \le k \le K} \mu\left(\theta_k\right)$, where $\mu\left(\theta\right) = \mathbb{E}_\theta\left(X\right)$. In ignorance of θ_k, Lai and Robbins [12] define the *regret* of an allocation rule by

$$R_n\left(\theta\right) \quad = \quad n\mu\left(\theta^*\right) - \mathbb{E}_\theta\left(S_n\right) = \sum_{k:\mu(\theta_k)<\mu(\theta^*)} \left(\mu\left(\theta^*\right) - \mu\left(\theta_k\right)\right) \mathbb{E}_\theta T_n\left(k\right),$$

where $T_n\left(k\right)$ is the number of patients receiving treatment k. They show that adaptive allocation rules can be constructed to attain the asymptotically minimal order of $\log n$ for the regret, in contrast to the regret of order n for the traditional equal randomization rule that assigns patients to each treatment with equal probability $1/K$. A subsequent refinement by Lai [13] shows the relatively simple rule that chooses the treatment with the largest upper confidence bound $U_k^{(n)}$ for θ_k to be asymptotically optimal if the upper confidence bound at stage n, with $n > k$, is defined by

$$U_k^{(n)} \quad = \quad \inf\left\{\theta \in A \ : \ \theta \ge \hat{\theta}_k \text{ and } 2T_n\left(k\right) I\left(\hat{\theta}_k, \theta\right) \ge h^2\left(T_n\left(k\right)/n\right)\right\},$$

where $\inf \emptyset = \infty$, A is some open interval known to contain θ, $\hat{\theta}_k$ is the maximum likelihood estimate of θ_k. $I\left(\theta, \lambda\right)$ is the Kullback-Leibler information number, and the function h has a closed-form approximation. For the first K stages, the K treatments are assigned successively. It is noted in [14, p. 97] that the upper confidence bound $U_k^{(n)}$ corresponds to inverting a generalized likelihood ratio (GLR) test based on the GLR statistic $T_n\left(k\right) I\left(\hat{\theta}_k, \theta\right)$ for testing $\theta_k = \theta$.

2.2. An adaptive design combining multiple objectives

The multi-arm bandit problem has the same "learn-as-we-go" spirit of the BATTLE trial and focuses on attaining the best response rate for patients in the trial. However, such a trial does not establish which treatment is the best for future patients, with a guaranteed probability of correct selection. We now describe a group sequential design for jointly developing and testing treatment recommendations for biomarker classes, while using multi-armed bandit ideas to provide sequentially optimizing treatments to patients in the trial. Thus, the design has to fulfill multiple objectives, which include (a) treating accrued patients with the best (yet unknown) available treatment, (b) developing a treatment strategy for future patients, and (c) demonstrating that the strategy developed indeed has better treatment effect than the historical mean effect of SOC plus a predetermined threshold. In a group sequential trial, sequential decisions are made only at times of interim analysis. Let n_i denote the total sample size up to the time of the ith analysis, $i = 1, \cdots, I$, so that n_I is the total sample size by the scheduled end of the trial, and let n_{ij} be the total sample size from biomarker class j up to the time of the ith analysis, hence $n_i = \sum_{j=1}^{J} n_{ij}$. Because of the need for informed consent, the treatment allocation that uses the aforementioned upper confidence bound rule is no longer appropriate. It is unlikely for patients to consent to being assigned to a seemingly inferior treatment for the sake of collecting more information to ensure that it is significantly inferior (as measured by the upper confidence bounds). Instead, randomization in a double blind setting is required, and the randomization probability $\pi_{jk}^{(i)}$, determined at the ith interim analysis, of assigning a patient in group j to treatment k cannot be too small to suggest obvious inferiority of the treatments being tried, that is,

$$\pi_{jk}^{(i)} \geq \epsilon \text{ for some } 0 < \epsilon < 1/K.$$

We now describe the adaptive randomization rule. The unknown mean treatment effect μ_{jk} of treatment k in biomarker class j can be estimated by the sample mean $\hat{\mu}_{ijk}$ at interim analysis i. Let $k_j = \arg\max_k \mu_{jk}$, which can be estimated by $\hat{k}_{ij} = \arg\max_k \hat{\mu}_{ijk}$ at the ith interim analysis. Analogy with multi-arm bandit theory suggests assigning the highest randomization probability to treatment $\hat{k}_{ij}$ and randomizing to the other available treatments in biomarker class j with probability ϵ. Because the randomization probabilities are only updated at interim analyses in a group sequential design and because $\hat{k}_{ij}$ may fluctuate over i among treatments whose treatment effects do not differ by more than δ_{ij}, it is more stable to lump these “nearby” treatments into the set

$$\mathcal{H}_{ij} = \left\{k \in \mathcal{K}_{ij} : \left|\hat{\mu}_{ij}^{*} - \hat{\mu}_{ijk}\right| \leq \delta_{ij}\right\}, \tag{1}$$

where $\hat{\mu}_{ij}^{*} = \hat{\mu}_{ij\hat{k}_{ij}}$ and $\mathcal{K}_{ij}$ is the set of available treatments in biomarker class j at interim analysis i. The randomization probabilities $\pi_{jk}^{(i)}$ are therefore determined at the ith interim analysis by

$$\pi_{jk}^{(i)} = \epsilon \text{ for } k \in \mathcal{K}_{ij} \backslash \mathcal{H}_{ij}, \ \pi_{jk}^{(i)} = \left(1 - \left|\mathcal{K}_{ij} \backslash \mathcal{H}_{ij}\right| \epsilon\right) / \left|\mathcal{H}_{ij}\right| \text{ for } k \in \mathcal{H}_{ij}, \tag{2}$$

where we use $|A|$ to denote the number of elements of a finite set A. Equal randomization is used up to the first interim analysis. In Section 3.1, we carry out a simulation study of the performance of this design for the objective of treating patients in the trial with the best available treatments, and compare it with an alternative adaptive randomization scheme proposed by Zhou et al. [9] for the BATTLE trial and modified by Lai et al. [7].

Besides treating patients in the trial with the best available treatment, the group sequential design can also be used to address testing and inference questions, with guaranteed error probabilities, that are of basic interest to personalized treatment selection for future patients based on their biomarkers. We use GLR statistics and modified Haybittle-Peto stopping rules introduced by Lai and Shih [15] to include early elimination of significantly inferior treatments from a biomarker class. Following [13] and [15], we assume an exponential family of distributions for the treatment effects, with density function $f_\theta(x) = e^{\theta x - \psi(x)}$ with respect to some probability measure ν on $\Theta = \left\{\theta : \int e^{\theta x} d\nu(x) < \infty\right\}$, where θ depends on the treatment k and biomarker class j and will be defined by θ_{jk}. In the exponential family, the mean μ is $\psi'(\theta)$ and $\theta = \theta_\mu = (\psi')^{-1}(\mu)$ since ψ^{-1} is a smooth increasing function on Θ. The maximum likelihood estimate (MLE) of μ is the sample mean $\hat{\mu}$, and we let $\hat{\mu}_{ijk}$ denote the average treatment effect of treatment k in biomarker class j at interim analysis i. The Kullback-Leibler information number is

$$I(\mu, \mu') = \mathbb{E}_{\theta_\mu}\left\{\log\left[f_{\theta_\mu}(X)/f_{\theta_{\mu'}}(X)\right]\right\} = (\theta_\mu - \theta_{\mu'})\mu - [\psi(\theta_\mu) - \psi(\theta_{\mu'})].$$

Let n_{ijk} be the total sample from biomarker class j receiving treatment k up to the ith interim analysis, so $n_{ij} = \sum_{k=1}^K n_{ijk}$. Let

$$\begin{aligned} l_j^i(k, k') &= n_{ijk}\left\{\hat{\mu}_{ijk}\theta_{\hat{\mu}_{ijk}} - \psi\left(\theta_{\hat{\mu}_{ijk}}\right)\right\} + n_{ijk'}\left\{\hat{\mu}_{ijk'}\theta_{\hat{\mu}_{ijk'}} - \psi\left(\theta_{\hat{\mu}_{ijk'}}\right)\right\} \\ &\quad - (n_{ijk} + n_{ijk'})\left\{\bar{\mu}\theta_{\bar{\mu}} - \psi(\theta_{\bar{\mu}})\right\}, \end{aligned} \tag{3}$$

where $\bar{\mu} = (n_{ijk}\hat{\mu}_{ijk} + n_{ijk'}\hat{\mu}_{ijk'})/(n_{ijk} + n_{ijk'})$. Let $\mu_{jk} = \psi'(\theta_{jk})$. As shown by Brezzi and Lai [14, p. 103] who also recommend constraining the MLE to a compact subset of $\psi(\Theta)$ on which ψ'' is uniformly continuous, $l_j^i(k, k')$ is the GLR statistic at the ith interim analysis for testing the null hypothesis $\mu_{jk} = \mu_{jk'}$ and plays a basic role in constructing the upper confidence bound rule in the multi-arm bandits from the exponential family.

We now propose an elimination scheme based on the GLR statistic (3) with a guaranteed probability of $1 - \alpha$ that the best for each biomarker class is not eliminated. At the ith analysis $(1 \leq i \leq I)$, treatment $k \neq \hat{k}_{ij}$ is eliminated for the biomarker class j if $l_j^i\left(k, \hat{k}_{ij}\right) \geq a_\alpha$. The computation of a_α is described in Section 3.2. This elimination scheme is also related to the second objective of the trial, which is inference, at the end of the trial, on which treatment strategy is best for future patients. To accomplish the above objective, we use subset selection ideas from the selection and ranking literature [16, 17], in which there are two approaches to selecting the best of K treatments with guaranteed probability of correct selection. One is the “indifference zone” approach, which guarantees that the probability of correctly selecting the best treatment exceeds $1-\alpha$ when the largest mean effect differs from the second largest by at least δ. In practice, however, one does not have any idea about the distance between the largest and second largest means. To address this difficulty, Chan and Lai [18] consider a stronger constraint that the probability of selecting a treatment whose mean effect is within δ of the largest is at least $1 - \alpha$. They also develop an efficient fully sequential procedure to attain this. Their procedure, however, cannot be extended to a group sequential design in which there is a prescribed upper bound on the total number of observations. An alternative to the indifference zone approach is subset selection, for which the goal is to select a subset of treatments, with a guaranteed probability of at least $1-\alpha$ that it contains the best treatment. In this approach, one also wants the expected size of the selected subset to be as small as possible in some sense.

We extend the subset selection approach to the setting of J biomarker classes in a group sequential design. Using the elimination scheme described in the proceeding paragraph, let $\mathcal{K}_{ij}$ be the set of surviving treatments for class j at the ith interim analysis. When $\mathcal{K}_{ij}$ consists only of $\hat{k}_{ij}$, the trial recommends using treatment $\hat{k}_{ij}$ for future patients. For notational simplicity, $\mathcal{K}_{Ij}$ at the Ith analysis by the trial's scheduled end will be denoted by $\mathcal{K}_j$, which may contain two or more treatments. Similarly we denote $\hat{\mu}_{Ijk}$ by $\hat{\mu}_{jk}$. The recommended set of treatments for class j is $\mathcal{K}_j$, with an overall probability guarantee of $1-\alpha$ to contain the best treatments for all classes. Whereas the probability α of incorrectly eliminating the best treatment in subset selection corresponds to type I error in hypothesis testing, $P\left(\bigcup_{j=1}^{J} B_j\left(\bar{\delta}\right) \neq \emptyset\right)$ is an analog of traditional type II error, where $B_j\left(\bar{\delta}\right) = \left\{k \in \mathcal{K}_j : \mu_{jk} < \max_{1\le k' \le K} \mu_{jk'} - \bar{\delta}\right\}$.

The third objective of this trial, which is to demonstrate that the developed treatment strategy improves the mean treatment effect of SOC by a prescribed margin, amounts to testing the null hypothesis H_0^* : $\sum_{j=1}^{J} \pi_j \max_{1\le k\le K} \mu_{jk} \le \gamma$, where π_j is the prevalence of biomarker class j and γ is the historical treatment effect of SOC plus a prescribed margin. The GLR statistic for testing H_0^* is

$$L^I \quad = \quad \sum_{j=1}^{J}\sum_{k=1}^{K} n_{Ijk}\left(\hat{\mu}_{jk}\theta_{\hat{\mu}_{jk}} - \psi\left(\theta_{\hat{\mu}_{jk}}\right)\right) - \sum_{j=1}^{J}\sum_{k=1}^{K} n_{Ijk}\left(\tilde{\mu}_{jk}\theta_{\tilde{\mu}_{jk}} - \psi\left(\theta_{\tilde{\mu}_{jk}}\right)\right),$$

where $\tilde{\mu}_{jk}$ is the MLE of μ_{jk} under the constraint $\sum_{j=1}^{J} \hat{\pi}_j \max_{1\le k\le K} \mu_{jk} \le \gamma$, in which $\hat{\pi}_j$ is the observed prevalence of biomarker class j at the Ith (i.e., terminal) analysis. With a prescribed type I error of $\tilde{\alpha}$, the GLR test rejects H_0^* if

$$L^I > d_{\tilde{\alpha}} \quad \text{and} \quad \sum_{j=1}^{J} \hat{\pi}_j \max_{1\le k\le K} \hat{\mu}_{jk} > \gamma. \tag{4}$$

The computation of $d_{\tilde{\alpha}}$ is described in Section 3.3.

3. Implementation and Simulation Studies

3.1. Comparison of adaptive randomization schemes

We first present a simulation of the performance of the preceding group sequential trial in treating patients who have been accrued to the trial, and its performance with respect to the inferential objectives relevant to future patients will be studied in Section 3.4. The adaptive randomization rule in the second paragraph of Section 2.2, denoted by AR1, does not involve elimination in the subsequent paragraphs, which will be studied in Section 3.4 and 3.5. It is a group sequential modification of the fully sequential upper confidence bound (UCB) allocation rule that has been shown to minimize asymptotically the regret in the multi-arm bandit problem, as we have noted earlier. Accordingly the simulation study will compare AR1, which uses $\epsilon = 0.1$ in (2), against the benchmark UCB rule in the response rate of patients receiving each treatment (including the best and the worst) for each biomarker class. Note that AR1 is quite different from the Bayesian adaptive allocation rule in the BATTLE trial described in Section 2.1, which assumes a hierarchical Bayesian probit model on the response rate p_{jk} of treatment k for biomarker class j and which uses randomization probabilities proportional to the posterior means of p_{jk} for different treatments in each biomarker class. Since these posterior distributions, evaluated by Markov chain Monte Carlo methods, are too computationally intensive for replicating them many times in a simulation study, we follow [7] and replace the posterior mean at the ith interim analysis by the maximum likelihood estimate $\hat{p}_{jk}^{(i)}$ of p_{jk}, under the constraint that p_{jk} has a priori bounds $b = 0.05$ and $B = 0.95$. The adaptive randomization rule that uses randomization probabilities proportional to $\hat{p}_{jk}^{(i)}$ between interim analyses i and $i+1$, denoted AR2, is also considered for comparison. In addition, we follow [19] and choose $\delta_{ij} = n_{ij}^{-2/5}$ so that $\sqrt{n_{ij}}\delta_{ij} \to \infty$ for biomarker class j at interim analysis i.

The simulation study considers $n = 1000$ and the cases $K = J = 3$ in Table 1 and $K = 4$, $J = 3$ or 4 in Table 2. In addition, it assumes $I = 5$ analyses (including the interim and final analyses), with equal group sizes $n_i - n_{i-1} = 200$ $(i = 1, \cdots, 5, n_0 = 0)$. Table 1 studies the following scenarios for the response rates p_{jk}, in which the class sizes are proportional to $3:2:1$ for $j = 1, 2, 3$.

$$S1: \quad p_{jk} = 0.7 \text{ for } j = k,\ p_{jk} = 0.2 \text{ for } j \neq k.$$

$$S2: \quad p_{jk} = 0.7 \text{ for } j = k,\ p_{12} = p_{23} = p_{31} = 0.5,\ p_{13} = p_{21} = p_{32} = 0.2.$$

$$S3: \quad p_{jk} = 0.7 \text{ for } j = k,\ p_{12} = p_{23} = p_{31} = 0.65,\ p_{13} = p_{21} = p_{32} = 0.2.$$

Thus, each biomarker class has a unique best treatment that is substantially better than other treatments in S1, there are treatments with moderate effectiveness between the best one and the worst ones for each biomarker class in S2, and there is a treatment which is close to the best for each biomarker class in S3. Table 2 considers scenarios S4 and S5 that are similar to the scenario 1 and 2 of the first simulation study of [7], and another scenario S6 similar to that in the BATTLE trial with the RXR/CyclinD1 class (that has a small size) and the all-negative biomarker class removed.

$S4:$ $p_{11} = p_{22} = 0.6$, $p_{33} = p_{44} = 0.75$, $p_{jk} = 0.3$ for $j,k \in \{1,2,3,4\} \times \{1,2\}$, $j \neq k$

$p_{jk} = 0.1$ for $j,k \in \{3,4\} \times \{1,2,3,4\}$, $j \neq k$; class size proportions are $15:20:30:25$.

$S5:$ $p_{11} = 0.8$, $p_{22} = p_{33} = p_{44} = 0.6$, $p_{jk} = 0.3$ for $j \neq k$

class size proportions are $15:20:30:25$.

$S6:$ $p_{13} = 0.6$, $p_{1k} = 0.4$ for $k \in \{1,2,4\}$; $p_{21} = p_{22} = 0.1$, $p_{23} = 0.3$, $p_{24} = 0.8$;

$p_{31} = p_{32} = 0.4$, $p_{33} = 0.1$, $p_{34} = 0.6$; class size proportions are $35:15:50$.

The results for each scenario in Tables 1 and 2 are based on 10000 simulations. For each allocation rule, besides the overall mean response of the $n = 1000$ subjects, the tables also give in parentheses the mean number of each (j,k) category of subjects in biomarker class j receiving treatment k and the mean response rate in this category. For each scenario in both tables, AR1 outperforms AR2 in terms of the overall mean response and the expected number of subjects receiving the best treatment in each biomarker class. Moreover, the benchmark UCB rule outperforms the adaptive randomization rules as expected but is inappropriate for applications to clinical trials that require informed consent and have operational difficulties in implementing fully sequential procedures.

3.2. Computation of a_α

The threshold a_α is determined by the constraint $P\,(\text{best treatment for some biomarker class is eliminated}) \leq \alpha$. Fix j and order the parameter configuration for the k treatments as $\theta_{j,[1]} \geq \cdots \geq \theta_{j,[K]}$. Assuming $\theta_{j,[1]} > \theta_{j,[2]}$, the event of eliminating the (unique) best treatment for biomarker class j is

$$A_j \quad = \quad \left\{ \max_{k \neq [1]} \left[l_j^i\left([1], k\right) \mathbf{1}_{\left\{\hat{\mu}_{ijk} > \hat{\mu}_{ij,[1]}\right\}} \right] \geq \alpha \text{ for some } 1 \leq i \leq I \right\}.$$

Letting $\theta_{j,[2]}$ approach $\theta_{j,[1]}$ implies that we can use $P_*\left(A_j\right)$ to bound the probability that the best treatment for biomarker class j is eliminated, where P_* is the probability measure satisfying $\theta_{j1} = \cdots = \theta_{jK}$ for all $1 \leq j \leq J$. Hence a_α can be determined by

$$\alpha = P_*\left(\bigcup_{j=1}^{J} A_j\right) = \sum_{j=1}^{J} P_*\left(A_j\right) - \sum_{j_1 < j_2} P_*\left(A_{j_1}\right) P_*\left(A_{j_2}\right) + \cdots + (-1)^{J+1} \prod_{j=1}^{J} P_*\left(A_j\right). \tag{5}$$

in which the last equality follows from the inclusion-exclusion principle and the independence of the events $A_1, \cdots, A_J$.

Table 1
Mean response rate and sample size (in parentheses) for scenarios S1-S3 involving $K = 3$ treatments

		Marker		Treatment		
		Class		1	2	3
S1	UCB	1		0.70 (485.8)	0.20 (7.1)	0.20 (7.1)
		2		0.20 (6.9)	0.70 (319.6)	0.20 (7.0)
		3		0.20 (6.6)	0.20 (6.6)	0.70 (153.4)
		Total	0.680 (1000)			
	AR1	1		0.70 (392.3)	0.20 (53.9)	0.20 (53.7)
		2		0.20 (37.5)	0.70 (258.2)	0.20 (37.5)
		3		0.20 (22.6)	0.20 (22.5)	0.70 (121.8)
		Total	0.586 (1000)			
	AR2	1		0.70 (290.5)	0.20 (104.7)	0.20 (104.9)
		2		0.20 (69.5)	0.70 (194.7)	0.20 (69.2)
		3		0.20 (34.2)	0.20 (34.5)	0.70 (97.9)
		Total	0.491 (1000)			
S2	UCB	1		0.70 (466.1)	0.50 (26.9)	0.20 (7.0)
		2		0.20 (6.9)	0.70 (300.3)	0.50 (26.2)
		3		0.50 (21.9)	0.20 (6.5)	0.70 (138.1)
		Total	0.675 (1000)			
	AR1	1		0.70 (332.5)	0.50 (114.1)	0.20 (53.6)
		2		0.20 (36.9)	0.70 (206.0)	0.50 (90.4)
		3		0.50 (54.8)	0.20 (21.4)	0.70 (90.3)
		Total	0.592 (1000)			
	AR2	1		0.70 (234.3)	0.50 (176.5)	0.20 (89.3)
		2		0.20 (59.3)	0.70 (156.4)	0.50 (117.6)
		3		0.50 (58.5)	0.20 (29.4)	0.70 (78.7)
		Total	0.541 (1000)			
S3	UCB	1		0.70 (360.0)	0.65 (133.0)	0.20 (6.9)
		2		0.20 (6.7)	0.70 (225.6)	0.65 (101.0)
		3		0.65 (58.5)	0.20 (6.3)	0.70 (102.1)
		Total	0.675 (1000)			
	AR1	1		0.70 (231.2)	0.65 (215.1)	0.20 (53.4)
		2		0.20 (36.1)	0.70 (153.3)	0.65 (144.2)
		3		0.65 (71.2)	0.20 (20.2)	0.70 (75.3)
		Total	0.624 (1000)			
	AR2	1		0.70 (214.8)	0.65 (201.4)	0.20 (83.9)
		2		0.20 (55.5)	0.70 (143.1)	0.65 (134.5)
		3		0.65 (67.4)	0.20 (27.6)	0.70 (71.7)
		Total	0.596 (1000)			

Table 2
Mean response rate and sample size (in parentheses) for scenarios S4-6 involving $K = 4$ treatments

		Marker Class		Treatment			
				1	2	3	4
S4	UCB	1		0.60 (125.7)	0.30 (13.7)	0.30 (13.6)	0.30 (13.6)
		2		0.30 (14.7)	0.60 (178.4)	0.30 (14.6)	0.30 (14.7)
		3		0.10 (4.9)	0.10 (4.9)	0.75 (318.5)	0.10 (4.9)
		4		0.10 (4.8)	0.10 (4.9)	0.10 (4.9)	0.75 (263.3)
		Total	0.647 (1000)				
	AR1	1		0.60 (69.8)	0.30 (32.2)	0.30 (32.3)	0.30 (32.3)
		2		0.30 (39.7)	0.60 (102.9)	0.30 (39.7)	0.30 (39.9)
		3		0.10 (30.3)	0.10 (30.4)	0.75 (242.4)	0.10 (30.4)
		4		0.10 (25.8)	0.10 (25.8)	0.10 (25.7)	0.75 (200.4)
		Total	0.518 (1000)				
	AR2	1		0.60 (63.4)	0.30 (34.7)	0.30 (34.5)	0.30 (34.2)
		2		0.30 (45.9)	0.60 (84.0)	0.30 (46.1)	0.30 (46.2)
		3		0.10 (41.9)	0.10 (42.0)	0.75 (207.2)	0.10 (42.0)
		4		0.10 (35.1)	0.10 (35.2)	0.10 (35.2)	0.75 (172.4)
		Total	0.469 (1000)				
S5	UCB	1		0.80 (147.7)	0.30 (6.3)	0.30 (6.3)	0.30 (6.3)
		2		0.30 (14.4)	0.60 (178.6)	0.30 (14.5)	0.30 (14.6)
		3		0.30 (15.5)	0.30 (15.5)	0.60 (287.0)	0.30 (15.4)
		4		0.30 (15.0)	0.30 (15.0)	0.30 (15.2)	0.60 (232.4)
		Total	0.583 (1000)				
	AR1	1		0.80 (104.3)	0.30 (20.7)	0.30 (20.8)	0.30 (20.8)
		2		0.30 (39.7)	0.60 (102.2)	0.30 (40.0)	0.30 (40.2)
		3		0.30 (51.3)	0.30 (51.4)	0.60 (179.3)	0.30 (51.5)
		4		0.30 (45.7)	0.30 (46.1)	0.30 (46.1)	0.60 (139.9)
		Total	0.479 (1000)				
	AR2	1		0.80 (72.9)	0.30 (31.1)	0.30 (31.3)	0.30 (31.2)
		2		0.30 (46.2)	0.60 (84.0)	0.30 (46.0)	0.30 (46.0)
		3		0.30 (69.4)	0.30 (69.3)	0.60 (125.0)	0.30 (69.8)
		4		0.30 (57.7)	0.30 (57.6)	0.30 (57.9)	0.60 (104.4)
		Total	0.431 (1000)				
S6	UCB	1		0.40 (26.2)	0.40 (26.0)	0.60 (272.1)	0.40 (25.8)
		2		0.10 (3.8)	0.10 (3.8)	0.30 (6.1)	0.80 (136.2)
		3		0.40 (28.2)	0.40 (28.3)	0.10 (7.0)	0.60 (436.4)
		Total	0.591 (1000)				
	AR1	1		0.40 (72.0)	0.40 (72.2)	0.60 (133.7)	0.40 (72.1)
		2		0.10 (14.8)	0.10 (14.8)	0.30 (20.2)	0.80 (100.3)
		3		0.40 (102.0)	0.40 (101.6)	0.10 (45.5)	0.60 (250.8)
		Total	0.493 (1000)				
	AR2	1		0.40 (79.5)	0.40 (79.5)	0.60 (111.9)	0.40 (79.3)
		2		0.10 (17.8)	0.10 (17.8)	0.30 (33.0)	0.80 (81.3)
		3		0.40 (130.8)	0.40 (130.8)	0.10 (53.3)	0.60 (185.0)
		Total	0.462 (1000)				

For fixed j, we can compute $P_*(A_j)$ in (5) by using recursive numerical integration as follows. Since $\theta_{j1} = \cdots = \theta_{jK}$, we can let $[1] = 1$ and approximate n_{ijk} by $(1 + o_p(1))\, n_{ij}/K$, as the adaptive randomization rule is asymptotically equivalent to equal randomization in this case. Moreover, the GLR statistic $l_j^i(1,k)$ can be approximated by

$$l_j^i(1,k) \quad = \quad (1 + o_p(1)) \frac{n_{ij}}{4K\psi''\left(\theta_{\mu_{j1}}\right)} \left(\hat{\mu}_{ijk} - \hat{\mu}_{ij1}\right)^2 = \frac{1}{2}\left(\Delta_{jk}^i\right)^2,$$

where $\Delta_{jk}^i = \left[n_{ij}/\left(2K\psi''\left(\theta_{\mu_{j1}}\right)\right)\right]^{1/2} \left(\hat{\mu}_{ijk} - \hat{\mu}_{ij1}\right)$, for $k \in \{2, \cdots, K\}$; see [5, p. 95]. Therefore

$$P_*(A_j) \quad \to \quad P_*\left\{\max_{1 \le k \le K} \Delta_{jk}^i \ge \sqrt{2a_\alpha} \text{ for some } 1 \le i \le I\right\}. \tag{6}$$

The above probability can be computed by applying the central limit theorem to $\sqrt{n_{ij}}\left(\Delta_{j2}^i, \cdots, \Delta_{jK}^i\right)$ that has independent increments in i. In particular, for $K = 3$, the conditional distribution of $\left(\Delta_{j2}^{i+1}, \Delta_{j3}^{i+1}\right)$ given $\left(\Delta_{j2}^i, \Delta_{j3}^i\right)$ is

$$\mathcal{N}\left(\sqrt{\frac{n_{ij}}{n_{i+1,j}}} \begin{bmatrix} \Delta_{j2}^i \\ \Delta_{j3}^i \end{bmatrix}, \frac{n_{i+1,j} - n_{ij}}{2n_{i+1,j}} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right). \tag{7}$$

Therefore the right-hand side of (6) can be computed by using recursive numerical integration; see [5, Sections 4.3.1 and 8.2.4] and [19, p. 452]. With this recursive procedure to compute $P_*(A_j)$, we can use bisection search to find the a_α that satisfies (5), noting that $P_*\left(\bigcup_{j=1}^J A_j\right)$ is non-increasing in a_α.

Instead of recursive numerical integration, $P_*(A_j)$ can alternatively be computed by Monte Carlo simulation of the multivariate normal Markov chain $\left(\Delta_{j2}^i, \cdots, \Delta_{jK}^i\right)$, $1 \le i \le I$. This is preferable to recursive numerical integration for $K > 3$; see [19, pp. 452-453] With $P_*(A_j)$ computed by Monte Carlo, we can again use bisection search to solve (5) for a_α.

3.3. Computation of $d_{\tilde{\alpha}}$

To compute the constrained MLE $\tilde{\mu}_{jk}$, note that the constraint $\sum_{j=1}^J \tilde{\pi}_j \max_{1 \le k \le K} \mu_{jk} \le \gamma$ is convex in the μ_{jk}. Since the log-likelihood function is concave, its maximizer $(\tilde{\mu}_{j1}, \cdots, \tilde{\mu}_{jK})$ subject to convex constraints can be computed by using constrained convex optimization solvers, such as `fmincon` with the "interior-point" option in MATLAB.

Since the function $g(\boldsymbol{\mu}) = \sum_{j=1}^J \pi_j \max_{1 \le k \le K} \mu_{jk}$ that defines the composite null hypothesis H_0^* is not smooth at the hyperplanes $\mu_{jk} = \mu_{jk'}$, $k \ne k'$, traditional likelihood theory that assumes a smooth region for the null hypothesis as in Section 4.2.4 of [5] does not apply to the GLR statistic L^I. In fact, $2L^I$ is no

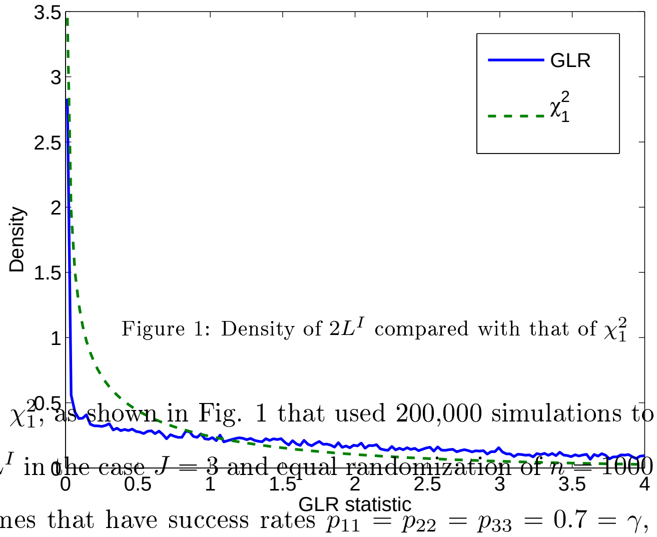

Figure 1: Density of $2L^I$ compared with that of χ_1^2

longer asymptotically χ_1^2, as shown in Fig. 1 that used 200,000 simulations to compute by Monte Carlo the density function of $2L^I$ in the case $J = 3$ and equal randomization of $n = 1000$ subjects to $K = 3$ treatments with Bernoulli outcomes that have success rates $p_{11} = p_{22} = p_{33} = 0.7 = \gamma$, $p_{jk} = 0.69$ for $j \neq k$. Fig. 1 corresponds to the case (C1) with $\gamma = 0.7$ in Table 4, for which we use corrections, due to Chernoff [20] and Self and Liang [21], of the χ_1^2 approximation to the null distribution of twice the GLR statistic for testing $g(\boldsymbol{\mu}) = \gamma$. Besides the central limit theorem, the main ingredient leading to the χ_1^2 approximation when g is smooth is the quadratic approximation of the GLR statistic around $\boldsymbol{\mu} = \boldsymbol{\mu}_o$ with $g(\boldsymbol{\mu}_o) = \gamma$. When the partial derivatives of g at $\boldsymbol{\mu}_o$ have jump discontinuities, creating a "kink" (local cone) of the type mentioned in [20] and [21] for the graph of the continuous function g near $\boldsymbol{\mu}_o$, the central limit theorem leads to the following limiting distribution of twice the GLR:

$$\inf_{\boldsymbol{\mu} \in C_0} \left\{ (\boldsymbol{Z} - \boldsymbol{\mu})^{'} (\boldsymbol{Z} - \boldsymbol{\mu}) \right\}, \tag{8}$$

where $\boldsymbol{Z}$ is multivariate standard normal, $\boldsymbol{I}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}_o}\right) = \boldsymbol{PDP}^{'}$ is the singular value decomposition of the Fisher information matrix, and C_0 is a cone with vertex at $\boldsymbol{D}^{1/2}\boldsymbol{P}^{'}\boldsymbol{\mu}_o$; see [21, p. 607]. In other words, the limiting distribution is the same as that of the GLR test $H_0 : \boldsymbol{\mu} \in C_0$ based on $\boldsymbol{Z}$; see [20].

For the special case of H_0^*, $g(\boldsymbol{\mu}_o) = \sum_{j=1}^{J} \pi_j \max_{1 \le k \le K} \mu_{jk}^o$ and $\boldsymbol{I}\left(\boldsymbol{\theta}_{\boldsymbol{\mu}_o}\right)$ is a diagonal matrix with diagonal elements $\psi^{''}\left(\theta_{\mu_{jk}^o}\right)$. Suppose $\max_{1 \le k \le K} \mu_{jk}^o$ is uniquely attained at $k = k_j$, for every j. Then there are no jump discontinuities of the gradient vector $\partial g / \partial \boldsymbol{\mu}$ at $\boldsymbol{\mu}_o$, and therefore the usual χ_1^2 approximation to $2L^I$ still applies as $n \to \infty$. In other words, C_0 in (8) can be expressed as a linear constraint of the form

$\sum_{j=1}^{J} \pi_j \mu_{j,k_j} = 0$. On the other hand, if $\max_{1 \le k \le K} \mu_{jk}^o$ is attained at m_j treatments $k^1, \cdots, k^{m_j}$, then C_0 is tantamount to the constraint $\sum_{j=1}^{J} \pi_j \max\left(\mu_{j,k^1}, \cdots, \mu_{j,k^{m_j}}\right) = 0$. Using the approximation (8) to the null distribution of $2L^I$, $d_{\tilde{\alpha}}$ can be determined as the $1 - \tilde{\alpha}$ quantile of (8) if π_j, m_j and $k^1, \cdots, k^{m_j}$ are specified. Although these parameters are not known a priori, π_j can be replaced by its consistent estimate $\hat{\pi}_j$ in the determination of $d_{\tilde{\alpha}}$. However, because $(\tilde{\mu}_{j1}, \cdots, \tilde{\mu}_{jK})$ can differ by $O_p(1/\sqrt{n})$ from $\left(\mu_{j1}^o, \cdots, \mu_{jK}^o\right)$, which may not belong to H_0^*, $\left(m_j, k^1, \cdots, k^{m_j}\right)$ cannot be estimated consistently. Feder [22] has derived the distribution of twice GLR when $\boldsymbol{\mu}_o$ is within $O_p(1/\sqrt{n})$ of the boundary $g(\boldsymbol{\mu}_*) = \gamma$, showing that it is basically a "noncentral version" of (8). For the special case of H_0^*, we can use this result to derive the following conservative estimate of $\left(m_j, k^1, \cdots, k^{m_j}\right)$.

Let $\tilde{k}_j = \max_{k \in \mathcal{K}_j} \tilde{\mu}_{jk}$ and let $\tilde{m}_j$ be the number of treatments $k \in \mathcal{K}_j$ such that $\tilde{\mu}_{jk} \ge \tilde{\mu}_{j,k_j} - \delta_j$, where $\delta_j = \delta_{Ij}$ is introduced in (1) and the first paragraph of Section 3.1. Note that H_0^* involves $\max(\mu_{j1}, \cdots, \mu_{jK})$, which is $\ge \max_{k \in \mathcal{K}} \mu_{jk}$ for any subset $\mathcal{K}$ of $\{1, \cdots, K\}$. Hence, choosing $\mathcal{K}$ to be the subset $\mathcal{K}_j$ of surviving treatments would lead to a conservative estimate of $d_{\tilde{\alpha}}$. Let $\tilde{k}^{(1)}, \cdots, \tilde{k}^{(\tilde{m}_j)}$ denote these treatments. Since $\sqrt{n_j}\delta_j \to \infty$, it follows from [22] that replacing $\left(\pi_j, m_j, k^1, \cdots, k^{m_j}\right)$ in $d_{\tilde{\alpha}}$ by $\left(\hat{\pi}_j, \tilde{m}_j, \tilde{k}^{(1)}, \cdots, \tilde{k}^{(\tilde{m}_j)}\right)$ for $1 \le j \le J$ yields an estimate $\tilde{d}_{\tilde{\alpha}}$ that is $\ge d_{\tilde{\alpha}} + o_p(1)$. Therefore, we compute $d_{\tilde{\alpha}}$ somewhat conservatively by using Monte Carlo simulations of (8), in which π_j, m_j, k^1, $\cdots$, k^{m_j} in the constraint $\sum_{j=1}^{J} \pi_j \max\left(\mu_{j,k^1}, \cdots, \mu_{j,k^{m_j}}\right) = 0$ are replaced by $\hat{\pi}_j$, $\tilde{m}_j$, $\tilde{k}^{(1)}$, $\cdots$, $\tilde{k}^{(\tilde{m}_j)}$ for $1 \le j \le J$.

3.4. A simulation study of inferences for future patients

In this section we present a simulation study of the inferential procedures in Section 2.2 in the case of $K = J = 3$, with $k = j$ being the best treatment for biomarker class j. We take $\alpha = 0.1$ and $\tilde{\alpha} = 0.05$. The class sizes are proportional to $5 : 4 : 1$ for $n = 1000$ subjects. We assume that $p_{jk} = 0.7$ if $j = k$ and the following five configurations of parameters p_{jk} for $j \ne k$:

($C1$) $p_{jk} = 0.69$ for $j \ne k$: Although each biomarker class has a unique best treatment, other treatments are almost as good.

($C2$) $p_{12} = p_{23} = p_{31} = 0.69$, $p_{13} = p_{21} = p_{32} = 0.2$: For each biomarker class, the best treatment has a close competitor, but the remaining treatment is substantially worse.

($C3$) $p_{jk} = 0.2$ for $j \ne k$: The best treatment is substantially better than the other treatments in each biomarker class.

($C4$) $p_{jk} = 0.45$ for $j \ne k$: This is a variant of ($C3$).

($C5$) $p_{12} = p_{23} = p_{31} = 0.5$, $p_{13} = p_{21} = p_{32} = 0.2$.

As in Section 3.1, we consider $I = 5$ analyses with $n_i - n_{i-1} = 200$, for $i = 1, \cdots, 5$. For each parameter configuration, we consider the probability $p_{\text{I}} = P\left(\bigcup_{j=1}^{J} A_j\right)$ that the best treatment is not included in the recommended set of treatments for some biomarker class, which is analogous to type I error, and the analog of type II error $p_{\text{II}} = P\left(\bigcup_{j=1}^{J} B_j\left(\bar{\delta}\right) \neq \emptyset\right)$, which is the probability that the recommended set contains an inferior treatment with $p_{jk} \leq p_{j,k_j} - \bar{\delta}$ for some j. Also given are $p_{\text{I},j} = P\left(A_j\right)$ and $p_{\text{II},j} = P\left(B_j\left(\bar{\delta}\right) \neq \emptyset\right)$ for each j. Table 3 gives the values of p_{I} and p_{II}, and also the expected size $\mathbb{E}\left|\mathcal{K}_j\right|$ of the recommended set of treatments for each biomarker class j, with $\bar{\delta} = 0.1$. Also given are the probabilities of rejecting H_0^* for different values of γ; in particular, the value $\gamma = 0.7$ corresponds to the type I error of the test. In addition, Table 3 also gives the mean response rate, overall and for each (j, k) category, as in Table 1 and 2. Each result is based on 10000 simulations.

Table 3 shows that p_{I} (in the row "Overall") indeed does not exceed the nominal value $\alpha = 10\%$ in all cases and that the type I error of the proposed test of H_0^* (in the column $\gamma = 0.70$) is maintained below the nominal value of $\tilde{\alpha} = 5\%$. The type II error of the proposed test and the values $p_{\text{II},1}$, $p_{\text{II},2}$, $p_{\text{II},3}$, and p_{II} (in the row "Overall") vary with the parameter configurations. The power of the GLR test of H_0^*, under the columns $\gamma = 0.63$ and $\gamma = 0.65$, are above 85% except for the parameter configuration (C4) and (C5), where they are close to 80%. The high values of $p_{\text{II},3}$ in (C4) and (C5) can be explained by the low prevalence of biomarker class $j = 3$, resulting in an expected number of 100 (out of a total of $n = 1000$) patients falling in the class. As a consequence, $\mathcal{K}_3$ contains an inferior treatment (with mean difference from the best exceeding $\bar{\delta} = 0.1$) with the high probability shown in $p_{\text{II},3}$. In fact, the values 2.25 and 1.82 for $\mathbb{E}\left|\mathcal{K}_3\right|$ in these cases suggest that the expected number of inferior treatments is 1.25 or 0.82. On the other hand, even though $\mathbb{E}\left|\mathcal{K}_j\right|$ is near 3 for every j in (C1) and close to 2 in (C2), $p_{\text{II},j} = 0$ in (C1) because there is no treatment whose mean differs from the best by more than 0.1, and $p_{\text{II},1} = p_{\text{II},2} = 0$, $p_{\text{II},3} = 0.6$ in (C2) because there is only one markedly inferior treatment.

The advantages of the proposed group sequential over the traditional design, which does not have interim analysis and uses equal randomization, can be seen by comparing Table 3 with Table 4 that gives corresponding results for the traditional design. Note that the traditional design is a special case of the group sequential design in Section 2.2 with $I = 1$. Because equal randomization dilutes the sample size for the best treatment, the power of the GLR test of H_0^* in Table 4 is lower than that in Table 5, while the overall response rate of patients in the trial is also substantially reduced as expected.

3.5. Is adaptive randomization really useful?

In their comparison of clinical trial designs with fixed sample sizes for testing whether a new treatment is better than a control treatment, Korn and Freidlin [23] have found no benefits in using (outcome-) adaptive

Table 3
Mean response rate for each treatment, probabilities $p_{\mathrm{I},j}$ and $p_{\mathrm{II},j}$ for subset selection in biomarker class j, expected subset size $\mathbb{E}\,|\mathcal{K}_j|$, and probability of rejecting H_0^* for $\gamma = 0.70$ (null), 0.65, 0.63 (alternative).

	Marker		Treatment						γ in H_0^*		
	Class		1	2	3	$p_{\mathrm{I},j}$	$p_{\mathrm{II},j}$	$\mathbb{E}\,\lvert\mathcal{K}_j\rvert$	0.63	0.65	0.70
C1	1		0.70	0.69	0.69	2.50%	0.00%	2.89			
			(171.9)	(164.2)	(163.8)						
	2		0.69	0.70	0.69	2.66%	0.00%	2.89			
			(131.2)	(137.3)	(131.5)						
	3		0.69	0.69	0.70	2.83%	0.00%	2.90			
			(33.0)	(33.0)	(33.9)						
	Overall	0.694 (1000)				7.78%	0.00%		98.7%	85.4%	3.4%
C2	1		0.70	0.69	0.20	1.50%	0.00%	1.95			
			(238.6)	(228.0)	(33.3)						
	2		0.20	0.70	0.69	1.63%	0.00%	1.95			
			(27.0)	(190.8)	(182.2)						
	3		0.69	0.20	0.70	1.70%	6.15%	2.02			
			(44.5)	(10.0)	(45.6)						
	Overall	0.660 (1000)				4.75%	6.15%		99.1%	87.5%	3.7%
C3	1		0.70	0.20	0.20	0.00%	0.00%	1.00			
			(432.3)	(33.8)	(33.7)						
	2		0.20	0.70	0.20	0.00%	0.00%	1.00			
			(27.7)	(344.8)	(27.6)						
	3		0.20	0.20	0.70	0.00%	11.52%	1.12			
			(11.6)	(11.6)	(76.9)						
	Overall	0.627 (1000)				0.00%	11.52%		99.2%	88.1%	2.9%
C4	1		0.70	0.45	0.45	0.00%	3.42%	1.04			
			(391.6)	(53.9)	(54.3)						
	2		0.45	0.70	0.45	0.00%	9.10%	1.10			
			(49.1)	(302.0)	(48.9)						
	3		0.45	0.45	0.70	0.04%	79.98%	2.25			
			(22.4)	(22.4)	(55.3)						
	Overall	0.637 (1000)				0.04%	82.42%		96.3%	78.0%	2.3%
C5	1		0.70	0.50	0.20	0.00%	6.98%	1.07			
			(393.5)	(73.0)	(33.8)						
	2		0.20	0.70	0.50	0.00%	12.86%	1.13			
			(27.5)	(305.5)	(66.8)						
	3		0.50	0.20	0.70	0.12%	72.97%	1.82			
			(28.8)	(11.2)	(60.0)						
	Overall	0.630 (1000)				0.12%	78.09%		96.8%	79.8%	2.2%

Table 4
Mean response rate for each treatment, probabilities $p_{\mathrm{I},j}$ and $p_{\mathrm{II},j}$ for subset selection in biomarker class j, expected subset size $\mathbb{E}\,|\mathcal{K}_j|$, and probability of rejecting H_0^* for $\gamma = 0.70$ (null), 0.65, 0.63 (alternative).

	Marker		Treatment						γ in H_0^*		
	Class		1	2	3	$p_{\mathrm{I},j}$	$p_{\mathrm{II},j}$	$\mathbb{E}\,\lvert\mathcal{K}_j\rvert$	0.63	0.65	0.70
C1	1		0.70	0.69	0.69	0.71%	0.00%	2.96			
			(166.6)	(166.6)	(166.7)						
	2		0.69	0.70	0.69	0.66%	0.00%	2.96			
			(133.4)	(133.4)	(133.2)						
	3		0.69	0.69	0.70	0.80%	0.00%	2.96			
			(33.3)	(33.3)	(33.4)						
	Overall	0.693 (1000)				2.15%	0.00%		98.7%	85.4%	4.1%
C2	1		0.70	0.69	0.20	0.35%	0.00%	1.99			
			(166.7)	(166.6)	(166.7)						
	2		0.20	0.70	0.69	0.31%	0.00%	1.99			
			(133.2)	(133.3)	(133.3)						
	3		0.69	0.20	0.70	0.50%	1.23%	2.00			
			(33.3)	(33.4)	(33.4)						
	Overall	0.530 (1000)				1.16%	1.23%		95.3%	75.2%	4.3%
C3	1		0.70	0.20	0.20	0.00%	0.00%	1.00			
			(166.6)	(166.5)	(166.7)						
	2		0.20	0.70	0.20	0.00%	0.00%	1.00			
			(133.5)	(133.2)	(133.6)						
	3		0.20	0.20	0.70	0.00%	7.72%	1.09			
			(33.4)	(33.3)	(33.3)						
	Overall	0.366 (1000)				0.00%	7.72%		76.4%	48.4%	2.2%
C4	1		0.70	0.45	0.45	0.00%	2.87%	1.03			
			(166.8)	(166.7)	(166.7)						
	2		0.45	0.70	0.45	0.00%	8.76%	1.10			
			(133.3)	(133.3)	(133.4)						
	3		0.45	0.45	0.70	0.00%	81.66%	2.34			
			(33.3)	(33.2)	(33.3)						
	Overall	0.533 (1000)				0.00%	83.75%		77.1%	48.6%	2.6%
C5	1		0.70	0.50	0.20	0.00%	11.15%	1.11			
			(166.6)	(166.4)	(166.7)						
	2		0.20	0.70	0.50	0.00%	20.85%	1.21			
			(133.4)	(133.3)	(133.4)						
	3		0.50	0.20	0.70	0.00%	79.83%	1.84			
			(33.5)	(33.3)	(33.3)						
	Overall	0.467 (1000)				0.00%	85.82%		76.4%	48.9%	2.4%

Table 5
Mean response rate for each treatment, probabilities $p_{\mathrm{I},j}$ and $p_{\mathrm{II},j}$ for subset selection in biomarker class j, expected subset size $\mathbb{E}\,|\mathcal{K}_j|$, and probability of rejecting H_0^* for $\gamma = 0.70$ (null), 0.65, 0.63 (alternative).

	Marker		Treatment						γ in H_0^*		
	Class		1	2	3	$p_{\mathrm{I},j}$	$p_{\mathrm{II},j}$	$\mathbb{E}\,\lvert\mathcal{K}_j\rvert$	0.63	0.65	0.70
C1	1		0.70	0.69	0.69	2.55%	0.00%	2.89			
			(168.1)	(166.1)	(165.7)						
	2		0.69	0.70	0.69	2.84%	0.00%	2.88			
			(132.8)	(134.3)	(132.8)						
	3		0.69	0.69	0.70	2.90%	0.00%	2.90			
			(33.3)	(33.3)	(33.5)						
	Overall	0.693 (1000)				8.06%	0.00%		98.9%	85.5%	4.3%
C2	1		0.70	0.69	0.20	1.33%	0.00%	1.95			
			(235.2)	(231.2)	(33.6)						
	2		0.20	0.70	0.69	1.41%	0.00%	1.96			
			(27.6)	(187.4)	(184.9)						
	3		0.69	0.20	0.70	1.49%	0.84%	1.97			
			(43.1)	(13.5)	(43.4)						
	Overall	0.658 (1000)				4.17%	0.84%		99.0%	87.1%	4.1%
C3	1		0.70	0.20	0.20	0.00%	0.00%	1.00			
			(429.5)	(35.3)	(35.4)						
	2		0.20	0.70	0.20	0.00%	0.00%	1.00			
			(30.5)	(339.2)	(30.2)						
	3		0.20	0.20	0.70	0.00%	3.29%	1.04			
			(17.4)	(17.3)	(65.2)						
	Overall	0.617 (1000)				0.00%	3.29%		98.6%	85.6%	2.6%
C4	1		0.70	0.45	0.45	0.00%	1.04%	1.01			
			(348.9)	(75.9)	(75.1)						
	2		0.45	0.70	0.45	0.00%	4.11%	1.05			
			(69.4)	(260.9)	(69.8)						
	3		0.45	0.45	0.70	0.09%	71.94%	2.14			
			(29.5)	(29.5)	(40.9)						
	Overall	0.613 (1000)				0.09%	73.37%		91.1%	69.0%	2.2%
C5	1		0.70	0.50	0.20	0.00%	2.34%	1.02			
			(358.6)	(106.8)	(34.5)						
	2		0.20	0.70	0.50	0.00%	6.20%	1.06			
			(29.1)	(273.1)	(97.8)						
	3		0.50	0.20	0.70	0.07%	67.32%	1.70			
			(36.9)	(16.0)	(47.2)						
	Overall	0.612 (1000)				0.07%	70.06%		93.4%	72.9%	2.3%

instead of traditional equal randomization, “in terms of required sample sizes, the numbers and proportions of patients having an inferior outcome.” Their results are in sharp contrast to the results of Tables 3 and 4. Note, however, that whereas Table 3 uses a group sequential design with $I = 5$ analyses and allows treatment elimination at each analysis, Table 4 uses a fixed sample size design that corresponds to the case $I = 1$. Following [23], it is natural to ask whether the advantages of the proposed design over the traditional design are mainly due to the group sequential feature that allows early termination of inferior treatments. We have therefore also tried the same group sequential design in conjunction with equal (instead of adaptive) randomization for the surviving treatments in each biomarker class. Note that the threshold a_α for treatment elimination remains the same, irrespective of equal or adaptive randomization. Moreover, the rejection threshold $d_{\tilde{\alpha}}$ for the group sequential GLR test with equal randomization can be determined in the same way as in Section 3.4. Comparison of Table 3 with Table 5, which gives the corresponding results for the group sequential design with equal randomization, shows that the marked improvements of adaptive randomization (Table 3) over equal randomization (Table 4) are substantially diminished when a group sequential design with early termination of significantly inferior treatments is used.

4. Discussion

The emerging field of biomarker-guided personalized therapies is an exciting new direction in translational medicine and poses new challenges to designing and analyzing clinical trials for their development and validation. While traditional designs often require large sample sizes, adaptive Bayesian designs such as that used by BATTLE, which “allows researchers to avoid being locked into a single, static protocol of the trial”, can yield breakthroughs, as pointed out in an April 2010 editorial in *Nature Reviews in Medicine* on such designs. In the same issue of the journal, Ledford [24] comments on these adaptive designs: “The approach has been controversial, but is catching on with both researchers and regulators as companies struggle to combat the nearly 50% failure rate of drugs in large, late-stage trials.” The BATTLE trial, however, is not associated with new drug development that is funded by a pharmaceutical company. For new drug development, we have described in Section 1 biomarker-guided accrual design for phase III trials. These designs are indeed promising in “driving down the cost of clinical trials 50-fold” in comparison with traditional clinical trials, which Ledford argues to be important in mitigating “the risk of developing a drug for these small numbers of patients.” The adaptive accrual designs actually do not have such risk as they are targeted towards the entire ITT population and switch to the Dx+ subpopulation only after the data show futility for ITT.

In the case of approved drugs, pharmaceutical companies would not sponsor clinical trials for developing and testing biomarker-guided personalized treatment selection strategies. Funding for such trials can come from private foundations and government agencies as in the case of the BATTLE trial, or from the Patient-Centered Outcomes Research Institute, established after the 2010 Patient Protection and Affordable Care Act to undertake comparative effectiveness research (CER). Fiore et al. [25] and Shih and Lavori [26] have recently proposed to use (a) the infrastructure of clinical experiments in natural clinical settings, such as POC (point of care) clinical trials, and (b) group sequential designs to conduct CER trials more easily and at a much lower cost than the traditional randomized clinical trial approach. The innovative designs introduced in Section 2.2 can be regarded as a continuation of that line of work, incorporating biomarkers into CER for personalized treatment selection. Their development and implementation have also led to new methodological advances in adaptive randomization, which is the focus of Section 2.1, and in sequential subset selection and testing non-smooth multiparameter hypotheses, which is treated in Sections 2.2, 3.2 and 3.3. In particular, we have demonstrated the statistical efficiency of the adaptive randomization rule proposed in Section 2.2 as a modification of the UCB rule in multi-arm bandit theory for clinical trials. It is much simpler than the Bayesian adaptive randomization rule used in the BATTLE trial, and is also convenient to use in conjunction with GLR statistics for group sequential testing and frequentist inference.

The group sequential design has an additional advantage that the cut-points used to define the biomarker classes do not have to be finalized until analyzing the data from the trial up to the time of the first interim analysis. The choice of these cut-points is normally based on data from previous early-phase trials with relatively small sample sizes in the literature. For example, Kim et al. [11, pp. 51-52] describe the measurement technology used in the BATTLE trial and the biomarker scoring methods used to develop the classifier. In particular, "combined expression of cytoplasmic and membrane staining" or "expression of nuclear staining" was examined for different proteins, and "all expression was assessed using semiquantitative analysis of intensity and extension" to derive a score ranging from 0 to 300, or expressed as a percentage for nuclear expression. "Cytoplasmic and membrane expression scores $>$100 were considered positive for VEGF and VEGFR-2, and scores $>$200 were considered positive for RXRβ and RXRγ." Moreover, "a nuclear score $>$30% was considered positive for RXRα, and a nuclear score $>$10% was considered positive for CyclinD1." Such semiquantitative classification is "unsupervised learning" based on heuristics and convenience. A supervised learning approach is proposed for BATTLE-2, which will "prespecify an extremely limited set of markers and will use the first half of the study population (approximately 200 patients) to conduct prospective testing of biomarkers/signatures" to guide "the second half of the study (approximately 200 patients)." Jiang et al. [27] have proposed to use the results of a phase III trial for a secondary analysis to identify the cut-points for defining biomarker classes in a future study. The initial stage of the group sequential design in Section 2.2 can be augmented to incorporate supervised learning of the biomarker classifier with cut-points chosen on the basis of clinical trial data up to the first interim analysis, which is analogous to the secondary analysis proposed in [27] and also to the first half of the BATTLE-2 design but is more flexible. Note that the initial stage (prior to the first interim analysis) uses equal randomization to the K treatments in the absence of a biomarker classifier. This is equivalent to the hypothetical version of SOC in [7], which is assumed to choose the treatments with equal probability. If one wants to test whether the BGS to be developed is significantly better than this hypothetical version of SOC, then one already has clinical trial data of the SOC and does not need to rely on historical data. Therefore, in addition to its multiple objectives listed in Section 2.2, the group sequential trial design proposed herein can also be used to build the biomarker classifiers on the basis of clinical trial data up to the first interim analysis and even to gather actual data about the SOC. Its sample size should be large enough to accomplish these goals, but it can be funded as a POC trial to improve the effectiveness of existing treatments, as discussed in the preceding paragraph.

Acknowledgments

This research was supported by the NSF grant DMS-1106535 and the NIH grant 1 P30 CA124435-01.

References

[1] Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. Stat Med. 2009;28(10):1445–1463.

[2] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986;73(3):751–754.

[3] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharm Stat. 2011;10(4):347–356.

[4] Wang SJ, O'Neill RT, Hung H. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat. 2007;6(3):227–244.

[5] Bartroff J, Lai TL, Shih MC. Sequential Experimentation in Clinical Trials. Springer; 2013.

[6] Simon R. Development and validation of biomarker classifiers for treatment selection. J Stat Plan Inference. 2008;138(2):308–320.

[7] Lai TL, Lavori PW, Shih MCI, Sikic BI. Clinical trial designs for testing biomarker-based personalized therapies. J Clin Trials. 2012;9(2):141–154.

[8] Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. J Clin Trials. 2010;7(5):567–573.

[9] Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer step toward personalized medicine. J Clin Trials. 2008;5(3):181–193.

[10] Lee JJ, Gu X, Liu S. Bayesian adaptive randomization designs for targeted agent development. J Clin Trials. 2010;7(5):584–596.

[11] Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1(1):44–53.

[12] Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. Adv Appl Math. 1985;6(1):4–22.

[13] Lai TL. Adaptive treatment allocation and the multi-armed bandit problem. Ann Stat. 1987;15(3):1091–1114.

[14] Brezzi M, Lai TL. Optimal learning and experimentation in bandit problems. J Econ Dyn Control. 2002;27(1):87–108.

[15] Lai TL, Shih MC. Power, sample size and adaptation considerations in the design of group sequential clinical trials. Biometrika. 2004;91(3):507–528.

[16] Gupta SS, Panchapakesan S. On a class of subset selection procedures. Ann Math Stat. 1972;43(3):814–822.

[17] Gupta SS, Panchapakesan S. Multiple decision procedures: theory and methodology of selecting and ranking populations. Wiley; 1979.

[18] Chan HP, Lai TL. Sequential generalized likelihood ratios and adaptive treatment allocation for optimal sequential selection. Seq Anal. 2006;25(2):179–201.

[19] Lai TL, Liao OYW. Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. Seq Anal. 2012;31(4):441–457.

[20] Chernoff H. On the distribution of the likelihood ratio. Ann Math Stat. 1954;p. 573–578.

[21] Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 1987;82(398):605–610.

[22] Feder PI. On the distribution of the log likelihood ratio test statistic when the true parameter is" near" the boundaries of the hypothesis regions. Ann Math Stat. 1968;39(6):2044–2055.

[23] Korn EL, Freidlin B. Outcome-adaptive randomization: Is it useful? J Clin Oncol. 2011;29(6):771–776.

[24] Ledford H. Clinical drug tests adapted for speed. Nature Rev Med. 2010;464(7293):1258.

[25] Fiore LD, Brophy M, Ferguson RE, D'Avolio L, Hermos JA, Lew RA, et al. A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. J Clin Trials. 2011;8(2):183–195.

[26] Shih MC, Lavori PW. Sequential methods for comparative effectiveness experiments: Point of care clinical trials. Stat Sin. To appear in 2013;.

[27] Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst. 2007;99(13):1036–1043.