Clinical Trial Designs for Testing Biomarker-Based Personalized Therapies

Tze Leung Lai[a,b], Philip Lavori[b,a], Mei-Chiung Shih[b,c] and Branimir Sikic[d]

October 17, 2011

[a]Department of Statistics, Stanford University, Stanford, California

[b]Department of Health Research and Policy, Stanford University, Stanford, California

[c]VA Cooperative Studies Program, Palo Alto, California

[d]Department of Medicine, Stanford University, Stanford, California

Running head: Testing biomarker-based personalized therapies

Author for correspondence: Mei-Chiung Shih, PhD, Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA. E-mail: meichiun@stanford.edu

Background Advances in molecular therapeutics in the past decade have opened up new possibilities for treating cancer patients with personalized therapies, using biomarkers to determine which treatments are most likely to benefit them, but there are difficulties and unresolved issues in the development and validation of biomarker-based personalized therapies. We develop a new clinical trial design to address some of these issues. The goal is to capture the strengths of the frequentist and Bayesian approaches to address this problem in the recent literature, and to circumvent their limitations.

Methods We use generalized likelihood ratio tests of the intersection null and enriched strategy null hypotheses to derive a novel clinical trial design for the problem of advancing promising biomarker-guided strategies toward eventual validation. We also investigate the usefulness of adaptive randomization and futility stopping proposed in the recent literature.

Results Simulation studies demonstrate the advantages of testing both the narrowly focused enriched strategy null hypothesis related to validating a proposed strategy and the intersection null hypothesis that can accommodate to a potentially successful strategy. Adaptive randomization and early termination of ineffective treatments offer increased probability of receiving the preferred treatment and better response rates for patients in the trial, at the expense of more complicated inference under small-to-moderate total sample sizes and some reduction in power.

Limitations The binary response used in the development phase may not be a reliable indicator of treatment benefit on long-term clinical outcomes. In the proposed design, the biomarker-guided strategy is not compared to "standard of care", such as physician's choice that may be informed by patient characteristics. Therefore a

positive result does not imply superiority of the biomarker-guided strategy to "standard of care." The proposed design and tests are valid asymptotically. Simulations are used to examine small-to-moderate sample properties.

Conclusion Innovative clinical trial designs are needed to address the difficulties and issues in the development and validation of biomarker-based personalized therapies. The paper shows the advantages of using likelihood inference and interim analysis to meet the challenges in the sample size needed and in the constantly evolving biomarker landscape and genomic and proteomic technologies.

Key words: Adaptive randomization; Biomarkers; Generalized likelihood ratio statistics; Futility stopping; Personalized therapies; Ovarian cancer.

Introduction

The current anticancer drug pipeline involves hundreds of kinase inhibitors and other targeted drugs that are in various stages of development. While the targeted treatments are devised to attack specific targets, the "one size fits all" treatment regimens commonly used may have diminished their effectiveness. Genome-guided and risk-adapted personalized therapies that are tailored for individual patients are expected to substantially improve the effectiveness of these treatments.

To achieve this potential for personalized therapies, the first step is to identify and measure the relevant biomarkers. The markers can be individual genes or proteins or gene expression signatures. How to measure them conveniently from patients is also an important consideration. One can use tissue samples from the tumor – fixed versus fresh, or circulating tumor cells, or from serum. One has to decide which biomarker measurement technology to use – quantitative rt-PCR, immunohistochemicals (IHC), phospho-flow, etc. The next step is to select drugs (standard cytotoxins, monoclonal antibodies, kinase inhibitors and other targeted drugs) based on the genetics of the disease in individual patients and biomarkers of drug sensitivity and resistance. The third step is to design clinical trials to provide data for the development and validation of personalized therapies. The designs for validation of the personalized, biomarker-guided strategy (BGS) are reviewed in the next section. The design of a Phase II clinical trial currently being planned at Stanford Cancer Center to test a biomarker-guided personalized therapy for platinum-resistant ovarian cancer has led us to consider a number of important issues not addressed in the literature and to explore some recently proposed related methods. We describe these issues and some proposals to address them. In particular, we show that using the data from such a trial, a generalized likelihood ratio (GLR) test of an "enriched strategy null hypothesis" can be carried out to test the proposed BGS, while another GLR test of an

"intersection null hypothesis" can be performed to demonstrate the efficacy of one biomarker strategy, not necessarily the BGS proposed from observational data on uncontrolled treatments in patients with stored samples. Simulation studies of these proposals and related methods in the literature are also presented.

Frequentist designs with fixed sample size

Clinical trial designs to test the effectiveness of biomarker-guided personalized therapies evolved from what are called "targeted designs" by Simon and Maitournam [1] for clinical trials with eligibility restricted to patients who are predicted to respond to the therapy being tested by using genomic technologies. The targeted designs are compared in [1] with traditional randomized designs having broader eligibility criteria, with regard to the number of patients required for randomization and for screening, in the setting of fixed sample size trials. Simon [2] subsequently considered the development and validation of biomarker classifiers for treatment selection, and in particular the design of validation studies for comparing a biomarker-based treatment strategy to "standard of care" that does not use the biomarkers to select treatment. He pointed out the inefficiency of the obvious design that randomizes patients to biomarker-based treatment selection and to treatment selection without the classifier, due to the overlap of treatments actually received by the two groups (Figure 1 of [2]). Such designs are now commonly called "biomarker-strategy designs" [3], and an example of biomarker-strategy designs is a prospective randomized controlled trial (RCT) comparing a biomarker-directed chemotherapy versus physicians' choice in patients with recurrent platinum-resistant ovarian cancer [4].

Simon [2] proposed a more efficient design (Figure 2 of [2]) which excludes patients for whom the biomarker-guided and standard of care treatment choices agree and which randomizes the remaining patients to the two treatment strategies. Such designs are now called "enrichment designs" [3, 5, 6]. Another design, called "biomarker-stratified design" in [3] and "all-comers

design" in [6], randomizes patients to the treatments regardless of biomarker status but the analysis plan focuses on the dependence of the treatment effects on the biomarker status.

Bayesian adaptive randomization designs

Besides frequentist designs for validation trials, there has been much interest in Bayesian adaptive randomization designs to test BGS in the recent literature. In particular, the BATTLE (biomarker-integrated approach of targeted therapy of lung cancer elimination) project, described in [7, 8, 9], consists of one umbrella trial and four parallel phase II studies with targeted therapies (erlotinib, sorafenib, vandetanib, and a combination of erlotinib and bexarotene), labeled Treatments 1, 2, 3, 4, respectively, for patients with advanced non-small cell lung cancer (NSCLC). It is assumed that there are four biomarker profiles so that a treatment can be identified to be most efficacious in patients whose biomarker profile matches the treatment's mechanism of action (Figure 1 of [7]). Those patients whose four biomarker profiles are all negative are assigned to biomarker group 0; we use 0 instead of 5 used in [7, 8] to be consistent with our subsequent notation. These patients are adaptively randomized to one of the four treatments according to their biomarker groups. Barker et al. [10] also describe a Bayesian clinical trial design involving adaptive randomization for I-SPY2 (investigation of serial studies to predict therapeutic response with imaging and molecular analysis) in breast cancer.

The adaptive randomization scheme like the one in [7] is based on a Bayesian probit model that is used to characterize the response rate p_{kj} of the jth patient subgroup to the kth treatment. Let Y_{kj} be the indicator variable (taking the value 1 or 0) for response of a patient in the jth subgroup receiving treatment k. The probit model in [7] relates Y_{kj} to a latent variable $Z_{kj} \sim N(\mu_{kj}, 1)$ such that $\{Y_{kj} = 1\} = \{Z_{kj} > 0\}$, and the Bayesian probit model assumes μ_{kj} to have a normal prior distribution $N(\phi_k, \sigma^2)$ with $\phi_k \sim N(0, \tau^2)$. The hyperparameters σ^2

and τ^2 of the hierarchical Bayes model can be chosen to be very large to approximate a vague prior. A more refined Bayesian logistic regression model with a multivariate normal prior for the regression parameters is assumed in [8]. The posterior mean $\gamma_{kj}^{(t)}$ of the response to treatment k in patient subgroup j given the observed responses, 1 or 0, of patients up to time t can be computed by Gibbs sampling. Letting $\hat{\gamma}_{kj}^{(t)} = \max(\gamma_{kj}^{(t)}, 0.1)$, the randomization proportion for a patient in subgroup j to receive treatment k at time $t+1$ is $\hat{\gamma}_{kj}^{(t)} / \sum_{h \in H_t} \hat{\gamma}_{hj}^{(t)}$, where H_t denotes the subset of all non-suspended treatments for that patient subgroup at $t+1$. The kth treatment is suspended at time $t+1$ for the jth patient subgroup if the posterior probability of response to treatment k for this patient subgroup has less than 10% chance of exceeding 0.5.

Let θ_0 be the response rate of the standard of care. In this Bayesian design, the biomarker-guided therapy is considered to be superior to standard of care if there is at least $100\delta\%$ chance that the posterior probability of response exceeds θ_0, where the cutoffs θ_0 and δ are chosen to match a pre-specified (e.g., 5%) frequentist type I error rate by simulations under some parameter configurations belonging to the null hypothesis. Note that the posterior probability refers to the conditional probability of an event given the data, and that in this Bayesian framework, "chance" refers to the probability given the data.

Planning a randomized trial to test a biomarker-guided treatment strategy for platinum-resistant ovarian cancer

Patient selection

A randomized trial that is currently being planned at Stanford Cancer Center involves patients with platinum-resistant ovarian cancer, defined as having progression during first-line therapy or within six months of its completion. These patients have several treatment options, most notably liposomal doxorubicin, topotecan and docetaxel, all of which lead to similar remis-

sion rates around 10-20% [11–15]. In our projections, we assume a remission rate of 15% for each drug. These three drugs, all approved for recurrent ovarian cancer, have well understood molecular targets, mechanisms of action, and mechanisms of resistance. Patients are required to have adequate renal, hepatic, and bone marrow function, and no prior treatment with liposomal doxorubicin, topotecan, or docetaxel. Patients with either measurable disease or a CA-125 marker-only relapse are eligible; CA-125 marker relapse is defined as an elevated CA-125 on two successive readings at least one month apart, two-fold above the prior lowest reading, and in the abnormal range. Eligible patients have tumors that are serous, endometrioid, mixed and undifferentiated (which usually cluster with serous in gene expression profiles), comprising approximately 80% of ovarian epithelial cancers. Clear cell, mucinous and low malignant potential tumors are excluded from this trial because of their different clinical behavior. Tissue blocks from the original diagnostic or debulking operation are available for biomarker analysis at the time of entry to this trial.

Biomarkers and patient stratification

Desirable characteristics of therapeutic biomarkers include heterogeneity of expression among cancers of a given type, robust assays for detection of the biomarker, and preclinical and clinical evidence for association of tumor response with marker expression. In addition, a practical consideration but also a critical point in biomarker development is that technologies such as gene expression signatures which require fresh tissue would markedly diminish patient eligibility for acceptance of the trial. A requirement for fresh biopsies would be expensive, invasive, and unacceptable to a high proportion of patients. The panel of biomarkers for this study uses quantitative reverse transcriptase PCR of gene expression in tumor sections from the original diagnostic or debulking surgery, for the following genes: ABCB1, TOPO2A, TOPO1, and TUBB3 [14, 16–26].

Essential to the design of this trial is the expected distribution of values for the biomarkers in ovarian cancer, and the selection of cutoffs for classification of the biomarkers as predicting either drug sensitivity or drug resistance. To approach this problem, we have reviewed the published data on expression of these genes in ovarian cancer [14, 16–19, 21, 23–25], the microarray data from our studies [27], and a recent comprehensive report from Tothill et al. of gene expression profiling of ovarian carcinomas, including 204 cases of serous papillary cancer [28]. The pairwise correlations of the expressions of these biomarkers are not significant in our and Tothill's datasets, and therefore we regard them as uncorrelated covariates for patient stratification, based on certain cutoffs for the expression levels of the 4 genes. The basic idea is to use expression levels beyond the cutoffs (above or below, depending on the case) to predict resistance (R) or sensitivity (S) to the three drugs liposomal doxorubicin (LD), topotecan (Top) and Docetaxel (Dxl). Table 1 summarizes the preferred drug for each of the 16 strata formed by these cutoffs of the four gene expressions.

INSERT TABLE 1 ABOUT HERE

In Table 1, the "pan-resistant" patients are those for whom the biomarkers predict all three treatments to be equally ineffective. The percentage of such patients is expected to be 21.6%. In addition, 9.6% of the patients are expected to be "pan-sensitive" in the sense that the patient is predicted to respond equally well to all three drugs. The other subgroups in Table 1 can be labeled as LD, Top, Dxl, LD or Top, LD or Dxl, Top or Dxl. For these 6 strata, the first 3 are "singlets", with respective percentages 9.6%, 14.4%, and 14.4% of the patients (totalling 38.4%). The last three are "doublets", with respective percentages 6.4%, 9.6%, and 14.4%, (totalling 30.4%). We can label the singlet strata as $j = 1, 2, 3$ and the treatments as $k = 1, 2, 3$ so that the biomarkers recommend treatment j for the jth stratum, and label the doublet strata as $j = 4, 5, 6$ so that the biomarkers recommend against treatment $j - 3$ for stratum

9

j. In addition, we label the subgroup consisting of pan-resistant or pan-sensitive patients as subgroup 0.

Design considerations

The BGS called for by Table 1 presents some new challenges for design of an appropriate clinical trial. First, there is no distinction between new treatments and standard treatments, because all three candidate treatments are used as standard of care. Second, there are strata where the BGS recommends only one treatment, (strata 1-3), and others where it recommends either of two treatments. Third, the "standard of care" against which the BGS is to be (ultimately) compared is not well-defined, even when restricted to the three treatments, since it is a physician's choice that may be informed by patient characteristics. Fourth, the data underlying Table 1 are preliminary and do not provide a uniform level of confidence in the recommendations made in each stratum.

The last two issues interact to discourage us from considering a "biomarker strategy" design, along the lines of [4]. Efficiency considerations, pointed out by Simon [2], encourage us to consider a design of the "biomarker stratified, enriched" type: randomizing patients to all three treatments in each stratum, excluding stratum 0 from the main study. The preliminary nature of the current classifier (issue four) has an impact on the choice of null hypothesis, as we discuss below, after introducing some general notation.

The intersection null and enriched strategy null hypotheses

Suppose a development study has come up with J patient subgroups and a biomarker-guided choice among K treatments for each subgroup. The ovarian cancer example of Table 1 is a special case, where $J = 6$ and $K = 3$. Let p_{kj} be the response rate, an unknown parameter, of the jth subgroup to the kth treatment. In addition, there is a subgroup 0 for which the

biomarkers do not have recommended treatment. In an enriched, biomarker stratified design, the comparison to a true physician's choice cannot be made, since that condition is not represented in the design. But there are other comparisons that can be used to represent different measures of the success or failure of the BGS. We use the ovarian cancer example to illustrate these comparisons.

In the ovarian cancer example, within each stratum where there are two BGS recommended treatments, the BGS response rate could be taken as the average of the two treatment-specific response rates. Corresponding to the "standard of care" response rate in a stratum, one might take the average response rate over all three treatments (lacking a true representation of a physician's choice condition). The overall effect of the BGS compared to that hypothetical version of standard of care might be defined as the average within-stratum difference of the BGS and "standard of care" response rates, weighted by the stratum proportions, a recipe that we make precise in the next paragraph, using the ovarian cancer example to fix ideas for the general case. One might instead generalize the "enrichment design" to define the effect of adhering to the BGS recommendation or doing precisely the opposite. That is, comparing the response rate (within each stratum) under the BGS recommended treatment to the response rate under the complementary treatment (not recommended by the BGS), averaging over treatments when necessary, and then averaging over strata. This is also made precise below.

The response rate p_{kj} here means rate of remission within six months for patients in the jth biomarker group receiving treatment k. Let R_j (N_j) be the set of indices of treatments recommended (not recommended) by the BGS, for patients in stratum j. Let $P_j = (\sum_{k \in R_j} p_{kj})/\|R_j\|$ be the average response for patients in stratum j to the treatments recommended by the BGS for such patients, where $\|A\|$ denotes the size of a set A. Note that for the ovarian example, $P_j = p_{jj}$ for $1 \leq j \leq 3$, and $P_j = (\sum_{k \neq j-3} p_{kj})/2$ for $4 \leq j \leq 6$. The BGS based on Table

1 assuming random choice among multiple recommended treatments has response rate P_j for the jth patient subgroup, and therefore its overall response rate is $\sum_{j=1}^{6} \pi_j P_j$, where π_j is the prevalence, or probability of occurrence, of subgroup j among the $i = 1, \ldots, J$ subgroups. To represent the "standard of care" response rate, one can define $\bar{p}_j = (\sum_{k \in R_j \cup N_j} p_{kj}) / \|R_j \cup N_j\|$) which is the response rate for the jth biomarker group if the treatments were assigned by clinicians with equal probability without the biomarker guidance. (If in each stratum j one of the treatments h were to be considered the standard, one could replace $\bar{p}_j$ by p_{hj}, and one could also use another convex combination of the p_{kj} to represent standard of care.)

A comparison of BGS to the hypothetical version of standard of care could be based on testing if an estimate of $\sum_{j=1}^{J} \pi_j (P_j - \bar{p}_j)$ is significantly positive. However, as noted in [2], it is more efficient to consider testing $\sum_{j=1}^{J} \pi_j (P_j - \bar{q}_j)$ instead, where

$$\bar{q}_j = (\sum_{k \in N_j} p_{kj}) / \|N_j\|. \tag{1}$$

Note that $P_j - \bar{q}_j$ is the difference in response rates between the treatments recommended and those not recommended by the BGS. This leads to the *enriched strategy null hypothesis*

$$H_0^* : \sum_{j=1}^{J} \pi_j (P_j - \bar{q}_j) \leq 0, \tag{2}$$

which is what an enrichment design attempts to test. However, it may be premature to focus on testing H_0^* to validate the BGS in a Phase II trial. In particular, in the ovarian cancer study, the proposed ovarian BGS is based on observational data, motivating a Phase II clinical trial whose data may be used to improve the BGS. For example, some other treatment than j may actually be better for some subgroup j ($j = 1, 2, 3$), and some other subgroup j ($j = 4, 5, 6$) may actually be pan-sensitive (i.e., properly belonging to subgroup 0).

To evaluate whether there is some value in the current BGS and to identify its correctible

shortcomings, we propose to also test the *intersection null hypothesis*

$$H_0 : p_{1j} = \cdots = p_{Kj} \text{ for } 1 \leq j \leq J. \tag{3}$$

Note that the intersection null hypothesis implies the enriched strategy null hypothesis. Rejection of H_0 implies that in at least one of the strata defined by the BGS, there is at least one superior and one inferior treatment, not necessarily the ones set up by the BGS. Rejection of H_0 also implies that there is some biomarker strategy, not necessarily the one set up for validation, that has better response rate than random assignment of the K treatments. At the sample sizes appropriate for a Phase II trial, rejection of the intersection null is likely only when there are some strata with large differences in response rates among the treatments. If the superior results coincide with the BGS recommendation, so much the better, but even if there are surprises, the result would guide further development. Since H_0 implies H_0^*, an alternative to H_0^* is also an alternative hypothesis of H_0.

Generalized likelihood ratio tests of H_0 and H_0^*

The GLR statistic for testing the intersection null hypothesis (3) after observing the outcomes of n_{kj} patients from subgroup j who have received treatment k, $1 \leq j \leq J$, $1 \leq k \leq K$, is

$$\Lambda = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{kj} \{\hat{p}_{kj} \log(\hat{p}_{kj}/\hat{p}_j) + (1 - \hat{p}_{kj}) \log[(1 - \hat{p}_{kj})/(1 - \hat{p}_j)]\}, \tag{4}$$

where $\hat{p}_{kj}$ is the proportion of responders among the n_{kj} patients (belonging to group j and receiving treatment k) and

$$\hat{p}_j = (\sum_{k=1}^{K} n_{kj}\hat{p}_{kj})/(\sum_{k=1}^{K} n_{kj}). \tag{5}$$

Note that $\hat{p}_{kj}$ is the sample mean of the independent Bernoulli random variables $Y_1, \ldots, Y_{n_{kj}}$ with mean p_{kj}. Using a traditional RCT to test the intersection null hypothesis H_0, with fixed sample size $n = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{kj}$, one can apply the χ^2-test based on GLR, which has an

13

approximate $\chi^2_{(K-1)J}$ distribution under H_0. The χ^2-approximation, however, is inappropriate when $\hat{p}_{kj}$ is near 0 and 1 (too few successes or failures in n_{kj} Bernoulli trials). An obvious way to ameliorate this is to truncate $\hat{p}_{kj}$ below and above by *a priori* bounds $0 < b < B < 1$, i.e., by redefining $\hat{p}_{kj}$ as b (or B) if it is below b (or above B).

Interim monitoring for biomarker trials has been advocated by Freidlin et al. [3] who argue that "the monitoring rule should be able to stop the study in the subgroup(s) for which the therapeutic question has been answered while continuing the subgroups that have open questions," and by Liu et al. [29] who propose a two-stage enrichment design which begins with enrollment from a certain subgroup of patients and extends to all subgroups after first-stage data show an estimated treatment improvement over a certain threshold, and which stops the trial for futility at the end of the first stage otherwise. We can use a GLR test that can stop for futility during interim monitoring, as in [3], [29] and Simon's [30] two-stage design in Phase II cancer trials. The basic methodology of group sequential GLR tests has already been developed in Section 3.4 of Lai and Shih [31], which also allows early stopping for efficacy, but here we only follow [31] to consider a specified class H_A of one-sided alternatives and to stop for futility when a group sequential GLR test of H_A rejects it at interim analysis. Note that the futility stopping in this case means stopping the entire trial early as in [29], rather than eliminating an apparently futile subgroup mid-course during the trial as in [3]. For the choice of H_A, let

$$H_A : \sum_{j=1}^{J} \pi_j \left(\max_{1 \leq k \leq K} p_{kj} - \widetilde{q}_j \right) \geq \varepsilon, \tag{6}$$

where letting $\widetilde{P}_j = \max_{1 \leq k \leq K} p_{kj}$, we define the following analog of (1):

$$\widetilde{q}_j = (\sum_{k:\ p_{kj} < \widetilde{P}_j} p_{kj})/(\sum_{k:\ p_{kj} < \widetilde{P}_j} 1).$$

In addition to the GLR test of the intersection null H_0, we can also carry out the GLR test of the enriched strategy null hypothesis H_0^*. We find the constrained maximum likelihood

estimates $\hat{p}_{kj}^*$ of p_{kj} $(1 \leq k \leq K, 1 \leq j \leq J)$ under the linear constraint $\sum_{j=1}^{J} \hat{\pi}_j (P_j - \bar{q}_j) \leq 0$, in which $\hat{\pi}_j$ is the relative frequency of patient subgroup j in the sample. This constrained maximization problem can be solved by standard convex programming algorithms, such as `constrOptim` in R. The GLR statistic for testing H_0^* is given by

$$\Lambda^* = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{kj} \{\hat{p}_{kj} \log(\hat{p}_{kj}/\hat{p}_{kj}^*) + (1 - \hat{p}_{kj}) \log[(1 - \hat{p}_{kj})/(1 - \hat{p}_{kj}^*)]\}. \tag{7}$$

The GLR statistic is approximately χ_1^2, or equivalently, the signed-root GLR statistic is approximately standard normal. Since H_0^* is a one-sided hypothesis, one should use a one-sided GLR test that rejects H_0^* at level α if the signed-root GLR statistic exceeds the normal quantile $z_{1-\alpha}$, or equivalently, if

$$\Lambda^* \geq \chi_{1;1-\alpha}^2 \quad \text{and} \quad \sum_{j=1}^{J} \hat{\pi}_j(\hat{P}_j - \hat{\bar{q}}_j) > 0, \tag{8}$$

where the notation $\hat{\ }$ denotes the corresponding (unconstrained) maximum likelihood estimates. Note that the constrained maximum likelihood estimates $\hat{p}_{kj}'$ of p_{kj} under the constraint $\sum_{j=1}^{J} \pi_j(\max_{1 \leq k \leq K} p_{kj} - \widetilde{q}_j) \geq \varepsilon$ can be computed similarly for the GLR statistic for testing H_A, for which $\hat{p}_{kj}^*$ is replaced by $\hat{p}_{kj}'$ in (7).

Although H_0 is a subset of H_0^*, which seems to suggest that it is easier to reject H_0 than H_0^* when H_0^* does not hold, the GLR test of H_0^* may actually have higher power at a given alternative than that of H_0 because the former is one-sided and uses a standard normal quantile, or a χ_1^2-quantile together with an inequality constraint, while the latter uses a $\chi_{(K-1)J}^2$-quantile. Another way to explain this is that H_0 considers the individual probabilities of the treatment-biomarker classes while H_0^* considers a linear combination of these probabilities. Note that futility stopping is only applied to the test of H_0. It means accepting H_0, which is a subset of H_0^*, upon stopping the trial. Deciding not to reject H_0^* does not suffice to stop the trial in our framework because continuing may still result in eventual rejection of H_0 and thereby demonstrating that there is some biomarker strategy that has better response rate than random

assignment of the K treatments.

Adaptive randomization and termination of failing treatments

We now consider adding two features to the RCT that have been claimed to be the advantages of the Bayesian approaches: outcome adaptive randomization and early termination of ineffective treatments. The χ^2-approximation to the null distribution of the GLR statistic can be extended to allow adaptive treatment allocation when the adaptive scheme satisfies certain conditions [32]. In particular, instead of the complicated Bayesian adaptive randomization scheme based on a Bayesian probit model as in [7], we can use a simple adaptive randomization scheme that involves the MLE $\hat{p}_{kj}^{(t)}$ rather than the posterior estimate $\hat{\gamma}_{kj}^{(t)}$ of p_{kj}. Specifically we use equal randomization to the K treatments until at least one patient is enrolled in each treatment in each subgroup, and then switch to adaptive randomization that assigns treatment k to a patient in subgroup j at time $t+1$ with probability

$$\hat{p}_{kj}^{(t)} / \sum_{h=1}^{K} \hat{p}_{hj}^{(t)}. \tag{9}$$

To reduce the chance that non-response in the first few subjects receiving a certain treatment in a particular biomarker subgroup prevents future subjects in this biomarker group from being randomized to this treatment again, the randomization probability in (9) can be truncated below by an *a priori* bound $\eta > 0$.

Analogous to [7, p.185], we can suspend randomization to treatment k for subgroup j by resetting $\hat{p}_{kj}^{(s)} = 0$ in (9) for $s > t$ if

$$\hat{p}_{kj}^{(t)} < p_* \quad \text{and} \quad n_{kj}^{(t)} > n_*, \tag{10}$$

where p_* is a "futility" threshold and n_* is a minimal sample size associated with an inferior estimate, analogous to the cutoffs θ_1 and δ_L in the Bayesian criterion for suspension in [7].

Note that with early suspension of treatment, even equal randomization could result in patient imbalance among the k treatments.

Simulation studies

This section consists of two simulation studies of the operating characteristics of the proposed methods for the design and analysis of RCT to test a BGS.

Simulation study 1

The first simulation study considers the same setting as that in [7] for a trial with $K = 4$ treatments and $J = 4$ patient subgroups for whom the biomarker strategy to be tested identifies the best treatment, together with a patient subgroup 0 for whom the strategy does not have a preferred treatment. The total sample size is 200, including those patients in subgroup 0 who are excluded from the trial. There are 15%, 20%, 30% and 25% of the patient population in the subgroups 1, 2, 3 and 4, and the subgroup 0 contains 10% of the patient population.

Table 2 gives the expected sample size in each treatment-patient group (k, j) in four scenarios. Each result is based on 5000 simulations. Also given in parentheses is the response rate for each group (k, j). A row showing the total expected sample size for each subgroup j is also included. This is followed by another row giving the total response rate (ResRate) and the probability of receiving the best treatment (Prob) in square brackets. As in [7], the equal randomization (ER) design with equal probability (1/4) of assignment to each of the four treatments is compared with the adaptive randomization (AR) design described in the preceding section, which uses MLEs in (9) in lieu of the Bayes esimates in [7]. The truncation bounds used in $\hat{p}_{kj}$ to modify the MLE are $b = 0.05$ and $B = 0.95$. The lower bound for randomization probabilities under AR is $\eta = 0.10$. In the first scenario (S1), the response rates are $p_{11} = p_{22} = 0.6$, $p_{33} = p_{44} = 0.75$, $p_{13} = p_{23} = p_{43} = p_{14} = p_{24} = p_{34} = 0.3$, and

$p_{13} = p_{23} = p_{43} = p_{14} = p_{24} = p_{34} = 0.1$. In the next two scenarios (S2 and S3), which have the same parameter configurations as Scenarios 1 and 2 of [7], the response rates are $p_{11} = 0.8$, $p_{22} = p_{33} = p_{44} = 0.6$, and $p_{kj} = 0.3$ for $k \neq j$. Scenario 3 suspends inferior treatments in both the ER and AR designs while Scenario 2 does not, and the suspension threshold and minimal sample size for suspension in (10) are $p_* = 0.28$ and $n_* = 5$. In the fourth scenario (S4), $p_{kj} = 0.3$ for all k and j. This parameter configuration belongs to H_0.

Table 2 shows that when the superior treatment in the subgroups has a much larger response rate (S1), AR randomizes substantially more subjects to the superior treatment, leading to a larger response rate in the study sample. For S2-S3, the increase in the probability of receiving the superior treatment by using the adaptive randomization scheme (9) are comparable to those reported in [7] which uses much more complex Bayesian adaptive randomization schemes. Allowing early suspension of inferior treatments in S3 yields higher response rates than those of S2, for both ER and AR. Note that in [7] the comparisons are to an external historical control rate, while in our setup we are testing randomized treatments against each other in a RCT.

In S4, the type I error probability is larger than the nominal 0.10 level under both ER and AR. The inflation of type I error rate is due to the fact that with the moderate total sample size considered here (and consequently the relatively small n_{kj}), the asymptotic χ^2-distribution is not a good approximation to the distribution of the GLR test statistic. This can be remedied in the fixed sample design with ER by using the Bartlett correction [33]. With Bartlett correction, the type I error (in S4) for ER is 0.11; the corresponding power for S1-S3 are 1.00, 0.86 and 0.93. The Bartlett correction, however, has not been extended to AR.

INSERT TABLE 2 ABOUT HERE

Simulation study 2

The second simulation study is related to the ovarian cancer study described earlier, with $K = 3$ treatments and $J = 6$ patient subgroups for whom the biomarker strategy to be tested identifies the best treatment(s), together with a patient subgroup 0 for whom the strategy does not have a preferred treatment. The total sample size is 300, including those patients in subgroup 0 who are excluded from the trial. There are 9.6%, 14.4%, 14.4%, 6.4%, 14.4% and 9.6% of the patient population in the subgroups 1 to 6, and the subgroup 0 contains 31.2% of the patient population, so about 200 patients are expected to be included in the trial.

As in Table 2, Table 3 gives the expected sample size and the response rate in each treatment-patient group (k, j) in four scenarios, for both the equal and adaptive randomization designs. Two fixed sample size GLR tests are conducted at the end of the study, of the intersection null hypothesis H_0 at 0.10 significance level and the enriched strategy null hypothesis H_0^*, which is one-sided, at 0.05 significance level. The probability of rejecting H_0 is denoted by Power, while that for H_0^* is denoted by Power*.

In the first scenario (S1), the biomarker strategy agrees with the actual p_{kj} values, which are equal to 0.3 for the preferred treatments and to 0.05 for the other treatments within each subgroup. In the second scenario (S2), the biomarker strategy is correct for subgroups 1-4 (with p_{kj}=0.3 for the preferred treatments and p_{kj}=0.05 for the other treatments), but there are no preferred treatments for subgroups 5 and 6, with $p_{kj} = 0.15$ for $j = 5, 6$ and all k, so the biomarker recommendation has no value in those subgroups. In the third scenario (S3), $p_{31} = p_{12} = p_{23} = p_{14} = p_{34} = p_{15} = p_{35} = p_{26} = p_{36} = 0.3$ and all other p_{kj}'s are 0.05. Thus, the biomarker strategy is correct only for subgroup 6, resulting in $\sum_{j=1}^{6} \pi_j (P_j - \bar{q}_j) = -0.05$. This parameter configuration belongs to H_0^* and falls outside H_0. Finally, in the fourth scenario (S4), $p_{kj} = 0.15$ for all k and j, which belongs to the intersection null hypothesis H_0. The truncation bounds used in $\hat{p}_{kj}$ to modify the MLE are $b = 0.05$ and $B = 0.95$. The truncation

bound for randomization probabilities under AR is $\eta = 0.20$.

In S1 and S2, for which the biomarker strategy is correctly specified for all or for most of the study subjects, the probability of rejecting H_0^* can even be higher than that of rejecting the smaller H_0, due to the larger degrees of freedom of the GLR test for testing H_0. On the other hand, in S3, for which the parameter configuration belongs to H_0^* and not to H_0, testing H_0 has a high probability of a positive conclusion (0.92 for ER and 0.89 for AR), while testing H_0^* has probability of only 0.01 of falsely rejecting it. Note that the probability of being assigned to the preferred treatment (when there is one) increases by about 0.10 from ER to AR in S1-S3; the total response rate increases by 0.01-0.02. In S4, the type I error (same as power) of the GLR test of H_0 is substantially smaller than the nominal level 0.10 in ER and AR, and that of H_0^* is also substantially smaller than the nominal level 0.05. This is due to the relatively small sample sizes n_{kj} for all groups, which tend to make the χ^2-approximation to the null distribution of the GLR statistic conservative when the $\hat{p}_{kj}$ are truncated at $b = 0.05$ and $B = 0.95$.

INSERT TABLE 3 ABOUT HERE

In Table 4, we repeat the scenarios of Table 3, but the design now includes early stopping for futility, with $\epsilon = 0.15$. The power and power* drop slightly, and under S4 (the null scenario) there is a modest reduction in the expected sample size, to 280 for ER and 274 for AR. In Table 5 we do not stop for futility, but suspend ineffective treatments within a biomarker group ($n_* = 5$, $p_* = 0.1$). This causes some reduction in power, but in each non-null scenario (S1-S3), the probability (Prob) of receiving the preferred treatment increases by about 0.15 in ER and 0.10 in AR, compared to the "no-suspension" design in Table 3. The corresponding increases in response rates, however, are considerably smaller.

INSERT TABLES 4 AND 5 ABOUT HERE

20

Discussion

Traditional clinical trial designs for development and validation of biomarker-guided therapies, as described for example in [2], often require large sample sizes; moreover, they cannot adapt to evolving knowledge about biomarkers. Adaptive randomization designs, such as those in BATTLE [7, 8, 9] and I-SPY2 [10], "which allows researchers to avoid being locked into a single, static protocol of the trial," can "yield breakthroughs, but must be handled with care" to ensure that they do not "inflate the risk of reaching a false positive conclusion," as pointed out in a recent editorial in *Nature* (April 2010). In the same issue of the journal, Ledford [34] reported BATTLE and I-SPY2, saying, "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 (cancer) drugs in large, late-stage trials." It is hoped that innovative designs can "drive down the cost of clinical trials 50-fold" in comparison with traditional designs for the development of personalized medicine, otherwise "drug companies (won't) be interested in taking the risk of developing a drug for these small number of patients" [34]. Lee et al. [8] argue that the "Bayesian framework is particularly suitable for adaptive designs because the inference does not depend on a particular, preset scheme" and that "traditional clinical trial designs are more rigid and can only answer a small number of well-formulated questions." On the other hand, a drawback of Bayesian designs is that their operating characteristics can be highly model-dependent. Mandrekar and Sargent [6] argue that "the gold standard for predictive marker validation continues (appropriately) to be a prospective RCT," and Simon [2] points out that such "external validation" is needed to "determine whether use of a completely specified diagnostic classifier for therapeutic decision making in a defined clinical context results in patient benefit."

While we agree with this criterion for validation, we also want to investigate potential ad-

vantages of outcome-adaptive randomization in the Bayesian approach. We therefore proceed by (a) using frequentist testing based on likelihood theory involving the GLR test statistics, (b) considering both the enriched strategy null hypothesis (2) related to validation and the intersection null hypothesis (3) that may be more appropriate in the development phase, (c) allowing futility stopping to stop early and accept the intersection null hypothesis, and (d) incorporating adaptive randomization and termination of failing treatments similar to the Bayesian approach but using much simpler procedures as in (9) and (10).

Our simulation studies show that AR offers increased probability of receiving the preferred treatment and better response rates than ER for patients in the trial. On the other hand, AR makes the estimates $\hat{p}_{kj}$ biased and the imbalance in patient allocation can complicate the analysis of the trial results. For example, with the moderate total sample size and the relatively small sample sizes for the various treatment-patient groups in the simulation studies, Bartlett corrections or saddlepoint approximations [33] may be needed to adjust the χ^2-approximations to the null distributions of the GLR statistics used to test H_0 or H_0^*. Although such adjustments are available for ER, their counterparts for AR are not yet developed because of the analytic difficulties caused by data-dependent randomization; the inadequacy of χ^2-approximations under AR is also noted in [35] for comparing two proportions at small samples sizes. This suggests using ER instead of AR in the ovarian cancer study, especially in view of only relatively small gains in response rates of AR that also accompany a decrease in power in Tables 3-5. That AR tends to decrease power of commonly used statistical tests has also been observed for the case $K = 2$ and $J = 1$ by others [36, 37], and that AR offers little improvement in response rate over ER for trial subjects in this setting has also been noted in [36]. See also [38] for an editorial on outcome-adaptive randomizations. While the asymptotic properties are a useful guide to choosing methods, we recognize the need to do simulations to explore small-sample behavior,

at least until the methods have been exercised enough to provide some sense of their utility in practice.

One of the clinical dilemmas that arises in studies of therapeutic biomarkers is the problem of what to do for patients in whom the biomarkers predict resistance to all of the drugs tested. We suppose that in preparation for a prospective randomized trial, such as we propose, investigators have some confidence in the stratification, based on a mix of biological knowledge and some data indicating the way the biomarkers predict response in observational studies of the outcomes of treatment given by physician's choice. The question of whether to include the "pan-resistant" patients in the main trial, contributing to the hypothesis testing, depends on the tradeoff of increased degrees of freedom for the tests versus capitalizing on an unexpected result (via a contribution to rejecting the intersection null, for example). When theory and prior data suggest that there are no differences in treatment outcomes in a subgroup, there is little justification in spending degrees of freedom on the small chance of a surprise. So we would exclude them from the main trial. Since their biomarker measurements are already taken, they can be enrolled in an auxiliary trial in which we randomize them among the K treatments. We conceive this trial as being at an earlier stage in development of the BGS, which is not expected to work in that patient population. Investigators can then measure other potential biomarkers from the tumor specimens, to explore their utility in refining the predictive model. For the ovarian cancer example, Table 1 shows that the percentage of such pan-resistant patients in the subgroup 0 is expected to be 21.6%. In the more general setting, patients belonging to the subgroup 0 that does not have a preferred treatment can be recruited to an auxiliary study, the goal of which is to reduce the size of the subgroup 0 so that it is decomposed as

$$0 = \{0\} \cup \{1\} \cup \cdots \cup \{J\}$$

in which $\{j\}$ represents the subset of subgroup 0 for which the auxiliary study assigns j as

the preferred treatment, allowing the subset $\{i\}$ to be empty if the development study cannot identify biomarkers that favor i.

Another clinical dilemma in the validation of a BGS is that classifiers cannot be expected to be perfect "straight from the box;" moreover, by the time a large validation study is completed, the classifier may be out of date, given the pace of biomarker development. This is why we recommend to use, in addition to the narrowly focused enriched strategy null hypothesis (2), the intersection null hypothesis (3) that can accommodate to a partially successful classifier. Such flexibility requires the development of novel statistical methods to analyze these designs. This paper represents a first step toward that development. The model we propose adapts to partial failure of the BGS by allowing rejection of the intersection null, even after failure to reject the strategy null. If that null is not rejected, then (up to type II error) the BGS fails more completely. In that case, the study data (and samples of tissue) can be used to investigate the failure in more detail, in order to come up with a better BGS, perhaps very different from the one proposed. But that exploration would be part of the initial development of the new BGS, which would then be a candidate for a new study of the type we propose. We do not attempt to build a model that can accommodate arbitrary revision of the BGS, since that would almost certainly lead to loss of control over operating characteristics.

Acknowledgments

References

[1] **Simon R., Maitournam A.** Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 2004; **10**:6759–6763.

[2] **Simon R.** Development and validation of biomarker classifiers for treatment selection. *J Statist Plan Infer* 2008; **138**:308–320.

[3] **Freidlin B, McShane LM, Korn EL.** Randomized clinical trials with biomarkers: Design issues. *Journal of National Cancer Institutes* 2010; **102**:152–160.

[4] **Cree IA, Kurbacher CM, Lamont A, et al.** A prospective randomized controlled trial of tunour chemosensitivity assay directed chemotherapy versus physician's choice in patients iwth recurrent platinum-resistant ovarian cancer. *Anti-Cancer Drugs* 2007; **18**:1093–1101.

[5] **Simon R.** Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized Medicine* 2010; **7**:33–47.

[6] **Mandrekar SJ, Sargent DJ.** Predictive biomarker validation in practice: lessons from real trials. *Clinical Trials* 2010; **7**:567–573.

[7] **Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ.** Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine. *Clinical Trials* 2008; **5**:181–193.

[8] **Lee JJ, Gu X, Liu S.** Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* 2010; **7**:584–596.

[9] **Kim ES, Herbst RS, Wistuba II, et al.** The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery* 2011; **1**:44–51.

[10] **Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ.** I-SPY2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics* 2009; **86**:97–100.

[11] **Berkenblit A, Seiden MV, Matulonis UA, et al.** A phase II trial of weekly docetaxel in patients with platinum-resistant epithelial ovarian, primary peritoneal serous cancer, or fallopian tube cancer. *Gynecol Oncol* 2004; **95**:624–631.

[12] **Dinh P, Harnett P, Piccart-Gebhart MJ, Awada A.** New therapies for ovarian cancer: cytotoxics and molecularly targeted agents. *Crit Rev Oncol Hematol* 2008; **67**:103–112.

[13] **Gordon AN, Fleagle JT, Guthrie D, Parkin DE, Gore ME, Lacave AJ.** Recurrent epithelial ovarian carcinoma: a randomized phase III study of pegylated liposomal doxorubicin versus topotecan. *J Clin Oncol* 2001; **19**:3312–3322.

[14] **Penson RT, Oliva E, Skates SJ, et al.** Expression of multidrug resistance-1 protein inversely correlates with paclitaxel response and survival in ovarian cancer patients: a study in serial samples. *Gynecol Oncol* 2004; **93**:98–106.

[15] **Rose PG, Blessing JA, Ball HG, et al.** A phase II study of docetaxel in paclitaxel-resistant ovarian and peritoneal carcinoma: a Gynecologic Oncology Group study. *Gynecol Oncol* 2003; **88**:130–135.

[16] **Faggad A, Darb-Esfahani S, Wirtz R, et al.** Topoisomerase IIalpha mRNA and protein expression in ovarian carcinoma: correlation with clinicopathological factors and prognosis. *Mod Pathol* 2009; **22**:579–588.

[17] **Ferrandina G, Petrillo M, Carbone A, et al.** Prognostic role of topoisomerase-IIalpha in advanced ovarian cancer patients. *Br J Cancer* 2008; **98**:1910–1915.

[18] **Baekelandt MM, Holm R, Nesland JM, Trope CG, Kristensen GB.** P-glycoprotein expression is a marker for chemotherapy resistance and prognosis in advanced ovarian cancer. *Anticancer Res* 2000; **20**:1061–1067.

[19] **Khalifa MA, Abdoh AA, Mannel RS, Walker JL, Angros LH, Min KW.** P-glycoprotein as a prognostic indicator in pre- and postchemotherapy ovarian adenocarcinoma. *Int J Gynecol Pathol* 1997; **16**:69–75.

[20] **Braun MS, Richman SD, Quirke P, et al.** Predictive biomarkers of chemotherapy efficacy in colorectal cancer: results from the UK MRC FOCUS trial. *J Clin Oncol* 2008; **26**:2690–2698.

[21] **Codegoni AM, Castagna S, Mangioni C, Scovassi AI, Broggini M, D'Incalci M.** DNA-topoisomerase I activity and content in epithelial ovarian cancer. *Ann Oncol* 1998; **9**:313–319.

[22] **Hari M, Yang H, Zeng C, Canizales M, Cabral F.** Expression of class III beta-tubulin reduces microtubule assembly and confers resistance to paclitaxel. *Cell Motility and the Cytoskeleton* 2003; **56**:45–56.

[23] **Holden JA, Rahn MP, Jolles CJ, Vorobyev SV, Bronstein IB.** Immunohistochemical detection of DNA topoisomerase I in formalin fixed, paraffin wax embedded normal tissues and in ovarian carcinomas. *Mol Pathol* 1997; **50**:247–253.

[24] **Kavallaris M, Kuo DY, Burkhart CA, et al.** Taxol-resistant epithelial ovarian tumors are associated with altered expression of specific beta-tubulin isotypes. *J Clin Investigation* 1997; **100**:1282–1293.

[25] **Koshiyama M, Fujii H, Kinezaki M, Yoshida M.** Correlation between Topo II alpha expression and chemosensitivity testing for Topo II-targeting drugs in gynaecological carcinomas. *Anticancer Res* 2001; **21**:905–910.

[26] **Mozzetti S, Ferlini C, Concolino P, et al.** Class III beta-tubulin overexpression is a prominent mechanism of paclitaxel resistance in ovarian cancer patients. *Clin Cancer Res* 2005; **11**:298–305.

[27] **Schaner ME, Ross DT, Ciaravino G, et al.** Gene expression patterns in ovarian carcinomas. *Mol Biol Cell* 2003; **14**:4376–4386.

[28] **Tothill RW, Tinker AV, George J, et al.** Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 2008; **14**:5198–5208.

[29] **Liu A, Liu C, Li Q, Yu KF, Yuan VW.** A threshold sample-enrichment approach in a clinical trial with heterogenous subpopulations. *Clinical Trials* 2010; **7**: 537–545.

[30] **Simon R.** Optimal two-stage designs for Phase II clinical trials. *Controlled Clinical Trials* 1989; **10**: 1–10.

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

[32] **Hu F, Rosenberger WF.** *The Theory of Response-Adaptive Radnomization in Clinical Trials.* 2006. Wiley, New York.

[33] **Reid N.** Saddlepoint Methods and Statistical Inference. *Statistical Science* 1988; **23**: 213–227.

[34] **Ledford H.** Clinical drug tests adapted for speed. *Nature* 2010; **464**: 1258.

[35] **Gu X, Lee J.** A simulation study for comparing testing statistics in response-adaptive randomziation. *BMC Medical research Methodology* 2010; **10**:48.

[36] **Korn EL, Freidlin B.** Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology* 2011; **29**: 771–776.

[37] **Morgan CC, Coad SD.** A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Statistics in Medicine* 2007; **26**: 1937–1954.

[38] **Berry DA.** Adaptive clinical trials: The promise and the caution. *Journal of Clinical Oncology* 2010; **29**:606–609.

Table 1. The percentage of patients expected to fit into each of 16 possible strata. Column 5 represents the expected percentage of patients who fit into each of the 16 combinations of sensitivity (S) and resistance (R). Column 6 presents the expected drug sensitivities for each grouping of biomarker expressions: LD (liposomal doxorubicin), Top (topotecan), and Dxl (docetaxel). ABCB1 is expected to be predictive for either LD or Dxl, TOPO2A only for Dxl, TOPO1 for topotecan, and TUBB3 for Dxl. For some drugs, the biomarker ABCB1 may result in clinical resistance even if the other biomarker (TOPO2A for LD or TUBB3 for Dxl) predicts clinical drug sensitivity.

ABCB1	TOPO2A	TOPO1	TUBB3	%	Predicted Drug
S	S	R	R	9.6	LD
S	S	S	R	6.4	LD or Top
S	S	S	S	9.6	LD or Top or Dxl
S	R	S	R	6.4	Top
S	R	S	S	9,6	Top or Dxl
S	R	R	S	14.4	Dxl
S	R	R	R	9.6	pan-resistant
S	S	R	S	14.4	LD or Dxl
R	R	R	R	2.4	pan-resistant
R	R	R	S	3.6	pan-resistant
R	R	S	S	2.4	Top
R	S	S	S	2.4	Top
R	S	R	R	2.4	pan-resistant
R	S	S	R	1.6	Top
R	S	R	S	3.6	pan-resistant
R	R	S	R	1.6	Top

Table 2. Expected sample size and response rate (in parentheses) for scenarios S1-S4 of Simulation Study 1. The randomization probabilities under AR are truncated below by $\eta = 0.10$.

	Trt	Subgroup 1 ER	Subgroup 1 AR	Subgroup 2 ER	Subgroup 2 AR	Subgroup 3 ER	Subgroup 3 AR	Subgroup 4 ER	Subgroup 4 AR
S1	1	7.5 (0.60)	9.6 (0.57)	10.0 (0.30)	9.0 (0.27)	15.0 (0.10)	11.1 (0.09)	12.4 (0.10)	9.4 (0.09)
	2	7.5 (0.30)	6.8 (0.26)	10.0 (0.60)	13.0 (0.58)	15.0 (0.10)	11.1 (0.09)	12.6 (0.10)	9.3 (0.09)
	3	7.5 (0.30)	6.8 (0.26)	10.0 (0.30)	9.0 (0.27)	15.0 (0.75)	26.8 (0.75)	12.6 (0.10)	9.3 (0.09)
	4	7.6 (0.30)	6.9 (0.26)	10.0 (0.30)	9.1 (0.27)	15.1 (0.10)	11.2 (0.09)	12.5 (0.75)	22.0 (0.74)
	Total	30.0	30.0	40.0	40.0	60.1	60.1	49.9	49.9
	ResRate [Prob]	0.31 [0.25]	0.39 [0.40]						
	Power	1.00	1.00						
S2	1	7.5 (0.80)	10.8 (0.79)	10.0 (0.30)	9.0 (0.27)	15.0 (0.30)	13.4 (0.28)	12.6 (0.30)	11.2 (0.27)
	2	7.5 (0.30)	6.4 (0.27)	10.0 (0.60)	13.0 (0.58)	15.0 (0.30)	13.4 (0.28)	12.4 (0.30)	11.2 (0.28)
	3	7.5 (0.30)	6.4 (0.26)	10.0 (0.30)	9.0 (0.27)	15.0 (0.60)	19.7 (0.59)	12.4 (0.30)	11.2 (0.28)
	4	7.6 (0.30)	6.5 (0.27)	10.0 (0.30)	9.1 (0.27)	15.1 (0.30)	13.5 (0.28)	12.5 (0.60)	16.3 (0.58)
	Total	30.0	30.1	40.0	40.2	60.1	60.1	49.9	49.9
	ResRate [Prob]	0.38 [0.25]	0.41 [0.33]						
	Power	0.89	0.93						
S3	1	8.5 (0.80)	11.7 (0.79)	9.1 (0.27)	8.2 (0.25)	12.4 (0.26)	11.1 (0.25)	10.8 (0.27)	9.7 (0.25)
	2	7.1 (0.29)	6.1 (0.25)	12.5 (0.59)	15.1 (0.58)	12.4 (0.26)	11.1 (0.25)	10.8 (0.27)	9.7 (0.25)
	3	7.1 (0.29)	6.1 (0.26)	9.1 (0.27)	8.3 (0.26)	22.3 (0.59)	25.9 (0.58)	10.8 (0.27)	9.7 (0.25)
	4	7.1 (0.29)	6.1 (0.26)	9.2 (0.28)	8.2 (0.25)	12.6 (0.26)	11.1 (0.25)	17.4 (0.59)	20.3 (0.58)
	Total	29.9	30.0	39.9	39.8	59.7	59.3	49.8	49.5
	ResRate [Prob]	0.41 [0.34]	0.44 [0.41]						
	Power	0.94	0.95						
S4	1	7.5 (0.30)	7.5 (0.26)	10.0 (0.30)	9.9 (0.27)	15.0 (0.30)	15.0 (0.28)	12.6 (0.30)	12.5 (0.28)
	2	7.5 (0.30)	7.4 (0.26)	10.0 (0.30)	10.0 (0.27)	15.0 (0.30)	15.0 (0.28)	12.4 (0.30)	12.5 (0.28)
	3	7.5 (0.30)	7.5 (0.26)	10.0 (0.30)	9.9 (0.27)	15.0 (0.30)	15.0 (0.28)	12.4 (0.30)	12.4 (0.27)
	4	7.6 (0.30)	7.6 (0.26)	10.0 (0.30)	10.1 (0.27)	15.1 (0.30)	15.0 (0.28)	12.5 (0.30)	12.5 (0.28)
	Total	30.0	30.0	40.0	40.0	60.1	60.1	49.9	49.9
	ResRate	0.30	0.30						
	Type I error	0.14	0.15						

Table 3. Expected sample size and response rate (in parentheses) in scenarios S1-S4 of Simulation Study 2 when there is no early stopping for futility or treatment suspension. The randomization probabilities under AR are truncated below by $\eta = 0.20$.

		Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4		Subgroup 5		Subgroup 6	
	Trt	ER	AR	ER	AR	ER	AR	ER	AR	ER	AR	ER	AR
S1	1	9.6 (0.30)	12.8 (0.28)	14.4 (0.05)	11.7 (0.04)	14.4 (0.05)	11.7 (0.04)	6.5 (0.30)	7.1 (0.27)	14.4 (0.30)	16.0 (0.28)	9.6 (0.05)	7.5 (0.04)
	2	9.6 (0.05)	8.0 (0.04)	14.4 (0.30)	19.7 (0.29)	14.4 (0.05)	11.7 (0.04)	6.4 (0.30)	7.0 (0.27)	14.4 (0.05)	11.2 (0.05)	9.6 (0.30)	10.7 (0.27)
	3	9.6 (0.05)	8.0 (0.04)	14.4 (0.05)	11.7 (0.04)	14.4 (0.30)	19.8 (0.29)	6.4 (0.05)	5.1 (0.04)	14.4 (0.30)	16.0 (0.28)	9.6 (0.30)	10.6 (0.27)
	Total	28.8	28.8	43.1	43.1	43.2	43.2	19.2	19.2	43.2	43.2	28.8	28.8
	ResRate [Prob]	0.17 [0.48]	0.19 [0.58]										
	Power | Power*	0.91 | 0.99	0.89 | 0.99										
S2	1	9.6 (0.30)	12.8 (0.28)	14.4 (0.05)	11.7 (0.04)	14.4 (0.05)	11.7 (0.04)	6.5 (0.30)	7.1 (0.27)	14.4 (0.15)	14.4 (0.14)	9.6 (0.15)	9.6 (0.13)
	2	9.6 (0.05)	8.0 (0.04)	14.4 (0.30)	19.7 (0.29)	14.4 (0.05)	11.7 (0.04)	6.4 (0.30)	7.0 (0.27)	14.4 (0.15)	14.4 (0.13)	9.6 (0.15)	9.6 (0.13)
	3	9.6 (0.05)	8.0 (0.04)	14.4 (0.05)	11.7 (0.04)	14.4 (0.30)	19.8 (0.29)	6.4 (0.05)	5.1 (0.04)	14.4 (0.15)	14.4 (0.13)	9.6 (0.15)	9.7 (0.13)
	Total	28.8	28.8	43.1	43.1	43.2	43.2	19.2	19.2	43.2	43.2	28.8	28.8
	ResRate [Prob]	0.15 [0.38]	0.16 [0.49]										
	Power | Power*	0.70 | 0.81	0.68 | 0.84										
S3	1	9.6 (0.05)	8.1 (0.04)	14.4 (0.30)	19.8 (0.29)	14.4 (0.05)	11.7 (0.04)	6.5 (0.30)	7.1 (0.27)	14.4 (0.30)	16.0 (0.28)	9.6 (0.05)	7.5 (0.04)
	2	9.6 (0.05)	8.0 (0.04)	14.4 (0.05)	11.6 (0.04)	14.4 (0.30)	19.8 (0.29)	6.4 (0.05)	5.2 (0.04)	14.4 (0.30)	16.1 (0.28)	9.6 (0.30)	10.7 (0.27)
	3	9.6 (0.31)	12.8 (0.29)	14.4 (0.05)	11.7 (0.05)	14.4 (0.05)	11.7 (0.04)	6.4 (0.30)	7.0 (0.27)	14.4 (0.05)	11.1 (0.05)	9.6 (0.30)	10.6 (0.27)
	Total	28.8	28.8	43.1	43.1	43.2	43.2	19.2	19.2	43.2	43.2	28.8	28.8
	ResRate [Prob]	0.17 [0.48]	0.19 [0.58]										
	Power | Power*	0.92 | 0.01	0.89 | 0.01										
S4	1	9.6 (0.15)	9.7 (0.13)	14.4 (0.15)	14.4 (0.13)	14.4 (0.15)	14.5 (0.13)	6.5 (0.15)	6.5 (0.13)	14.4 (0.15)	14.4 (0.14)	9.6 (0.15)	9.6 (0.13)
	2	9.6 (0.15)	9.5 (0.13)	14.4 (0.15)	14.4 (0.13)	14.4 (0.15)	14.3 (0.13)	6.4 (0.15)	6.4 (0.12)	14.4 (0.15)	14.4 (0.13)	9.6 (0.15)	9.6 (0.13)
	3	9.6 (0.15)	9.7 (0.13)	14.4 (0.15)	14.3 (0.14)	14.4 (0.15)	14.4 (0.13)	6.4 (0.15)	6.4 (0.13)	14.4 (0.15)	14.4 (0.13)	9.6 (0.15)	9.7 (0.13)
	Total	28.8	28.8	43.1	43.1	43.2	43.2	19.2	19.2	43.2	43.2	28.8	28.8
	ResRate	0.15	0.15										
	Power | Power*	0.06 | 0.02	0.03 | 0.02										

Table 4. Expected sample size and response rate (in parentheses) in scenarios S1–S4 of Simulation Study 2 when there is early stopping for futility (with $\varepsilon = 0.15$) but no treatment suspension. The randomization probabilities under AR are truncated below by $\eta = 0.20$.

		Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4		Subgroup 5		Subgroup 6	
	Trt	ER	AR	ER	AR	ER	AR	ER	AR	ER	AR	ER	AR
S1	1	9.5 (0.30)	12.5 (0.28)	14.2 (0.05)	11.5 (0.04)	14.2 (0.05)	11.5 (0.04)	6.4 (0.30)	6.9 (0.26)	14.2 (0.30)	15.8 (0.28)	9.5 (0.05)	7.5 (0.04)
	2	9.5 (0.05)	7.9 (0.04)	14.2 (0.30)	19.5 (0.29)	14.2 (0.05)	11.5 (0.04)	6.3 (0.30)	6.9 (0.27)	14.2 (0.05)	11.0 (0.05)	9.5 (0.30)	10.5 (0.27)
	3	9.4 (0.05)	7.9 (0.04)	14.2 (0.05)	11.5 (0.04)	14.2 (0.30)	19.5 (0.29)	6.3 (0.05)	5.1 (0.05)	14.2 (0.30)	15.7 (0.28)	9.5 (0.29)	10.4 (0.27)
	Total	28.5	28.4	42.6	42.5	42.6	42.6	19.0	19.0	42.7	42.6	28.5	28.4
	ResRate [Prob]	0.17 [0.48]	0.19 [0.58]										
	Power	Power*	0.90	0.97	0.87	0.96										
S2	1	9.3 (0.29)	12.2 (0.28)	13.9 (0.05)	11.3 (0.04)	14.0 (0.05)	11.3 (0.04)	6.3 (0.30)	6.9 (0.26)	14.0 (0.15)	13.9 (0.13)	9.3 (0.15)	9.3 (0.13)
	2	9.3 (0.05)	7.7 (0.04)	13.9 (0.29)	19.0 (0.28)	13.9 (0.05)	11.2 (0.04)	6.2 (0.29)	6.7 (0.26)	14.0 (0.15)	13.9 (0.13)	9.3 (0.15)	9.2 (0.13)
	3	9.3 (0.05)	7.8 (0.04)	13.9 (0.05)	11.2 (0.04)	13.9 (0.29)	19.0 (0.29)	6.2 (0.05)	5.0 (0.04)	13.9 (0.15)	13.8 (0.13)	9.3 (0.15)	9.3 (0.13)
	Total	27.9	27.8	41.8	41.5	41.8	41.6	18.6	18.5	41.8	41.6	27.9	27.8
	ResRate [Prob]	0.14 [0.38]	0.16 [0.49]										
	Power	Power*	0.68	0.77	0.67	0.80										
S3	1	9.5 (0.05)	7.9 (0.04)	14.2 (0.30)	19.4 (0.29)	14.2 (0.05)	11.5 (0.04)	6.4 (0.30)	7.0 (0.27)	14.2 (0.30)	15.9 (0.28)	9.5 (0.05)	7.5 (0.04)
	2	9.5 (0.05)	7.9 (0.04)	14.2 (0.05)	11.4 (0.04)	14.2 (0.30)	19.5 (0.29)	6.3 (0.05)	5.1 (0.04)	14.2 (0.30)	15.7 (0.28)	9.5 (0.30)	10.5 (0.27)
	3	9.4 (0.30)	12.5 (0.28)	14.2 (0.05)	11.6 (0.04)	14.2 (0.05)	11.6 (0.04)	6.3 (0.30)	6.9 (0.26)	14.2 (0.05)	11.0 (0.05)	9.5 (0.30)	10.4 (0.27)
	Total	28.5	28.4	42.6	42.5	42.6	42.5	19.0	19.0	42.7	42.6	28.5	28.4
	ResRate [Prob]	0.17 [0.48]	0.19 [0.58]										
	Power	Power*	0.90	0.01	0.87	0.01										
S4	1	9.0 (0.15)	8.8 (0.13)	13.4 (0.15)	13.2 (0.13)	13.5 (0.15)	13.2 (0.13)	6.1 (0.15)	5.9 (0.13)	13.5 (0.15)	13.2 (0.13)	9.0 (0.14)	8.8 (0.13)
	2	9.0 (0.15)	8.8 (0.13)	13.4 (0.15)	13.1 (0.13)	13.4 (0.14)	13.0 (0.13)	6.0 (0.15)	5.8 (0.12)	13.5 (0.15)	13.2 (0.13)	9.0 (0.15)	8.8 (0.13)
	3	9.0 (0.15)	8.7 (0.13)	13.4 (0.15)	13.1 (0.13)	13.4 (0.14)	13.2 (0.13)	6.0 (0.15)	5.9 (0.12)	13.4 (0.15)	13.1 (0.13)	9.0 (0.15)	8.8 (0.12)
	Total	26.9	26.4	40.3	39.4	40.4	39.5	18.0	17.6	40.4	39.5	27.0	26.4
	ResRate	0.15	0.15										
	Power	Power*	0.06	0.02	0.03	0.02										

33

Table 5. Expected sample size and response rate (in parentheses) in scenarios S1-S4 of Simulation Study 2 when there is no early stopping for futility but with treatment suspension ($n_* = 5$, $p_* = 0.1$). The randomization probabilities under AR are truncated below by $\eta = 0.20$.

	Trt	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4		Subgroup 5		Subgroup 6	
		ER	AR	ER	AR	ER	AR	ER	AR	ER	AR	ER	AR
S1	1	12.7 (0.28)	14.4 (0.27)	9.2 (0.03)	7.8 (0.03)	9.2 (0.03)	7.8 (0.03)	6.7 (0.30)	7.2 (0.27)	17.0 (0.28)	17.4 (0.27)	7.4 (0.04)	6.3 (0.04)
	2	7.7 (0.04)	6.6 (0.04)	22.8 (0.28)	24.1 (0.26)	9.2 (0.03)	7.8 (0.03)	6.7 (0.30)	7.0 (0.26)	8.8 (0.04)	7.4 (0.04)	10.6 (0.29)	11.1 (0.27)
	3	7.7 (0.04)	6.6 (0.04)	7.5 (0.03)	7.7 (0.03)	22.9 (0.28)	24.0 (0.27)	5.7 (0.05)	4.8 (0.04)	17.0 (0.28)	17.4 (0.27)	10.7 (0.29)	11.2 (0.27)
	Total	28.1	27.6	41.1	39.6	41.3	39.6	19.1	19.0	42.9	42.2	28.8	28.6
	ResRate [Prob]	0.21 [0.63]	0.22 [0.68]										
	Power | Power*	0.89 | 0.99	0.79 | 0.98										
S2	1	12.8 (0.28)	16.4 (0.27)	9.3 (0.03)	7.7 (0.03)	9.2 (0.03)	7.7 (0.03)	6.7 (0.30)	7.0 (0.27)	13.6 (0.12)	12.9 (0.11)	9.4 (0.13)	9.3 (0.12)
	2	7.7 (0.04)	6.6 (0.03)	22.7 (0.28)	24.2 (0.27)	9.2 (0.03)	7.8 (0.03)	6.6 (0.29)	7.2 (0.27)	13.7 (0.12)	13.1 (0.11)	9.4 (0.13)	9.2 (0.12)
	3	7.7 (0.04)	6.6 (0.04)	9.2 (0.03)	7.8 (0.03)	22.8 (0.28)	24.0 (0.27)	5.8 (0.05)	4.8 (0.04)	13.5 (0.12)	12.9 (0.11)	9.4 (0.13)	9.2 (0.12)
	Total	28.1	27.7	41.2	39.7	41.3	39.5	19.1	19.0	40.9	38.9	28.2	27.7
	ResRate [Prob]	0.17 [0.55]	0.18 [0.61]										
	Power | Power*	0.69 | 0.86	0.53 | 0.83										
S3	1	7.7 (0.04)	6.6 (0.04)	22.7 (0.28)	23.8 (0.26)	9.2 (0.03)	7.8 (0.03)	6.7 (0.30)	7.1 (0.26)	17.1 (0.29)	17.4 (0.27)	7.4 (0.04)	6.3 (0.04)
	2	7.7 (0.04)	6.6 (0.03)	9.3 (0.03)	7.7 (0.03)	22.9 (0.28)	24.0 (0.26)	5.8 (0.05)	4.8 (0.04)	17.1 (0.28)	17.4 (0.26)	10.6 (0.29)	11.2 (0.26)
	3	12.8 (0.29)	14.5 (0.27)	9.2 (0.03)	7.8 (0.03)	9.2 (0.03)	7.8 (0.03)	6.7 (0.29)	7.1 (0.27)	8.8 (0.03)	7.3 (0.03)	10.7 (0.29)	11.2 (0.27)
	Total	28.2	27.7	42.2	39.3	41.3	39.6	19.1	19.2	43.0	41.2	28.7	28.7
	ResRate [Prob]	0.21 [0.63]	0.22 [0.68]										
	Power | Power*	0.89 | 0.01	0.79 | 0.01										
S4	1	9.3 (0.13)	9.3 (0.12)	13.8 (0.12)	13.1 (0.11)	13.7 (0.12)	12.9 (0.11)	6.4 (0.15)	6.3 (0.13)	13.7 (0.12)	13.0 (0.11)	9.4 (0.13)	9.2 (0.12)
	2	9.4 (0.13)	9.1 (0.12)	13.5 (0.12)	12.9 (0.11)	13.7 (0.12)	12.9 (0.11)	6.4 (0.14)	6.4 (0.12)	13.6 (0.12)	13.0 (0.11)	9.3 (0.13)	9.2 (0.12)
	3	9.3 (0.13)	9.2 (0.12)	13.4 (0.12)	13.1 (0.11)	13.6 (0.12)	12.8 (0.11)	6.4 (0.14)	6.3 (0.12)	13.7 (0.12)	12.8 (0.11)	9.3 (0.13)	9.3 (0.12)
	Total	28.1	27.6	40.7	39.1	41.0	38.6	18.2	19.1	41.0	38.7	28.0	27.6
	ResRate	0.15	0.15										
	Power | Power*	0.01 | 0.03	0.01 | 0.03										