Duke University
Duke Biostatistics and Bioinformatics (B&B) Working Paper Series

Year 2012 *Paper* 17

A Bayesian Hierarchical Model for Adaptive Biomarker Strategies in Randomized Phase II Studies

William T. Barry* Charles M. Perou† P. Kelly Marcom‡
Lisa Carey** Joseph G. Ibrahim††

*Dana Farber Cancer Institute, bbarry@jimmy.harvard.edu
†University of North Carolina, Chapel Hill, chuck_perou@med.unc.edu
‡Duke University, kelly.marcom@dm.duke.edu
**University of North Carolina, Chapel Hill, lisa_carey@med.unc.edu
††The University of North Carolina at Chapel Hill, ibrahim@bios.unc.edu

http://biostats.bepress.com/dukebiostat/art17

A Bayesian Hierarchical Model for Adaptive Biomarker Strategies in Randomized Phase II Studies

William T. Barry, Charles M. Perou, P. Kelly Marcom, Lisa Carey, and Joseph G. Ibrahim

Abstract

The role of biomarkers has increased in cancer clinical trials such that novel designs are needed to efficiently answer questions of both drug effects and biomarker performance. We advocate Bayesian hierarchical models for response-adaptive randomized phase II studies integrating single or multiple biomarkers. Prior selection allows one to control a gradual and seamless transition from randomized-blocks to marker-enrichment during the trial. Adaptive randomization is an efficient design for evaluating treatment efficacy within biomarker subgroups, with less variable final sample sizes when compared to nested staged designs. Inference based on the Bayesian hierarchical model also has improved performance in identifying the sub-population where therapeutics are effective over independent analyses done within each biomarker subgroup.

The use of Bayesian hierarchical models for adaptive randomization in biomarker-driven phase II studies

Running title: Bayesian hierarchical models for adaptive randomization

William T. Barry[a,*,**], Charles M. Perou[b], P. Kelly Marcom[c], Lisa A. Carey[d], and Joseph G. Ibrahim[e]

Acknowledgments: Computation resources for simulations were through the Duke Scalable Computing Resource funded by NIH (grant number 1S10RR025590-01) and the North Carolina Biotechnology Center (grant number 2009-IDG-1002). This work was funded in part by a Partners in Excellence grant from the V Foundation for Cancer Research and by the CJL Foundation.

Key words: Response adaptive; integral biomarkers; phase II trials.

[a] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston MA

[b] Departments of Genetics and Pathology, University of North Carolina at Chapel Hill, Chapel Hill, NC

[c] Department of Medicine, Duke University Medical Center, Durham, NC

[d] Division of Hematology/Oncology, University of North Carolina at Chapel Hill, Chapel Hill, NC

[e] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC

[*] Correspondence to: William T. Barry, 450 Brookline Avenue, CLS 11007, Boston, MA 02215-5450

[**] E-mail: bbarry@jimmy.harvard.edu

ABSTRACT

The role of biomarkers has increased in cancer clinical trials such that novel designs are needed to efficiently answer questions of both drug effects and biomarker performance. We advocate Bayesian hierarchical models for response-adaptive randomized phase II studies integrating single or multiple biomarkers. Prior selection allows one to control a gradual and seamless transition from randomized-blocks to marker-enrichment during the trial. Adaptive randomization is an efficient design for evaluating treatment efficacy within biomarker subgroups, with less variable final sample sizes when compared to nested staged designs. Inference based on the Bayesian hierarchical model also has improved performance in identifying the sub-population where therapeutics are effective over independent analyses done within each biomarker subgroup.

1 Introduction

Clinical trials in cancer are designed to rigorously monitor and assess health interventions, whether as observational studies or randomized controlled trials. With expansive research in tumor biology over the the past decades, cancer has increasingly been recognized as a biologically heterogeneous disease (Golub et al., 1999; Perou et al., 2000; Vogelstein and Kinzler, 2004). Tissue and specimen collection are now commonplace in therapeutic trials, to answer correlative scientific objectives about the disease process and patient-specific responses. At the same time, the availability and decreasing costs of high-throughput technologies have enabled the evaluation of the entire genome, and of other cellular compartments such as the transcriptome, proteome, metabolome or secretome, and has vastly increased the amount of molecular data derived from biospecimen. Guidelines have been issued on the collection and use of biospecimen for biomarker development (McShane et al., 2005; Schmitt et al., 2004; Simon et al., 2009); and ultimately, the molecular characterization of tumors has been postulated as providing information at the individual patient level to optimize care, and be a critical component of personalized medicine (Hamburg and Collins, 2010).

Biomarkers are broadly defined as chemical, physical or biological assessments used as

an indicator of a patients disease state. Their application in medicine is delineated as: *prognostic markers* providing information about the overall risk of a clinical outcome (e.g. cancer recurrence); or *predictive markers* providing information about the specific effect of a therapeutic intervention (e.g. response to a targeted therapy, or treatment-related toxicity). Many laboratory-based assays have been proposed as prognostic or predictive biomarkers of cancer (Ross et al., 2003; Amado et al., 2008), and some have been shown to have both prognostic and predictive value in specific clinical settings (Albain et al., 2010). Molecular assays can also serve as *surrogate markers* when they correlate with clinical outcomes of primary interest (e.g. overall survival). Thus, they can substitute as an earlier endpoint for evaluating therapeutic benefit, or be incorporated into the design that directs on-going treatment regimen. For example, based on the results of ACOSOG Z1031 (Ellis et al., 2011), Ki-76 is proposed in the neoadjuvant ALTERNATE trial as a surrogate for response so that it directs the treatment course of patients on trial (DeCensi et al., 2011).

Traditionally, the predictive and prognostic value of molecular assays have been investigated in a retrospective manner, where biospecimen are banked during the course of the trial and evaluated on completion. This allows for a variety of study designs, e.g. nested case-control that can draw from larger randomized or observational studies when clinical outcome is rare (Pepe et al., 2001), or when laboratory resources are limited. However, only a prospective application of the biotechnologies will fully evaluate their clinical utility as an assay. This includes the accessibility of the biospecimen, evaluation of quality control of the assay, and the feasibility of making determinations from the molecular output (Simon et al., 2009) .

Response-adaptive trials designs have been advocated as a way to allocate patients such that more receive the better treatment. Wei and Durham (1978) extended the stochastic play-the-winner process of Zelen (1969) to randomization using urn models. These strategies were later used in developing the randomized Polya urn (Durham et al., 1998) and drop-the-loser rules (Ivanova, 2003) and the concept of optimal allocation was introduced by

Rosenberger et al. (2001). As a second general approach, the doubly adaptive biased coin design was introduced by (Eisele and Woodroofe, 1995) and was further developed by (Hu and Zhang, 2004) among others.

Bayesian methods for clinical trials have been well established in the statistical literature. For interim monitoring of trials, Spiegelhalter et al. (1986) advocated the use of predictive power for making decisions of early stopping. The Bayesian model was also used to determine sample size requirements during trial development (Spiegelhalter and Freedman, 1986). Many subsequent methods were developed for sample size determination as summarized in the review given by Adcock (1997). Bayesian models have been proposed for alternative study designs including non-inferiority trials of therapeutics and medical devices (Spiegelhalter et al., 2004; Chen et al., 2011). Bayesian models have been applied to seamless phase II/III designs (Inoue et al., 2002), and to adaptive designs to drop treatment arms or modify randomization (Berry, 2005, 2006).

Under the Bayesian paradigm, Kass and Steffey (1989) established as a class "conditionally independent hierarchical models" for observations drawn from distinct units (e.g. sites, clusters or geographic regions). More recently, this class of models has been proposed for phase II and III clinical trials with integral biomarkers. Thall et al. (2003) proposed the use of a hierarchical model for single arm phase II trials when subjects have multiple subtypes of the disease. Zhou et al. (2008) extended the hierarchical structure to consider multiple treatments in a probit regression model for the randomized phase II trial: Biomarker-integrated approaches of targeted therapy of lung cancer elimination (BATTLE). The book by Berry (2011) includes several illustrations for using hierarchical models to borrow information across components of a trial, and most recently, the Bayesian paradigm is used to consider an evolving series of novel therapeutics and biomarkers in I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy (Barker et al., 2009).

In the following sections, we state the motivation for considering adaptive- randomization

strategies for biomarker-driven trials in the phase II setting. Using the general notation of Kass and Steffey (1989) we define the Bayesian components of the trial. We then use simulation to summarize operating characteristics under a variety of scenarios that represent combinations of predictive biomarkers. In particular, we argue that informative prior distributions are needed for adaptive-randomization and interim monitoring to control treatment assignment early in the trial, while final evaluations of efficacy should rely on non-informative priors when the frequentist paradigm for inference is desired. Lastly, using a specific investigation of a novel targeted therapy in metastatic breast cancer, we contrast the performance of the adaptive approach against traditional staged designs for phase IIs nested within the biomarker-defined subgroups (Mandrekar and Sargent, 2010).

2 Motivation

In cancer clinical trials, the research and regulatory environment have divided the process for evaluating new therapeutics into four phases, with phase II and III studies designated for giving preliminary and definitive evidence of efficacy, respectively. For efficacy trials that incorporate prospective biomarkers, the National Cancer Institute has designated two types. *Integrated studies* involve assays clearly identified as part of the primary objective of a clinical trial, and are often intended to validate biomarkers prior to their use in future trials. As such, they should be hypothesis-testing in nature, and not hypothesis-generating and motivated by discovery. Assays are to be performed in real time and include complete plans for specimen collection, laboratory measurements and statistical analysis. *Integral studies* have many of the same elements, but are also designed such that the assay must be completed before patients can proceed on the trial. Examples include biomarkers to establish eligibility, biomarkers used for patient stratification, and biomarkers that inform treatment assignment. The most common trial designs with integral biomarkers are listed below, with representative schema in Figure 1 (Freidlin et al., 2010).

- Randomized-block designs are where the biomarker is used to define a stratification factor for randomization, but equivalent schemes are used within strata, such that globally, treatment assignment does not vary by biomarker status.

- Marker-enrichment designs are used to select a sub-population for investigation, whether it be a predictive marker for patient sensitivity to treatment, or prognostic markers to identify high-risk patients in which a new therapeutic may have the most clinical benefit.

- Marker-directed designs are where treatment assignment is determined by the integral biomarker; for example, assigning marker positive patients to the hypothesized optimal treatment (predictive marker), or to the more aggressive treatment (prognostic marker).

[Insert Figure 1]

In deciding among the different integral biomarker designs, one must weigh the relative importance of validating the prognostic or predictive value of the biomarker, versus using the information it provides to optimize efficacy of the treatments. Randomized block designs provide the only direct evidence of marker performance, but are less efficient in terms of evaluating efficacy within target biomarker subgroups when compared to marker-directed designs. Conversely, it can be argued that in the phase II setting, where the goal is to provide evidence of efficacy for future phase III studies, marker-directed designs are restrictive in terms of the possible outcomes from conducting the study. A positive level of efficacy would lead to a randomized controlled trial within the marker subgroup, while insufficient levels of efficacy would not support moving to any phase III study. Based on these distinctions, the sensitivity and specificity of the assays should be known in advance of selecting between a randomized-block or marker-directed design. With this information, the efficiency

of enrichment can be weighed against the fact that some patients who truly benefit from treatments would be excluded from receiving the regimen.

Because these are difficult considerations to establish when developing phase II trials for new drugs or indications, we propose an adaptive strategy which allows for efficacy to be evaluated across all targeted subpopulations in an efficient manner. In essence, these methods allow for a single trial to gradually and seamlessly transition from a randomized block design to a marker-directed design. As a result, more patients are randomized to optimal therapy when, and only when, biomarkers are predictive. The actual size of the trial will also vary less than a randomized-block design that uses multi-stage tests to reach similar levels of efficiency. Lastly, by using Bayesian models, trial flexibility that is induced by the data-driven adaptations will be taken into account in the statistical inferences.

3 A phase II response-adaptive design

In a randomized phase II trial with integral biomarkers, suppose we have K patient subgroups that are mutually exclusive and exhaustive for all possible assay results. A total of J treatment regimens are to be considered in the randomized trial, whether they be designated as experimental or control arms. The primary objective is to evaluate the efficacy of each drug within the biomarker subgroups, i.e. a non-comparative multi-arm phase II (Rubinstein et al., 2005; Mandrekar and Sargent, 2010).

Here, the primary clinical endpoint is considered to be a binary outcome, $y \in \{0, 1\}$. The target response rate of an effective treatment in a given subgroup will be defined as $\pi_{1,jk}$, while an unacceptable response rate is defined as $\pi_{0,jk}$. Without loss of generality, we will assume throughout that there are common target rates of interest

$$\pi_{1,jk} = \pi_1 \quad \text{and} \quad \pi_{0,jk} = \pi_0 \qquad \forall jk$$

although we note that for prognostic markers it may be more applicable to have different targets for the high- and low-risk patient subgroups.

Under the formulation of Kass and Steffey (1989), a random vector of n observations, y_n, is conditionally independent given parameters, θ. Further, conditional on hyperparameters, ϕ, the $\{\theta_i\}$ are i.i.d., such that the elements of y_n are exchangeable with a common density $p(\cdot)$.

$$y_n|\theta \sim p(y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta_i)$$
$$\theta|\phi \sim p(\theta|\phi) = \prod_{i=1}^{n} p(\theta_i|\phi)$$

3.1 Hierarchical model for binary data

Let j denote treatment arm, $j = 1 \ldots J$; and k denote biomarker group, $k = 1 \ldots K$. Nested within treatment j and biomarker k, patients are indexed by i, $i = 1 \ldots n_{jk}$, $n_j = \sum_k n_{jk}$, and $n = \sum_j n_j$. We will use n to refer to the number of patients at any point during enrollment up to a final sample size, N. The observed responses are denoted as:

$$y_{ijk} = \begin{cases} 1 & \text{if patient } i \text{ within marker } k \text{ had a response to treatment } j \\ 0 & \text{otherwise} \end{cases}$$

Let π_{ijk} be the response probability for y_{ijk} and a binary model with the link function $\theta_{ijk} = f(\pi_{ijk})$. The proposed hierarchical structure for multiple treatment and biomarker groups is

$$\theta_{ijk} \sim N(\mu_{jk}, 1)$$
$$\mu_{jk} \sim N(\phi_j, \sigma^2)$$
$$\phi_j \sim N(\alpha, \tau^2)$$

with hyperparameters, $\phi = \{\alpha, \sigma^2, \tau^2\}$. The variance parameter σ^2 controls the extent of borrowing across marker groups within each treatment; α and τ^2 represent the second-stage prior distribution to the hierarchical model.

Bayesian binary hierarchical models are well characterized, and can be implemented in specialized software including BUGS (Lunn et al., 2000) or JAGS (Plummer, 2008). For the

special case of a probit model, $f(\cdot) = \Phi^{-1}(\cdot)$, the Gaussian priors are conjugate such that the full conditional distributions have closed forms. Correcting for an error that appears in Zhou et al. (2008) and keeping hyperparamaters unspecified, they take the form

$$\theta_{ijk}|y_{ijk}, \mu_{jk} \propto \begin{cases} N(\mu_{jk}, 1) \cdot I(-\infty, 0) & y_{ijk} = 0 \\ N(\mu_{jk}, 1) \cdot I(0, \infty) & y_{ijk} = 1 \end{cases}$$

$$\mu_{jk}|\theta_{ijk}, \phi_j \sim N\Big(\frac{\sigma^2 \cdot \sum_{i=1}^{n_{jk}} \theta_{ijk} + \phi_j}{n_{jk} \cdot \sigma^2 + 1}, \frac{1}{n_{jk} + \sigma^{-2}}\Big)$$

$$\phi_j|\mu_{jk} \sim N\Big(\frac{\tau^2 \cdot \sum_{k=1}^{K} n_{jk} \cdot \mu_{jk} + \alpha}{n_j \cdot \tau^2 + 1}, \frac{1}{n_j + \tau^{-2}}\Big)$$

We provide the Gibbs sampler for the probit model in the statistical language and environment R (see Appendix for source code). This code was used to run the simulations on scalable computing resources at the author institution.

3.2 Adaptive randomization

Because the general hypothesis is that patients with certain biomarker profiles respond differently to the targeted treatments, randomization is conditional on biomarker group.

Without a prior assumption of increased efficacy of certain treatments, equal randomization (ER) occurs at the beginning of the trial. After at least one patient is assessed for response in each treatment by biomarker group $\{n_{jk} \geq 1\}$, the trial moves to adaptive randomization (AR). Under the Bayesian paradigm, randomization ratios at each step in enrollment, r_n, are based on posterior distributions for θ. The functional relationship one chooses for θ and r_n was described by Rosenberger (1993) as the treatment effect mapping.

Here, we formulate two mappings to θ. Let $\Omega_{k,n}$ represent the subset of non-suspended treatments for marker group k at the time of randomization for patient n. For the BATTLE trial, randomization was based proportionally on the posterior mean for the response rate to each treatment

$$\mathrm{r}_{jk,n} = \frac{\hat{\pi}_{jk,n}}{\sum_{w \in \Omega_{k,n}} \hat{\pi}_{wk,n}}$$

where $\hat{\pi}_{jk,n} = E[f^{-1}(\mu_{jk})|\mathbf{y}_n]$. With non-informative priors to the model, this formulation (we term "ratio-mapping") is equivalent to the sequential maximum likelihood procedure (Rosenberger et al., 2001). Alternatively, one could base randomization on the probability a treatment is superior to all others (we term "max-mapping"),

$$\mathrm{r}_{jk,n} = Pr(\prod_{\substack{j' \neq j \\ j' \in \Omega_{k,n}}} \mu_{jk,n} > \mu_{j'k,n} \mid y_n)$$

which is derived from the full posterior distribution to θ. In contrasting the two formulations, we note that max-mapping will always approach 1 when one therapy is superior to all others, whereas the value ratio-mapping approaches will depend on J, π_0, and π_1. For this reason we favor max-mapping, and is used for the proposed trial in Section 5.

One criticism of Bayesian adaptive designs is that they are unstable for small amounts of data. A heuristic solution is to delay AR until a fixed number of patients are enrolled, and Cheung et al. (2006) suggested waiting until at least 10 patients are observed for every group. However, for phase II trials with integral biomarker, this will typically not be feasible. For instance, in the BATTLE trial AR did not begin until 97 of 255 patients were enrolled (Kim et al., 2011), due to the requirement that $n_{jk} \geq 1 \quad \forall jk$ for the Gibbs sampler defined in Zhou et al. (2008). We note that even at the completion of the trial, $n_{jk} < 10$ in many of the $J * K = 20$ subgroups. For this reason, we advocate the use of a class of informative prior distributions, termed "balanced priors": $\phi_{bal} = \{\alpha = \frac{f(\pi_1)+f(\pi_0)}{2},\ 0 < \sigma^{-2},\ 0 < \tau^{-2}\}$. By increasing τ^{-2}, one stabilizes the model so that equal randomization occurs until data is accumulated from enough patients showing a difference in response rates.

3.3 Interim monitoring of efficacy

During AR all active treatment arms are continuously monitored in order to update randomization ratios. Although biomarker subgroups will assign fewer patients to ineffective treatments as the trial proceeds, for administrative purposes it may be valuable to permanently suspend treatment arms once there is sufficient evidence of ineffectiveness. Under the

Bayesian paradigm, one can compute posterior odds or Bayes factors for hypotheses in ineffectiveness. Alternatively, the frequentist paradigm approach can be mirrored by defining a threshold for futility, and use the prior distributions and all accumulated data to compute credible sets for efficacy.

Decisions based on Bayesian interval estimation were proposed in Zhou et al. (2008) and can be generalized to binary models with $f^{-1}(\mu_{jk})$ as

$$F_{n,jk} = \begin{cases} 1 & \text{if } \Pr\left(\mu_{jk} \geq f(\pi_1)|y_n\right) \leq \delta_L \\ 0 & \text{otherwise} \end{cases}$$

where $(1-\delta_L)$ is the size of a one-sided credible set, and $F_{n,jk}$ is an indicated of suspension of assignment to treatment j in biomarker group k after n patients are enrolled on the trial. We further denote $F_{jk} = \bigcup_{n=1}^{N} F_{n,jk}$ as the cumulative event of suspension at any point in the trial. If all J treatments are suspended, then patients in marker group k are excluded from enrolling on the trial. In order to be conservative about suspension with small n, we advocate using informative "skeptical" priors (Spiegelhalter et al., 1994) which would be paradoxically centered around π_1: $\phi_{skep} = \{\alpha = f(\pi_1),\ 0 < \sigma^{-2},\ 0 < \tau^{-2}\}$.

3.4 Final determination of efficacy

A final evaluation is performed for all non-suspended treatments after reaching target accrual, N, and once complete clinical information is obtained. Again, models can be contrasted using Bayes factors, or a determination of efficacy can be defined under the hierarchical model when a $(1-\delta_U)$ sized one-sided credible set to $f^{-1}(\mu_{jk})$ excludes the unacceptable response rate,

$$S_{jk} = \begin{cases} 1 & \text{if } \Pr\left(\mu_{jk} \geq f(\pi_0) \quad | \ y_N\right) > \delta_U \\ 0 & \text{otherwise} \end{cases}$$

For the final analysis, a non-informative prior where τ^{-2} approaches zero allows for the data from the trial to drive all inferences.

Using these interim and final analysis plans, there is no early stopping for highly effective treatments, which is analogous to frequentist staged designs as developed by Simon (1989).

We advocate this for phase II trials, because any treatments demonstrating benefit within (or across) marker subgroups will have greater numbers of patients assigned, and consequently, a more precise declaration of efficacy in the final analysis. This provides the optimal information to support the development of a phase III trial, whether it be in a general or selected patient population.

The main study characteristics of interest are common to non-comparative phase II designs: true positive and true negative findings of efficacy. Using the decision criteria noted above, the probabilities of making correct determinations of efficacy in each treatment and biomarker combination are

$$\begin{aligned} \mathrm{P1}_{jk} &= \Pr\left(S_{jk} = 1 \mid \mu_{jk} = f(\pi_1)\right) \\ \mathrm{P2}_{jk} &= \Pr\left(S_{jk} = 0 \mid \mu_{jk} = f(\pi_0)\right) \end{aligned}$$

The complementary probabilities are analogous to the frequentist definitions of Type I and II error.

We can also define probabilities that are complementary to family-wise error rates, which relate to the chance of making correct determinations of efficacy across all marker subgroups where a treatment is effective (P3), or not effective (P4). Likewise, the overall probability of having both true positive and negative findings is their union (P5).

$$\begin{aligned} \mathrm{P3}_j &= \Pr\left(\bigcup_{k:\mu_{jk}=f(\pi_1)} S_{jk} = 1\right) \\ \mathrm{P4}_j &= \Pr\left(\bigcup_{k:\mu_{jk}=f(\pi_0)} S_{jk} = 0\right) \\ \mathrm{P5} &= \Pr\left(\bigcup_{jk:\mu_{jk}=f(\pi_1)} S_{jk} = 1 \quad \cdot \bigcup_{jk:\mu_{jk}=f(\pi_0)} S_{jk} = 1\right) \end{aligned}$$

Operating characteristics and sample-size determinations for the proposed design can be determined by simulating a series of relevant scenarios to the trial design.

4 Simulation

The following are two simplified scenarios where $J = K = 2$ that are representative of the general research setting of predictive biomarkers in multi-arm trials: (a) evaluating a novel targeted agent against standard-of-care with a single predictive biomarker; and (b) selecting among multiple targeted agents specific to complementary predictive biomarkers. A global null to each scenario would be no increased efficacy with either agent. To illustrate how simulation is used to tune model parameters and select sample-size, we will explore each scenario with true unacceptable and acceptable rates of response of ($\pi_0 = 0.25,\ \pi_1 = 0.5$), and ($\pi_0 = 0.05,\ \pi_1 = 0.2$).

Characteristics are drawn from B=1000 simulations, where marker status is first sampled from a multinomial distribution defined by marker prevalence, $\mathbf{p}$, which is here set to be $p = (0.5, 0.5)$. Treatment assignment is made under the randomization scheme, and the observed responses are sampled as independent Bernoulli variables with $\{\pi_{jk}\}$.

[Insert Figure 2]

Figure 2 displays the average randomization rates under the single-marker scenario for ratio- and max-mapping to the two sets of target response rates. Within each panel, trajectories are drawn for models using balanced priors: $\phi_{bal} = \{\alpha = (\Phi^{-1}(\pi_1)+\Phi^{-1}(\pi_0))/2,\ \sigma^{-2} = 1,\ \tau^{-2} = 100\}$, or using non-informative priors with $\tau^{-2} = 0.01$. With balanced priors, there is attenuation in the rate at which randomization approaches the true treatment effect to each mapping. Importantly, in subgroups where there is no increased efficacy, randomization ratios remain centered around 0.5 throughout enrollment. With ratio-mapping and balanced priors, randomization rates to the effective treatment approach the true ratios of 0.67 and 0.8 for $\pi_1 = 0.5$ and $\pi_1 = 0.2$, whereas max-mapping approaches 1 in both cases. Lastly, Table 1 shows that with a strong balanced prior, randomization has minimal variation (IQR < 0.02) when the number of patients on study is very small (n = 5), but that

an unacceptably large variation (IQR > 0.5) is seen early on with non-informative priors, which is only partially attenuated using a moderate prior with $\tau = 1$.

[Insert Table 1]

Next, we evaluated the probabilities of truly and falsely determining efficacy (P1 and $1 - \text{P2}$) when using the monitoring plans outlined above. Simulations focused on designs using balanced priors and max-mapping for randomization.

[Figure 3]

By plotting P1 and $1 - \text{P2}$ over a range of target sample sizes, one can use simulation to select the desired operating characteristics to a trial. For the target rates $\pi_1 = 0.5$ and $\pi_0 = 0.25$ we found that assessing futility with a threshold of $\delta_L = 0.025$ and a skeptical prior: $\phi_{skep} = \{\alpha = \Phi^{-1}(\pi_1),\ \sigma^2 = 1,\ \tau^{-2} = 100\}$ and making a final determination of efficacy using non-informative hyperprior $\phi_{non} = \{\alpha = \Phi^{-1}(\pi_0),\ \sigma^2 = 1,\ \tau^{-2} = 0.01\}$ and $\delta_U = 0.9$ provided a good balance between controlling for false positive and negative results. In particular, P1 $\geq 80\%$ and $1 - \text{P2} \leq 10\%$ is achieved with N = 55 in the single marker scenario and with N = 59 patients in the complementary marker scenario. By simulating under a null of no efficacy, we note the probability of early stoppage before N = 55 or N = 59 is 47% and 55%, such that the average sample size would be 48.4 and 50.3 respectively.

We next compare our method to independent Simon "optimal" two-stage tests performed within a randomized-block design, as an efficient non-adaptive approach to minimize sample-size when treatments are ineffective. Under a null, $H_0 : \pi_{jk} = 0.25$, and powered on the alternative $H_1 : \pi_{jk} = 0.5$, this requires 8 subjects in the first stage and 21 subjects total per arm (target N = 84) in order to control Type I and II errors at 10% and 20% respectively. Under the respective alternative hypotheses to the single and complementary marker

scenarios, the expected sample sizes to the two-staged design are 57.1 and 65.4, respectively, and 48.7 when there is truly no efficacy with either agent. Thus, marginal improvements in efficiency are seen with our adaptive approach. As advantages, resources would need not be budgeted for the larger target sample size, and more importantly, considerably less variation is seen under our simulations than the actual sample sizes that can occur with 4 independent two-stage tests (Figure 3).

Simulations under other true effective response rates show a slight attenuation in power when compared to the larger staged-tests: under $\pi_1 = 0.45$, P1 ranged from 0.64 to 0.67 versus power of 0.69 with the Simon design. With a larger true effect size ($\pi_1 = 0.55$), P1 ranged from 0.86 to 0.88 versus power of 0.89 with a Simon design. The small differences may be due to P2 being slightly lower than the Type I error to the Simon design, or may be reflective of tuning the parameters and size of the adaptive design to optimize characteristics against the target response rates.

For target response rates of $\pi_0 = 0.05$ and $\pi_1 = 0.2$, simulations were repeated to parameterize the model and select samples sizes. Figure 2 and Table 2 show that informative balanced priors are needed to stabilize $\{r_{jk,n}\}$ early in the trial and remain 1:1 on average in the non-target subgroup, and we focus on max-mapping to increase allocation to optimal therapy. Despite the lower event rates, similar gains in efficiency can be seen in the adaptive design when allowing for a higher false positive rate. Using thresholds of $\delta_L = 0.025$ and $\delta_U = 0.8$, we find that N = 74 and 71 control P1 $\geq 80\%$ and $1 - \text{P2} \leq 15\%$ for the two scenarios. In comparison, Simon two-stages tests would require a target N = 108 ($E[N] =$ 70.2 and 82.8, for the two scenarios) to control Type I and II errors at this level.

Lastly, despite using non-informative priors for determinations of efficacy, the posterior means for the response rate are biased slightly downward for $J = K = 2$, as is known to occur with adaptive randomization (Rosenberger and Lachin, 2002). At $n = 100$, median relative risks of 0.976 and 0.959 are seen to $\pi_1 = 0.5$ and $\pi_1 = 0.2$, respectively, after randomizing patients under max-mapping and balanced priors (Table 2). The extent of bias must be

carefully considered if one reports Bayesian point-estimates from the hierarchical model at the completion of the study.

5 Example

Increasingly, both clinicians and laboratory scientists have recognized that breast cancer is a heterogeneous disease, which poses a challenge to the development of new therapies and to the appropriate application of existing treatments to individual patients. Using DNA microarray technology, Sorlie et al. (2001) identified five major subtypes of breast tumors, including basal-like, Her2 over expressing, luminal-like (including luminal A and B), and normal breast tissue-like. It was later shown that luminal B subtype tumors have a poor prognosis relative to other ER+/Her2- breast cancers, and represent a population that may derive benefit from novel treatments in the locally advanced setting (Bild et al., 2009).

Phosphatidylinositol 3-kinases (PI3Ks) have come to attention as both a marker of prognosis and a potential target for therapy in a variety of human cancers (Vanhaesebroeck et al., 2010). Once activated, these kinases phosphorylate membrane lipids which in turn trigger a complex signaling cascade leading to cell cycle entry, growth and survival. Mutations leading to constitutive activation of the pathway have been observed, with early studies reporting a 40% rate of somatic mutations in the gene in breast cancer, especially hormone receptor-positive breast cancer (Campbell et al., 2004). Multiple inhibitors of the PI3K pathway are in development that demonstrate anti-tumor activity in pre-clinical and clinical studies (Markman et al., 2010; Baselga et al., 2011). Among the most interesting targeted strategies for PI3K inhibition is the luminal B subtype of breast cancer. Although typically hormone receptor-positive, this subtype is more chemosensitive than luminal A breast cancer (Fan et al., 2006), and recent studies implicate PI3K pathway signaling in proliferation and cell survival in this subtype (Bild et al., 2009). However, aberrations of PI3K pathway signaling are common across breast cancer subtypes, and a selection strategy for identifying those

most likely to respond to inhibition of the PI3K pathway has not yet been defined.

We propose a randomized phase II to evaluate a PI3K inhibitor in advanced hormone refractory breast cancer patients. Activity of the agent will be assessed in combination with standard capecitabine in ER+/HER2- breast cancer defined by standard histological methods. Integral biomarkers will be used to evaluate whether increased efficacy is seen in molecular subgroups of greatest potential to provide a selection strategy. This includes intrinsic subtypes by mRNA expression and PI3K DNA sequencing, with the scientific hypothesis that greater efficacy is seen with either PI3K mutations over wild-type, or with luminal B and other subtypes relative to luminal A tumors.

The primary clinical endpoint for evaluating patient response to capecitabine alone (X) and capecitabine plus PI3K inhibitor (XP) will be objective response. Based on prior knowledge of the efficacy of capecitabine, we will consider a response rate of $\theta_0 = 0.25$ as unacceptable, and $\theta_1 = 0.5$ as a target level of efficacy for treatments within all marker subgroups.

[Insert Figure 4]

5.1 Design and operating characteristics

In the Bayesian AR design, we set a threshold probability of $\delta_L = 0.01$ for the futility monitoring, and $\delta_U = 0.9$ for the threshold for concluding efficacy. The balanced, skeptical, and non-informative priors described above are used for randomization, interim monitoring and final analysis respectively.

One heuristic rule is applied over the AR scheme to further control enrollment to the trial. Since there are no interim rules for stopping for superiority, the total number of patients enrolled into a single treatment by subgroup will be capped at 35 to avoid oversampling. This threshold was selected under a reduced Bayesian models for a single treatment and single biomarker subgroup, as providing greater than 95% posterior probability of concluding efficacy when $\theta = \Phi^{-1}(\pi_1)$.

Simulations were run to select a maximum target sample size based on the probabilities of truly and falsely concluding efficacy. Specifically, six scenarios define different relationships between clinical benefit of XP and the two integral biomarkers, as enumerated in Table 2. Based on anticipated accrual, and the length of follow-up needed to observe objective response, a lag of 10 patients is included into the simulation for randomization and interim monitoring of futility.

[Insert Table 2.]

Table 3 shows that with a target sample size of N = 168, in all scenarios probabilities of falsely concluding efficacy in each ineffective treatment is less than 10%, while probabilities of concluding success in each effective treatment ranges from 82.1% to 92.8% varying largely by the marker prevalence. Across simulations, effective combinations were stopped at rates between 3.7% to 6.4% while ineffective treatments were stopped at some point during the AR phase 17.8% to 87.3% of the time.

[Insert Table 3]

In comparison, parallel Simon two-stage designs require greater maximum target sample sizes, needing to allocate $23 \cdot 8 = 192$ patients to control Type I and II errors at 0.1 and 0.15 in every group. An even greater number of patients is needed to match the exact operating characteristics to each scenario that is given in Table 3, although the discrete binomial distribution prevents a direct comparison.

Finally, there is a distinctive advantage of using all available data across biomarker subgroups when making inferences under the hierarchical model (Table 4). For each scenario, the joint probabilities of correctly identifying all subgroups where XP is effective (P3), and where XP or X are ineffective (P4). Results are superior to independent analysis with the

larger Simon two-stage designs. The largest improvements are see when multiple biomarker groups demonstrate increased efficacy. For instance, if intrinsic subtype and PI3K mutation are equally predictive (Scenario 5), the probability of identifying all three subgroups increases from 0.618 to 0.694, while under a global null (Scenario 1), the chance of a false discovery decreases from 54.4% down to 42.5%.

[Insert Table 4]

6 Discussion

We have presented a novel approach to studying the efficacy of novel agents in the context of integral biomarkers. By adopting a Bayesian response adaptive model, flexibility in the trial design allows for a seamless transition from investigating agents in a general population toward a marker-directed strategy where patients are randomized with greater probability to their optimal therapy. To meet the requirements of randomized phase II studies, the model incorporates a continuous monitoring for futility and a final analysis of efficacy that are conditioned on the integral biomarkers. Simulations demonstrate the properties of the model, and its advantages over using parallel and independent staged designs.

Adaptive trial designs give a framework whereby the mathematical models account for flexibility required in phase II screening trials, and with modern computational resources the numerical routines can be implemented as easily as exact binomial tests. Adaptive trials do require a larger informatics structure to continuously monitor enrolled patients in order to maximize gains in efficiency. However, adaptive approaches can be seamless and do not require suspension of enrollment until complete outcome information is obtained and evaluated, thus removing a large operational barrier to the study team and common hindrance to study accrual with staged phase II trials.

We have shown under simulation that adapting with a Bayesian hierarchal model lowers the total target sample sizes over traditional designs. Further, in staged designs, interim

looks that occur early in the trial to optimize the characteristics can cause wide variations in final sample sizes. Flexibility and robust performance of our Bayesian AR model is demonstrated by the consistent operating characteristics seen across a variety of relationships between treatment efficacy and biomarker subgroups. Conversely, the feasibility of using parallel multi-stage tests to efficiently evaluate efficacy across biomarker groups will be more sharply impacted by unequal prevalences. We also propose that adaptive designs will be more robust to marker misspecification than a randomized-block design, based on the flexibility and gains in power from the hierarchical model. Future simulation studies are planned to demonstrate and quantify this assertion based on the joint distributions of biomarkers seen in previous integral studies Kim et al. (2011). All these points allow for such trials to be planned and budgeted for more easily using Bayesian hierarchical models and response-adaptive randomization.

The greatest benefit of our approach is that by jointly modeling efficacy of treatments in the Bayesian hierarchical model, improved statistical inferences can be made about the predictive or prognostic value of biomarkers over designs that focus on efficacy within or across patient subgroups. This will be critical for clinical contexts where integral biomarkers can be used to identify the proper study population for definitive phase III studies of efficacy. Finally, we note that as a conservative element to the adaptive approach, if the clinical data are missing or delayed (completely at random to treatment assignment), the adaptive randomization will transition more slowly from equal randomization.

Future efforts are to apply the Bayesian hierarchical structure to statistical models for other clinical endpoints that are continuous and right-censored. However, the advantages of adaptive design are maximized when endpoint can be assessed early. With the expansion of rationally identified therapeutic targets, the simultaneous identification of rational biomarkers naturally follows. Indeed, the FDA has released a draft guidance document "In Vitro Companion Diagnostic Devices" to encourage development of biomarkers (molecular or otherwise) as diagnostics for guiding treatment decisions and patient selection. The flexibility

and efficiency of adaptive clinical trial designs provide important advances for guiding and accelerating this more complex co-development process.

References

Adcock, C. J. (1997). Sample size determination: A review. *Statistician*, 46(2):261–283.

Albain, K. S., Barlow, W. E., Shak, S., Hortobagyi, G. N., Livingston, R. B., Yeh, I. T., Ravdin, P., Bugarini, R., Baehner, F. L., Davidson, N. E., Sledge, G. W., Winer, E. P., Hudis, C., Ingle, J. N., Perez, E. A., Pritchard, K. I., Shepherd, L., Gralow, J. R., Yoshizawa, C., Allred, D. C., Osborne, C. K., and Hayes, D. F. (2010). Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol*, 11(1):55–65.

Amado, R. G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D. J., Juan, T., Sikorski, R., Suggs, S., Radinsky, R., Patterson, S. D., and Chang, D. D. (2008). Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology*, 26(10):1626–34.

Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., and Esserman, L. J. (2009). I-spy 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100.

Baselga, J., Campone, M., Piccart, M., Burris, H. A., Rugo, H. S., Sahmoud, T., Noguchi, S., Gnant, M., Pritchard, K. I., Lebrun, F., Beck, J. T., Ito, Y., Yardley, D., Deleu, I., Perez, A., Bachelot, T., Vittori, L., Xu, Z., Mukhopadhyay, P., Lebwohl, D., and Hortobagyi, G. N. (2011). Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. *N Engl J Med*.

Berry, D. A. (2005). Introduction to bayesian methods iii: use and interpretation of bayesian tools in design and analysis. *Clinical Trials*, 2(4):295–300; discussion 301–4, 364–78.

Berry, D. A. (2006). Bayesian clinical trials. *Nat Rev Drug Discov*, 5(1):27–36.

Berry, S. M. (2011). *Bayesian adaptive methods for clinical trials*. Chapman & Hall/CRC biostatistics series. CRC Press, Boca Raton.

Bild, A. H., Parker, J. S., Gustafson, A. M., Acharya, C. R., Hoadley, K. A., Anders, C., Marcom, P. K., Carey, L. A., Potti, A., Nevins, J. R., and Perou, C. M. (2009). An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. *Breast Cancer Res*, 11(4):R55.

Campbell, I. G., Russell, S. E., Choong, D. Y., Montgomery, K. G., Ciavarella, M. L., Hooi, C. S., Cristiano, B. E., Pearson, R. B., and Phillips, W. A. (2004). Mutation of the pik3ca gene in ovarian and breast cancer. *Cancer Research*, 64(21):7678–81.

Chen, M. H., Ibrahim, J. G., Lam, P., Yu, A., and Zhang, Y. (2011). Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics*, 67(3):1163–70.

Cheung, Y. K., Inoue, L. Y. T., Wathen, J. K., and Thall, P. F. (2006). Continuous bayesian adaptive randomization based on event times with covariates. *Statistics in Medicine*, 25(1):55–70.

DeCensi, A., Guerrieri-Gonzaga, A., Gandini, S., Serrano, D., Cazzaniga, M., Mora, S., Johansson, H., Lien, E. A., Pruneri, G., Viale, G., and Bonanni, B. (2011). Prognostic significance of ki-67 labeling index after short-term presurgical tamoxifen in women with er-positive breast cancer. *Ann Oncol*, 22(3):582–7.

Durham, S. D., Flournoy, N., and Li, W. (1998). A sequential design for maximizing the probability of a favourable response. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 26(3):479–495. Times Cited: 29.

Eisele, J. R. and Woodroofe, M. B. (1995). Central limit-theorems for doubly adaptive biased coin designs. *Annals of Statistics*, 23(1):234–254.

Ellis, M. J., Suman, V. J., Hoog, J., Lin, L., Snider, J., Prat, A., Parker, J. S., Luo, J. Q., DeSchryver, K., Allred, D. C., Esserman, L. J., Unzeitig, G. W., Margenthaler, J., Babiera, G. V., Marcom, P. K., Guenther, J. M., Watson, M. A., Leitch, M., Hunt, K., and Olson, J. A. (2011). Randomized phase ii neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: Clinical and biomarker outcomes and predictive value of the baseline pam50-based intrinsic subtype-acosog z1031. *Journal of Clinical Oncology*, 29(17):2342–2349.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med*, 355(6):560–9.

Freidlin, B., McShane, L. M., and Korn, E. L. (2010). Randomized clinical trials with biomarkers: Design issues. *Journal of the National Cancer Institute*, 102(3):152–160.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7.

Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *N Engl J Med*, 363(4):301–4.

Hu, F. F. and Zhang, L. X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Annals of Statistics*, 32(1):268–301. Times Cited: 53.

Inoue, L. Y., Thall, P. F., and Berry, D. A. (2002). Seamlessly expanding a randomized phase ii trial to phase iii. *Biometrics*, 58(4):823–31.

Ivanova, A. (2003). A play-the-winner-type urn design with reduced variability. *Metrika*, 58(1):1–13.

Kass, R. E. and Steffey, D. (1989). Approximate bayesian-inference in conditionally independent hierarchical-models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

Kim, E. S., Herbst, R. S., Wistuba, I., Lee, J. J., Blumenschein, G. R., J., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., J., Gupta, S., Alden, C. M., Liu, S., Tang, X., Khuri, F. R., Tran, H. T., Johnson, B. E., Heymach, J. V., Mao, L., Fossella, F., Kies, M. S., Papadimitrakopoulou, V., Davis, S. E., Lippman, S. M., and Hong, W. K. (2011). The battle trial: personalizing therapy for lung cancer. *Cancer Discov*, 1(1):44–53.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Mandrekar, S. J. and Sargent, D. J. (2010). Randomized phase ii trials: time for a new era in clinical trial design. *J Thorac Oncol*, 5(7):932–4.

Markman, B., Atzori, F., Perez-Garcia, J., Tabernero, J., and Baselga, J. (2010). Status of pi3k inhibition and biomarker development in cancer therapeutics. *Ann Oncol*, 21(4):683–91.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., and Clark, G. M. (2005). Reporting recommendations for tumor marker prognostic studies (remark). *J Natl Cancer Inst*, 97(16):1180–4.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M., and Yasui, Y. (2001). Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*, 93(14):1054–61.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–52.

Plummer, M. (2008). *JAGS Version 1.0.3 Manual*. IARC, Lyon, France.

Rosenberger, W. F. (1993). Asymptotic inference with response-adaptive treatment allocation designs. *Annals of Statistics*, 21(4):2098–2107.

Rosenberger, W. F. and Lachin, J. M. (2002). *Randomization in clinical trials : theory and practice*. Wiley series in probability and statistics. Wiley, New York.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., and Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913.

Ross, J. S., Fletcher, J. A., Bloom, K. J., Linette, G. P., Stec, J., Clark, E., Ayers, M., Symmans, W. F., Pusztai, L., and Hortobagyi, G. N. (2003). Her-2/neu testing in breast cancer. *Am J Clin Pathol*, 120 Suppl:S53–71.

Rubinstein, L. V., Korn, E. L., Freidlin, B., Hunsberger, S., Ivy, S. P., and Smith, M. A. (2005). Design issues of randomized phase ii trials and a proposal for phase ii screening trials. *Journal of Clinical Oncology*, 23(28):7199–206.

Schmitt, M., Harbeck, N., Daidone, M. G., Brynner, N., Duffy, M. J., Foekens, J. A., and Sweep, F. C. (2004). Identification, validation, and clinical implementation of tumor-associated biomarkers to improve therapy concepts, survival, and quality of life of cancer patients: tasks of the receptor and biomarker group of the european organization for research and treatment of cancer. *Int J Oncol*, 25(5):1397–406.

Simon, R. (1989). Optimal 2-stage designs for phase-ii clinical-trials. *Controlled Clinical Trials*, 10(1):1–10.

Simon, R. M., Paik, S., and Hayes, D. F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*, 101(21):1446–52.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–74.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health care evaluation*. Statistics in practice. Wiley, Chichester ; Hoboken, NJ.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical-trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical-trials - conditional or predictive power. *Controlled Clinical Trials*, 7(1):8–17.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 157:357–387.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). Hierarchical bayesian approaches to phase ii trials in diseases with multiple subtypes. *Statistics in Medicine*, 22(5):763–780.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Vanhaesebroeck, B., Guillermet-Guibert, J., Graupera, M., and Bilanges, B. (2010). The emerging mechanisms of isoform-specific pi3k signalling. *Nat Rev Mol Cell Biol*, 11(5):329–41.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat Med*, 10(8):789–99.

Wei, L. J. and Durham, S. (1978). Randomized play-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843. Times Cited: 150.

Zelen, M. (1969). Play winner rule and controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131. Times Cited: 213 Zelen, m.

Zhou, X., Liu, S. Y., Kim, E. S., Herbst, R. S., and Lee, J. L. (2008). Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine. *Clinical Trials*, 5(3):181–193.

APPENDIX: Sample R code:

```
####################
## Dependent function for simulation in R
####################

MCMCfun <- function(n_i,y,group2,theta.0,theta.1,phi){
  require(msm)
  alpha <- phi[1]; sigma2 <- phi[2]; tau2 <- phi[3]
  mu <- pr.eff <- pr.stop <- pihat <- rmax <- rep(0,J*K)
  psi <- rep(0,J)
  n.jk <- table(group2)
  n.j <- tapply(n.jk,rep(1:J,each=K),sum)
  sd.mu <- (n.jk + 1/sigma2)^(-.5)
  sd.phi <- (n.j + 1/tau2)^(-.5)

  for (b in 1:(n.burn+(skip+1)*n.iter)){
    z <-  rtnorm(n_i,mu[group2],lower = c(-Inf,0)[1+y], upper = c(0,Inf)[1+y])
    mu <- rnorm(J*K,mean = (sigma2 * tapply(z,group2,sum) + rep(psi,each=K) ) /
                           (sigma2 * n.jk + 1),sd=sd.mu)
    psi <- rnorm(J,mean = (tau2 * tapply(mu*n.jk,rep(1:J,each=K),sum) + alpha ) /
                           (tau2 * n.j + 1),sd=sd.phi)
    if(b > n.burn & trunc((b-n.burn)/(skip+1))==(b-n.burn)/(skip+1)){
      pr.eff  <- pr.eff  + (mu > qnorm(theta.0))/n.iter
      pr.stop <- pr.stop + (mu > qnorm(theta.1))/n.iter
      pihat   <- pihat   + pnorm(mu) / n.iter
      rmax    <-  rmax   + (mu == rep(tapply(mu,rep(1:K,J),max),J)) / n.iter
    }
   }
  return(list(pr.eff = pr.eff, pr.stop = pr.stop, pihat = pihat, rmax = rmax))
}

####################
## Simultations for Figures 2 and 3, and Table 1
##   (parameterized for left-most column
####################

## Parameters
Nmax = 100
J = 2; K = 2;                           ## Indexes for groups
prob.K = c(0.5,0.5)                     ## Proportion of genotype groups (length K)


p0 = 0.25; p1 = 0.2                     ## Target response rates
off = 0                                 ## Offset between target and true rate.

pi     = c(p1+off, p0,                  ## True response rates
           p0    , p1+off)              ##  ordered as trt(group)

phi.r = c(alpha  = (qnorm(p0)+qnorm(p1))/2,
          sigma2 = 1, tau2   = 0.01)    ## Hyperparameters for randomization
r.method = 2                            ## Mapping (1=ratio, 2=max)

phi.f = c(alpha  = qnorm(p1),
          sigma2 = 1, tau2   = 0.01)    ## Hyperparameters for futility

phi.e = c(alpha  = qnorm(p0),
          sigma2 = 1, tau2   = 100)     ## Hyperparameters for efficacy

delta.U = 0.9                           ## Decision rule [efficacy]
delta.L = 0.025                         ## Decision rule [stop]

n.burn = n.iter = 5000; skip = 0        ## MCMC parameters

#### Simulation
set.seed(seed)
```

```
group <- assign <- group2 <- y <- rep(NA,Nmax)
stop1 <- rep(0,J*K); fail1 <- 0
theta.0 = rep(p0,J*K); theta.1 = rep(p1,J*K)

group[1:(J*K)]  <- rep(1:K,J)
assign[1:(J*K)] <- rep(1:J,each=K)
group2[1:(J*K)] <- group[1:(J*K)] + K * (assign[1:(J*K)]-1)
y[1:(J*K)] <-      runif(J*K) < pi[group2[1:(J*K)]]

## Adaptive Randomization
i <- J*K
while((i < Nmax) & (fail1==0)){
  post.f <- MCMCfun(i,y[1:i],group2[1:i],theta.0,theta.1,phi.f)  ## 1. Run MCMC for futility
  stop1[stop1==0] <- (post.f[[2]][stop1==0] < delta.L) * i       ## 2. Check futility in active arms
  drop  <- tapply(stop1,rep(1:K,J),prod)                        ## 3. Drop groups with stopped arms
  if(prod(drop)) {fail1 <- 1} else {                            ## 4. If all arms not dropped
    if(sum(!drop)>1){                                             ## 4a. Draw new patients group
      group[i+1] <- sample((1:K)[!drop],1,prob=prob.K[!drop])
    } else group[i+1] <- (1:K)[!drop]
    post.r <- MCMCfun(i,y[1:i],group2[1:i],theta.0,theta.1,phi.r)  ## 4b. Run MCMC for randomization
    if(r.method == 1){
      rand <- post.r[[3]]
    } else if(r.method == 2) rand <- post.r[[4]]
    rand[stop1>0] <- 0
    assign[i+1] <- sample(1:J,1,prob=rand[rep(1:K,J)==group[i+1]]) ## 4c. Assign treatment
    group2[i+1] <- group[i+1] + K*(assign[i+1]-1)
    y[1+i] <- runif(1) < pi[group2[1+i]]                           ## 4d. Simulate outcome
    post.e <- MCMCfun(i+1,y[1:(i+1)],group2[1:(i+1)],theta.0,theta.1,phi.e)
    write(paste(c(i+1,                                             ## 4e. Output:
                  table(group2[1:(1+i)]),                            # Sizes
                  (post.e[[1]] > delta.U)*(stop1==0),                # Dec of Eff
                  (stop1>0),                                         # Dec of Fut
                  post.e[[3]],                                       # PostMean of Eff
                  rand),                                             # Rand weights
                collapse=" "),outfile,append=T)
    print(paste("     ",i+1,"patients analyzed"))
  }
  i <- i + 1
}


####################
## Simulation of PI3K trial design: Scenario #3: LumB ONLY
####################

## Parameters
Nmax = 200                          ## Maximum possible total sample size
J = 2; K = 4                        ## Indexes for groups
prob.K = c(0.161,0.393,0.200,0.244)    ## Proportion of biomarker subgroups

pi     = c(0.25,0.25,0.25,0.25,        ## True response rates (length J*K)
           0.50,0.50,0.25,0.25)        ##  ordered as trt(group) -

r.method = 2                        ## Treatment effect mapping
phi.r = c(alpha = (qnorm(0.25)+qnorm(0.5))/2,
          sigma2 = 1,tau2  = 0.01)     ## Hyperparameters for rand
phi.f = c(alpha = qnorm(0.5),
          sigma2 = 1,tau2  = 0.01)     ## Hyperparameters for fut
phi.e = c(alpha = qnorm(0.25),
          sigma2 = 1,tau2  = 100)      ## Hyperparameters for eff

delta.U <- 0.90                     ## Decision rule [success]
delta.L <- 0.02                     ## Decision rule [stop]

lag  <- 10                          ## Lag - estimated accrual before ORR
Imin <- 0                           ## Minimum number of patients before AR
cap  <- 35                          ## Maximum number of patients per arm
```

```
n.burn = n.iter = 5000; skip = 0       ## MCMC parameters

#### Simulation
set.seed(seed)

theta.0 <- rep(0.25,J*K); theta.1 <- rep(0.5,J*K)
group  <- sample(1:K,Nmax,replace=T,prob=prob.K)
stop1  <- stop2 <- rep(0,J*K); screen <- fail1 <- 0
assign <- y <- rep(NA,Nmax)

## Phase 1) ER phase until rule for interim monitoring triggered
  group2 <- factor(assign,levels=1:(J*K))
  i <- 0;
  while(i < (Nmax-lag-1) & (sum(table(group2)==0) | i < (Imin))){
    i <- i + 1
    assign[i] <- sample(1:J,1)
    group2[i] <- group[i] + K*(assign[i]-1)
    y[i] <- runif(1) < pi[group2[i]]
  }
  start <- (i+lag)
  print(paste(" ",start,"patients in ER phase"))

  assign[i+(1:lag)] <- sample(1:J,lag,replace=T)
  group2[i+(1:lag)] <- group[i+(1:lag)] + K*(assign[i+(1:lag)]-1)
  y[i+(1:lag)]      <- runif(lag) < pi[group2[i+(1:lag)]]

## Phase 2) AR phase, arms are dropped by futility analysis
  while((i < (Nmax-lag)) & (!fail1)){
    post.f <- MCMCfun(i,y[1:i],group2[1:i],theta.0,theta.1,phi.f)
    stop1[stop1==0] <- (post.f[[2]][stop1==0] < delta.L) * i
    stop2 <- table(group2[1:(i+lag)]) >= cap
    drop  <- tapply(stop1+stop2,rep(1:K,J),prod)
    if(prod(drop)){ fail1 <- i } else {
      j <- i + 1 + lag
      if(drop[group[j]]){
        screen <- screen + 1
        c <- 1;  while(c){
          group[j] <- sample((1:K),1,prob=prob.K)
          if(drop[group[j]]) screen <- screen + 1 else c <- 0
        }
      }
      post.r <- MCMCfun(i,y[1:i],group2[1:i],theta.0,theta.1,phi.r)
      if(r.method == 1){
        rand <- post.r[[3]]
      } else if(r.method == 2) rand <- post.r[[4]]
      rand[(stop1>0)|stop2] <- 0
      assign[j] <- sample(1:J,1,prob=rand[rep(1:K,J)==group[j]])
      group2[j] <- group[j] + K*(assign[j]-1)
      y[j] <- runif(1) < pi[group2[j]]
      post.e <- MCMCfun(j,y[1:j],group2[1:j],theta.0,theta.1,phi.e)
      print(paste("     ",j,"patients analyzed"))
      write(paste(c(j,screen, table(group2[1:j]),       ## Total, screened and subgroup sizes
                    (post.e[[1]] > delta.U)*(stop1==0), ## Decision of efficacy
                    (stop1>0),                          ## Decision of futility
                    rand),                              ## Randomiztion ratios
                    collapse="\t"),outfile,append=T)
    }
    i <- i + 1
  }
```

Table 1: Characteristics of response-adaptive randomization within the single-marker scenario with ($\pi_0 = 0.25,\ \pi_1 = 0.5$), and ($\pi_0 = 0.05,\ \pi_1 = 0.2$). Medians and interquartile ranges from 1000 simulations are given under differing priors and treatment effect mapping for (a) randomization ratios at varying n, (b) final allocation to treatment arm, and (c) posterior means the the response rate in each subgroup.

		Max-mapping			**Ratio-mapping**		
		Balanced $\tau^{-2} = 100$	**Moderate** $\tau^{-2} = 1$	**Noninform.** $\tau^{-2} = 0.01$	**Balanced** $\tau^{-2} = 100$	**Moderate** $\tau^{-2} = 1$	**Noninform.** $\tau^{-2} = 0.01$
Non-target subgroup ($\pi_{11} = \pi_{21} = 0.25$)							
Rand. ratio	$n = 5$	0.50 (0.49, 0.51)	0.51 (0.36, 0.65)	0.52 (0.24, 0.95)	0.50 (0.49, 0.51)	0.50 (0.39, 0.62)	0.51 (0.35, 0.93)
	$n = 20$	0.50 (0.29, 0.73)	0.60 (0.27, 0.83)	0.77 (0.15, 0.97)	0.50 (0.38, 0.62)	0.55 (0.39, 0.73)	0.61 (0.31, 0.94)
	$n = 100$	0.51 (0.22, 0.78)	0.55 (0.22, 0.87)	0.71 (0.21, 0.98)	0.50 (0.42, 0.58)	0.51 (0.42, 0.60)	0.54 (0.42, 0.78)
Post. mean	$\hat{\pi}_{11,100}$	0.24 (0.17, 0.29)	0.24 (0.14, 0.29)	0.19 (0.02, 0.28)	0.25 (0.19, 0.30)	0.24 (0.18, 0.30)	0.22 (0.08, 0.29)
	$\hat{\pi}_{21,100}$	0.22 (0.13, 0.30)	0.21 (0.13, 0.29)	0.18 (0.02, 0.28)	0.24 (0.17, 0.32)	0.23 (0.16, 0.30)	0.21 (0.07, 0.29)
Allocation to Trt 2		0.51 (0.30, 0.71)	0.54 (0.30, 0.79)	0.68 (0.28, 0.94)	0.50 (0.40, 0.59)	0.53 (0.42, 0.65)	0.57 (0.39, 0.86)
Target subgroup ($\pi_{21} = 0.25 \quad \pi_{22} = 0.5$)							
Rand. ratio	$n = 5$	0.50 (0.49, 0.83)	0.51 (0.48, 0.88)	0.55 (0.35, 0.98)	0.50 (0.50, 0.69)	0.51 (0.48, 0.72)	0.54 (0.40, 0.95)
	$n = 20$	0.77 (0.54, 0.90)	0.82 (0.51, 0.93)	0.91 (0.30, 0.99)	0.61 (0.51, 0.72)	0.66 (0.52, 0.80)	0.71 (0.46, 0.96)
	$n = 100$	0.94 (0.86, 0.98)	0.95 (0.87, 0.98)	0.97 (0.84, 0.99)	0.66 (0.60, 0.73)	0.67 (0.60, 0.76)	0.69 (0.60, 0.88)
Post. mean	$\hat{\pi}_{12,100}$	0.25 (0.18, 0.30)	0.25 (0.20, 0.31)	0.24 (0.17, 0.30)	0.25 (0.20, 0.31)	0.25 (0.20, 0.31)	0.25 (0.19, 0.30)
	$\hat{\pi}_{22,100}$	0.49 (0.43, 0.54)	0.49 (0.42, 0.54)	0.48 (0.40, 0.54)	0.49 (0.43, 0.55)	0.49 (0.42, 0.54)	0.49 (0.42, 0.55)
Allocation to Trt 2		0.83 (0.70, 0.89)	0.84 (0.71, 0.92)	0.88 (0.61, 0.96)	0.63 (0.55, 0.71)	0.66 (0.55, 0.77)	0.69 (0.53, 0.90)
Non-target subgroup ($\pi_{11} = \pi_{21} = 0.05$)							
Rand. ratio	$n = 5$	0.50 (0.49, 0.51)	0.50 (0.49, 0.52)	0.52 (0.43, 0.64)	0.50 (0.49, 0.51)	0.50 (0.49, 0.52)	0.52 (0.41, 0.67)
	$n = 20$	0.50 (0.43, 0.58)	0.60 (0.41, 0.75)	0.91 (0.24, 0.97)	0.50 (0.42, 0.58)	0.58 (0.41, 0.73)	0.86 (0.29, 0.95)
	$n = 100$	0.50 (0.25, 0.77)	0.79 (0.28, 0.92)	0.96 (0.22, 0.99)	0.51 (0.35, 0.67)	0.64 (0.38, 0.85)	0.92 (0.39, 0.98)
Post. mean	$\hat{\pi}_{11,100}$	0.04 (0.00, 0.07)	0.02 (0.00, 0.06)	0.01 (0.00, 0.05)	0.04 (0.01, 0.07)	0.03 (0.00, 0.07)	0.00 (0.00, 0.05)
	$\hat{\pi}_{21,100}$	0.04 (0.00, 0.06)	0.03 (0.00, 0.06)	0.01 (0.00, 0.05)	0.04 (0.02, 0.07)	0.03 (0.00, 0.06)	0.00 (0.00, 0.04)
Allocation to Trt 2		0.51 (0.35, 0.67)	0.68 (0.37, 0.81)	0.88 (0.32, 0.93)	0.51 (0.40, 0.60)	0.61 (0.43, 0.75)	0.84 (0.41, 0.90)
Target subgroup ($\pi_{21} = 0.05 \quad \pi_{22} = 0.2$)							
Rand. ratio	$n = 5$	0.50 (0.49, 0.51)	0.50 (0.49, 0.52)	0.52 (0.43, 0.66)	0.50 (0.50, 0.51)	0.50 (0.49, 0.52)	0.52 (0.41, 0.67)
	$n = 20$	0.68 (0.50, 0.84)	0.79 (0.47, 0.92)	0.96 (0.31, 0.99)	0.64 (0.50, 0.75)	0.73 (0.48, 0.85)	0.94 (0.31, 0.98)
	$n = 100$	0.94 (0.86, 0.97)	0.97 (0.89, 0.99)	0.99 (0.85, 1.00)	0.80 (0.68, 0.86)	0.87 (0.73, 0.95)	0.97 (0.76, 0.99)
Post. mean	$\hat{\pi}_{12,100}$	0.05 (0.03, 0.08)	0.05 (0.03, 0.08)	0.05 (0.01, 0.07)	0.05 (0.03, 0.08)	0.05 (0.03, 0.08)	0.05 (0.02, 0.07)
	$\hat{\pi}_{22,100}$	0.19 (0.14, 0.24)	0.19 (0.14, 0.23)	0.18 (0.11, 0.23)	0.19 (0.14, 0.24)	0.19 (0.14, 0.23)	0.18 (0.13, 0.23)
Allocation to Trt 2		0.80 (0.69, 0.86)	0.85 (0.69, 0.91)	0.91 (0.66, 0.96)	0.70 (0.60, 0.77)	0.79 (0.64, 0.86)	0.89 (0.63, 0.94)

Table 2: Hypothetical relationships between intrinsic subtype, PI3K mutation status and efficacy of the inhibitor ($\pi_{XP,k}$ below). Subgroups with clinical benefit over capecitabine alone ($\pi_{X,k} = 0.25$ in all subgroups) are highlighted in gray. The joint prevalence was reported by The Cancer Genome Atlas Network (2012), and accounts for inclusion into the luminal B* subgroup basal and Her2-enriched subtypes which are seen more rarely in ER+/Her2- disease by IHC.

	Luminal B*		**Luminal A**	
	PI3K mut.	**PI3K wt.**	**PI3K mut.**	**PI3K wt.**
Prevalence	**16.1%**	**39.3%**	**20.0%**	**24.4%**
Global Null	0.25	0.25	0.25	0.25
No Biomarker	0.5	0.5	0.5	0.5
Single Biomarker				
Luminal B only	0.5	0.5	0.25	0.25
PI3K mut. only	0.5	0.25	0.5	0.25
Joint Biomarker				
Either marker	0.5	0.5	0.5	0.25
Both markers	0.5	0.25	0.25	0.25

Table 3: Probabilities of concluding efficacy by treatment and biomarker subgroup under the six scenarios defined in Table 2. All effective treatments by subgroups per scenario are shaded in gray.

	Luminal B*		**Luminal A**	
	PI3K mut.	**PI3K wt.**	**PI3K mut.**	**PI3K wt.**
Global Null				
XP	0.058	0.073	0.072	0.066
X	0.071	0.069	0.076	0.063
No Biomarker				
XP	0.821	0.928	0.892	0.899
X	0.057	0.085	0.053	0.06
Luminal B only				
XP	0.856	0.928	0.094	0.094
X	0.064	0.066	0.061	0.059
PI3K mt only				
XP	0.870	0.085	0.909	0.069
X	0.074	0.059	0.07	0.055
Either marker				
XP	0.847	0.923	0.884	0.085
X	0.061	0.074	0.068	0.072
Both markers				
XP	0.874	0.075	0.067	0.074
X	0.076	0.066	0.057	0.072

Table 4: Family-wise operating characteristics of the AR design versus parallel Simon two-stage designs.

	P3	**P4x**	**P4xp**	**P5**
Global Null				
AR	-NA-	0.748	0.760	0.575
Simon	-NA-	0.676	0.676	0.456
No Biomarker				
AR	0.625	0.771	-NA-	0.497
Simon	0.527	0.676	-NA-	0.356
Luminal B only				
AR	0.798	0.786	0.820	0.536
Simon	0.726	0.676	0.822	0.403
PI3K mt only				
AR	0.789	0.766	0.855	0.521
Simon	0.726	0.676	0.822	0.403
Either marker				
AR	0.694	0.750	0.915	0.485
Simon	0.618	0.676	0.907	0.379
Both markers				
AR	0.874	0.763	0.802	0.526
Simon	0.852	0.676	0.745	0.429

Figure 1: Schema for integral biomarker trials designs that incorporate randomized treatment arms, including randomized-block (left panel), marker-enrichment (top-right), and marker-directed designs (bottom-right).

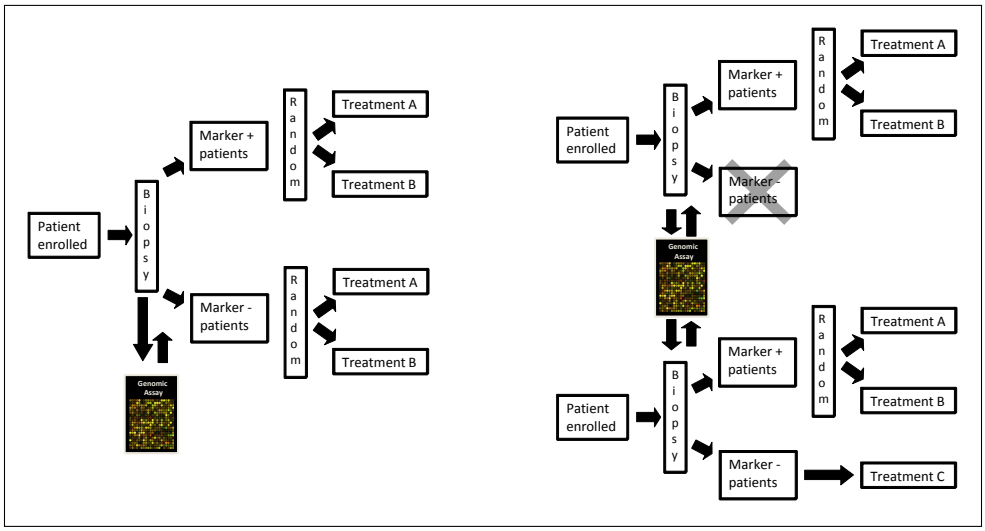

Figure 2: The average randomization ratio from N = 5 to N = 100 under the single marker scenario for the target subgroup (solid line) versus non-target subgroup (dotted-line). In each panel trajectories are drawn for non-informative priors ($\tau^{-2} = 0.01$, red) and for balanced priors ($\tau^{-2} = 100$, green). Results are displayed for ratio-mapping (left panels) and max-mapping (right panels); and for true efficacy levels of $\pi_0 = 0.25$ and $\pi_1 = 0.5$ (top panels) and for $\pi_0 = 0.05$ and $\pi_1 = 0.2$ (bottom panels).

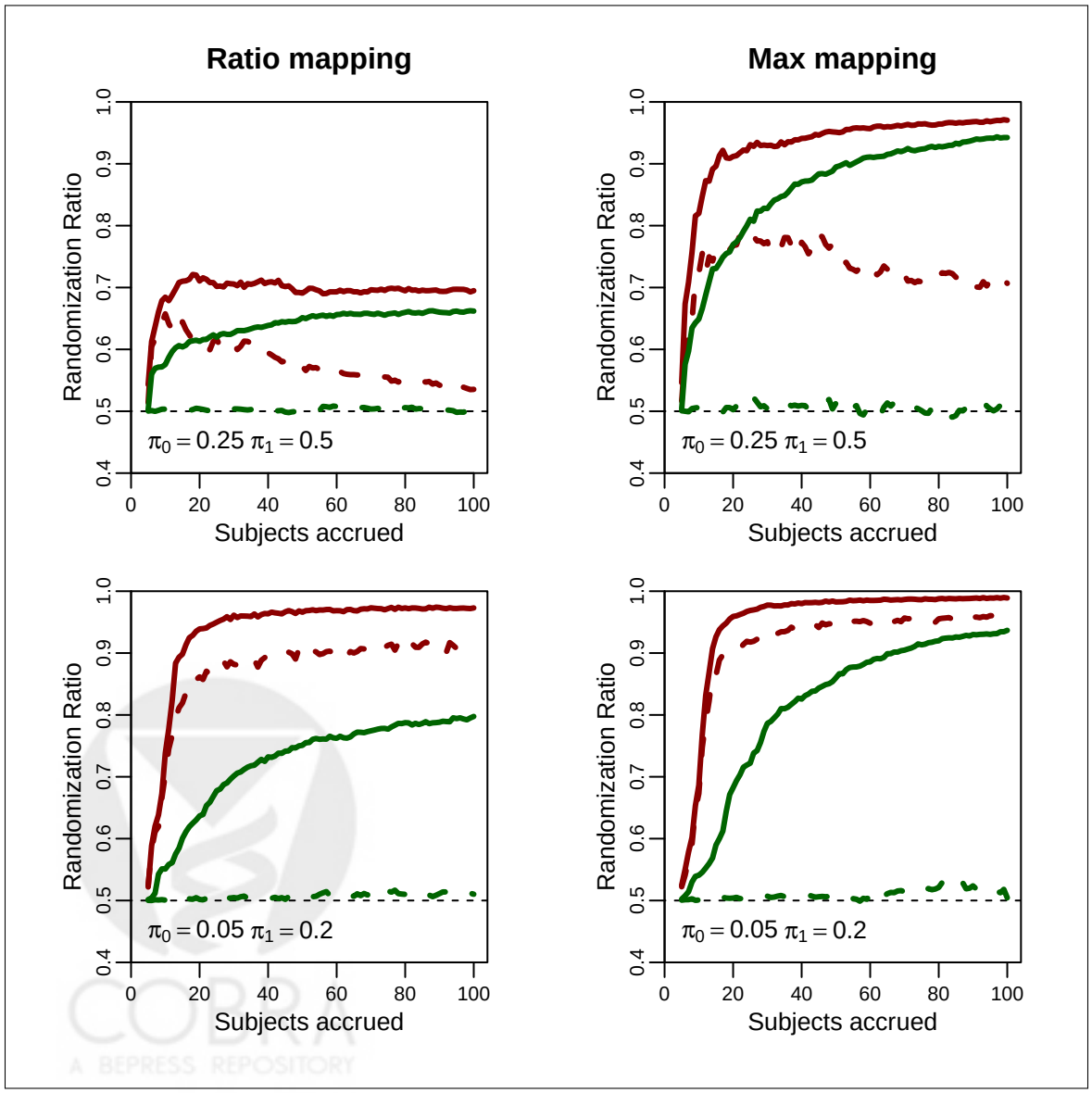

Figure 3: Operating characteristics of Bayesian adaptive versus fixed staged designs. The probabilities of determining efficacy are shown for target samples sizes ranging from N = 5 to 100. In both the single marker (left panels) and complementary marker (right panels) scenarios, effective treatment-marker combinations are shown in green, versus ineffective combinations in red. Vertical lines show the target and expected sample sizes (dark and light gray) that give 80% power and control Type I error at 10% in four parallel Simon two-stage tests. Lower panels display the cumulative distribution function (CDF) of sample sizes for the parallel Simon design (gray) under each scenario, versus sample sizes seen under simulation for adaptive designs (blue) with target N = 55 and 59, respectively.

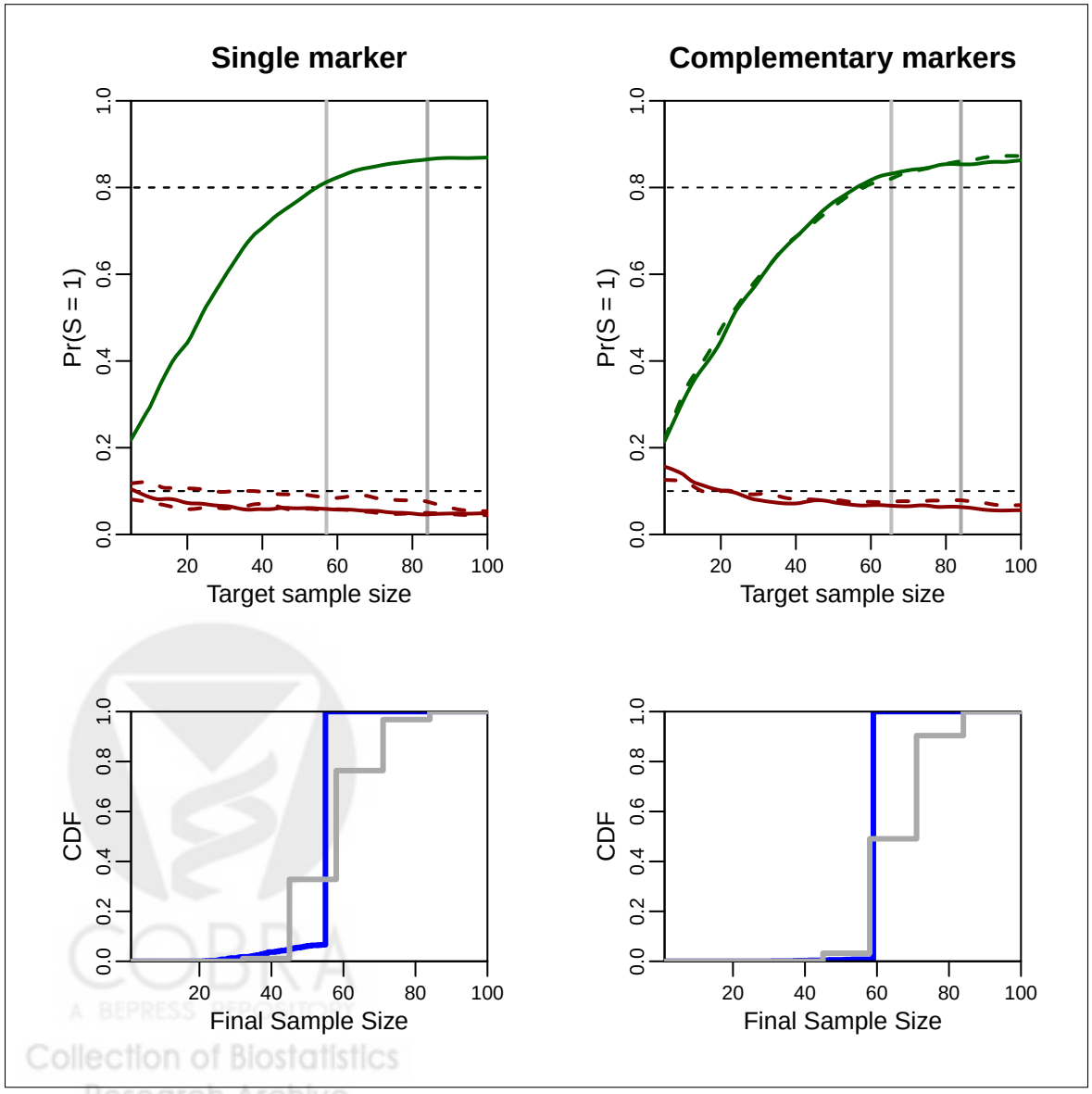

Figure 4: Schema for the adaptive randomized phase II to evaluate capecitabine with and without a PI3K inhibitor across four biomarker-defined subgroups of ER+/Her2- breast cancer.