1-29-2018

PHASE II ADAPTIVE ENRICHMENT DESIGN TO DETERMINE THE POPULATION TO ENROLL IN PHASE III TRIALS, BY SELECTING THRESHOLDS FOR BASELINE DISEASE SEVERITY

Yu Du
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Gary L. Rosner
Johns Hopkins University School of Medcine and Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mrosenbl@jhsph.edu

Phase II Adaptive Enrichment Design to Determine the Population to Enroll in Phase III Trials, by Selecting Thresholds for Baseline Disease Severity

Yu Du, Gary L. Rosner, Michael Rosenblum

January 27, 2018

Abstract

We propose and evaluate a two-stage, phase 2, adaptive clinical trial design. Its goal is to determine whether future phase 3 (confirmatory) trials should be conducted, and if so, which population should be enrolled. The population selected for phase 3 enrollment is defined in terms of a disease severity score measured at baseline. We optimize the phase 2 trial design and analysis in a decision theory framework. Our utility function represents a combination of the cost of conducting phase 3 trials and, if the phase 3 trials are successful, the improved health of the future population minus the cost of treatment. Given such a utility function and a discrete prior distribution on the conditional treatment effect, we compute the Bayes optimal adaptive design. The resulting design is compared to simpler designs in simulation studies. We also apply

the designs to resampled data from a completed, phase 2 trial evaluating a new surgical intervention for stroke.

Keywords: Adaptive Enrichment Design, Backward Induction, Treatment Effect Heterogeneity

1 Introduction

A new treatment may be effective only for a subset of the overall study population. This poses important challenges for designing a randomized clinical trial to evaluate such a treatment. On the one hand, ignoring participant heterogeneity and enrolling the overall population can lead to a dilution of the treatment effect, and may create an unethical situation where participants who do not benefit from the treatment are treated. On the other hand, enrolling only a narrow proportion of the overall population would not answer the question of whether the treatment benefits the larger population; furthermore, such restrictive enrollment could lead to slower recruitment and longer trial duration.

Our work is motivated by a multicenter, randomized trial where the new surgical intervention called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage Evacuation (MISTIE) was compared to standard medical care (Hanley et al., 2016). An important baseline (i.e., pre-randomization) characteristic is intracerebral hemorrhage (ICH) volume, which is one measure of disease severity. Based on their understanding of brain hemorrhage, the clinical investigators conjectured that the new treatment may have different effects depending on a participant's pre-randomization ICH volume. We compare different types of phase

2, randomized trial designs whose goal is to inform a future recommendation about which (if any) range of ICH volume should be used as the enrollment criterion in future phase 3, confirmatory trials. For computational reasons, we discretize ICH volume by partitioning the range of its possible values; in general, if one has a continuous-valued baseline score, our approach can be applied to a discretized version of it.

We consider two-stage, adaptive enrichment designs for the phase 2 randomized trial. Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on data accrued in an ongoing trial (Wang et al., 2009). We hypothesized that adaptive enrichment in the phase 2 trial might be useful for making an optimal recommendation for the population (if any) to enroll in future phase 3 confirmatory trials. In such a phase 2 design, information from the first stage of the phase 2 trial can be used to target who to enroll in the second stage of the phase 2 trial. For example, if early phase 2 data indicated a treatment benefit only for those with high baseline scores, participants near the boundary of such scores could be oversampled in the second stage of the phase 2 trial; this may lead to improved information for recommending which population to enroll in phase 3. We investigate whether such an adaptive feature adds value (in terms of expected utility, defined below) or not, compared to a non-adaptive phase 2 design.

Related work aimed at determining the population who benefits from a treatment includes, e.g., Jiang et al. (2007), Zhou et al. (2008), Barker et al. (2009), Freidlin et al. (2010), Lee et al. (2010), Kim et al. (2011), Cai et al. (2011), Lai et al. (2014), Xu et al. (2014), Ohwada and Morita (2016), Spencer et al. (2016). Our approach differs from these in

that we explicitly define the performance goal of our phase 2 design (the objective function) in a decision theory framework and compute the Bayes optimal adaptive enrichment design over a class of designs defined in Section 2.

We next discuss related work that uses a decision theory approach for optimizing trial designs. Colton (1963, 1965) aim to select the best of two treatments; Banerjee and Tsiatis (2006) aim to minimize expected sample size; Cheng and Berry (2007) aim to maximize the expected number of effectively treated participants in the trial; Hampson and Jennison (2015) aim to optimize power to detect the best of multiple treatments. Each of the aforementioned references involves a single population. In contrast, we consider multiple populations defined by a baseline score, and aim to determine which population (if any) to recommend for a pair of future phase 3 trials. Our designs also differ from the above related work in terms of the types of designs we consider, i.e., phase 2 designs that can adapt the population enrolled (called adaptive enrichment). The work of Graf et al. (2015), Krisam and Kieser (2015), Götte et al. (2015), and Rosenblum et al. (2016) involves adaptive enrichment using a binary biomarker that divides participants into two subpopulations. In contrast, by allowing the population selected for phase 3 to be an interval of the baseline score, we allow a larger number of possible subpopulations.

In Section 2, we define the class of trial designs that we optimize over. The trial design optimization problem and our approach to solving it are given in Section 3. In Section 4, we use simulations to compare the performance of an optimized, two-stage, adaptive design versus an optimized (non-adaptive) one-stage design. The main result is that the two-stage,

adaptive design did not improve expected utility, where the utility function measures the quality of the recommendation for who to enroll in future phase 3 trials; however, as shown in Section 4.4, the two-stage, adaptive design leads to fewer participants assigned to a non-efficacious or harmful treatment during phase 2. We apply the designs to resampled data from a completed, phase 2 trial (MISTIE II) evaluating a new surgical intervention for stroke, in Section 5. In Section 6, we present areas for future research.

2 Data Generating Process and Phase 2 Trial Designs

2.1 Overview of Fixed and Adaptive Designs

The populations we consider are based on a predefined variable that is obtained by discretizing a continuous-valued baseline score. The score could measure, e.g., disease severity. Our running example is pre-randomization ICH volume in the MISTIE II trial.

We focus on phase 2 randomized trials comparing a new treatment to standard of care (control) with a 1:1 randomization ratio. We compare two-stage, adaptive enrichment designs versus one-stage, fixed designs. Both designs have the same total sample size (denoted as $2n$), and have the goal of making an optimal recommendation for the population (if any) to enroll in two, future phase 3 trials. Two, future phase 3 trials are considered since this is what the U.S. Food and Drug Administration typically requires for approval of a new drug; the phase 3 trials are assumed to be fixed (non-adaptive) and have predefined sample size N.

Figure 1 illustrates two types of phase 2 designs: the one-stage, fixed design and the two-stage, adaptive enrichment design, referred to as the fixed design and adaptive design, respectively. In the fixed design, shown in Figure 1a, $2n$ participants are enrolled from the overall population. A preplanned rule is then applied to the data from this trial, in order to select the population to enroll in two, future phase 3 trials. The population is defined by a range of the baseline score.

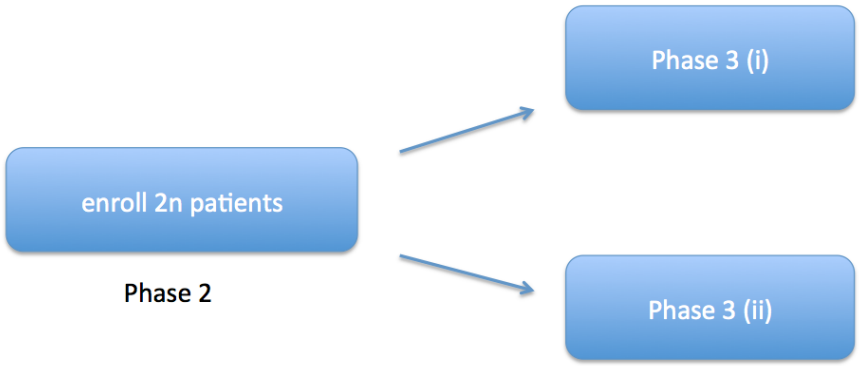
The adaptive design, shown in Figure 1(b), has two stages. In Stage A, n participants are recruited from the overall population. At the end of Stage A, a decision is made as to which population to enroll in Stage B, in terms of the baseline score. This allows an enrollment modification for Stage B (which always has n participants), with the goal of providing more targeted information to help in the population selection at the end of Stage B for the future phase 3 trials.

For both the fixed and adaptive phase 2 designs, the data from the $2n$ participants are ultimately used to select a range of the baseline score to set as the enrollment criterion for 2 future phase 3 trials. If the range selected is the empty set, this corresponds to not conducting any phase 3 trials; this option is important since one of the main goals of phase 2 is to weed out useless treatments.

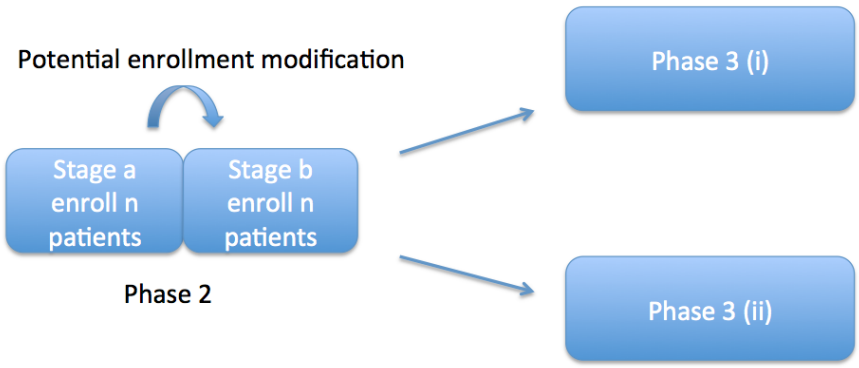
For the phase 2 adaptive design, the only feature that is adapted is the Stage B enrollment criterion. The decisions involved in the phase 2 adaptive designs are:

(i) after collecting the Stage A data, what the enrollment criterion will be for Stage B;

(ii) after Stage B data has been collected, what will be the enrollment criterion for the



(a) Fixed Design, $2n$ participants are enrolled in phase 2 from the overall population. At the end, a recommendation of the population (if any) to enroll in two, future phase 3 trials is made.



(b) Adaptive Design, n participants are enrolled in each of Stages A and B in phase 2 with the Stage B enrollment criterion depending on the results of Stage A. After Stage B, a recommendation of the population (if any) to enroll in two, future phase 3 trials is made.

Figure 1: Phase 2 fixed design (top-left) and phase 2 adaptive design (bottom-left), each of which may be followed by two phase 3 trials.

future phase 3 trials.

2.2 Data Collected on Each Participant

Let R denote the continuous-valued baseline score, e.g., ICH volume in the MISTIE trial example. We assume the population distribution of R is uniform on the interval $(0, 1)$, which can be achieved by a quantile transformation of any continuous variable. In order to make our computations feasible, we discretize R by partitioning $(0, 1)$ into M consecutive, equal length intervals. Let $\tilde{R} = \lceil RM \rceil$ denote the corresponding interval number, which has values in $\{1, \ldots, M\}$. We refer to $\tilde{R}$ as the discrete score.

Let T denote the treatment arm indicator, where $T = 1$ means assignment to treatment and $T = 0$ means assignment to control. Let $Y \in \mathbb{R}$ denote the primary outcome, which we assume to be measured on each participant relatively soon after her/his enrollment. Let $\Delta : (0, 1) \to \mathbb{R}$ denote the conditional treatment effect function $\Delta(r) = E(Y|T = 1, R = r) - E(Y|T = 0, R = r)$ given the continuous-valued baseline score $r \in (0, 1)$. Define the discrete analog of Δ, which represents the average treatment effect for stratum $\tilde{R} = \tilde{r}$, as:

$$\tilde{\Delta}(\tilde{r}) = E(Y|T = 1, \tilde{R} = \tilde{r}) - E(Y|T = 0, \tilde{R} = \tilde{r}) = M \int_{(\tilde{r}-1)/M}^{\tilde{r}/M} \Delta(r)dr, \tag{1}$$

for any $\tilde{r} \in \{1, \ldots, M\}$.

Conditional on the discrete baseline score and treatment indicator, the primary outcome Y is assumed to be a random, independent draw from a normal distribution with common variance σ^2. That is, we assume the conditional distribution of $Y|T = t, \tilde{R} = \tilde{r}$ is Normal$(\mu_t(\tilde{r}), \sigma^2)$, for unknown mean functions $\mu_0(\tilde{r}), \mu_1(\tilde{r})$ satisfying $\mu_1(\tilde{r}) - \mu_0(\tilde{r}) = \tilde{\Delta}(\tilde{r})$ and common, conditional variance σ^2. For simplicity, we assume $\mu_1(\tilde{r}) = \tilde{\Delta}(\tilde{r})/2$ and $\mu_0(\tilde{r}) = -\tilde{\Delta}(\tilde{r})/2$.

The data vector contributed by each participant i, used as input to the decision rules in the trial design, is denoted $V_i = (\tilde{R}_i, T_i, Y_i)$. Let $X^{(A)}$ and $X^{(B)}$ denote the sets of data vectors V_i collected during Stage A and Stage B, respectively. Let $X = (X^{(A)}, X^{(B)})$ denote the entire data set at the end of Stage B.

2.3 Definition of Phase 2 Adaptive Designs

Let $\mathcal{E}^{(A)}$ and $\mathcal{E}^{(B)}$ denote the allowed enrollment choices at the end of Stage A and Stage B, respectively. The action sets $\mathcal{E}^{(A)}, \mathcal{E}^{(B)}$ each consist of a prespecified, finite set of intervals (r_l, r_u) with endpoints $r_l, r_u \in \{0, 1/M, 2/M, \ldots, 1\}$ and $r_l \leq r_u$; each interval represents a range of the baseline score. If the interval $(r_l, r_u) \in \mathcal{E}^{(A)}$ is selected at the end of Stage A, it means that n participants will be enrolled during Stage B using inclusion criterion $R \in (r_l, r_u)$. If the interval $(r_l, r_u) \in \mathcal{E}^{(B)}$ is selected at the end of Stage B, it means that the two, phase 3 trials will enroll N participants using inclusion criterion $R \in (r_l, r_u)$. We assume that $\mathcal{E}^{(A)}$ contains the full interval $(0, 1)$ and that $\mathcal{E}^{(B)} = \mathcal{E}^{(A)} \cup \{\emptyset\}$, where the empty set $\emptyset$ represents not conducting any phase 3 trials. By convention, we represent the empty set by an interval (r_l, r_u) with $r_l = r_u$.

An adaptive design is defined as a pair of decision rules $(d^{(A)}, d^{(B)})$, where $d^{(A)}$ is a map from the Stage A data $X^{(A)}$ to the set of enrollment choices $\mathcal{E}^{(A)}$ and $d^{(B)}$ is a map from the cumulative data X to $\mathcal{E}^{(B)}$. We assume these maps are measurable, which is generally required for adaptive designs (Liu et al., 2002). Whenever maxima are taken over pairs of decision rules $(d^{(A)}, d^{(B)})$, it is assumed that this is over all possible pairs of such measurable

maps.

After collecting the Stage A data $X^{(A)}$, the prespecified decision rule $d^{(A)}$ is applied to select the action from $\mathcal{E}^{(A)}$; this determines the population to be enrolled during Stage B. At the end of Stage B, the cumulative data $X = (X^{(A)}, X^{(B)})$ has been collected, and the prespecified decision rule $d^{(B)}$ is applied to determine the action in $\mathcal{E}^{(B)}$; this action represents the enrollment criterion for the two, future phase 3 confirmatory trials. The fixed design is a special case of the adaptive design with $d^{(A)}$ always mapping to the action $(0, 1)$.

2.4 Prior Distribution on Conditional Treatment Effect Function

We assume a prior π on the conditional treatment effect function Δ. In our simulations, π is set to be a finite set of point masses on the functions $\Delta = \delta_1, \ldots, \delta_6$, shown in Figure 2. We refer to each δ_k as a possible state of nature.

In Figure 2, the function δ_1 represents the conditional treatment effect being 1 for all values of the baseline score R in $(0, 1)$. The function δ_2 represents the treatment only benefiting the population with baseline scores greater than $1/2$, while the function δ_4 represents the treatment benefiting the population whose baseline scores are between $1/4$ and $3/4$. Under the conditional treatment effect function δ_5, the treatment benefit is -1, i.e., the treatment is harmful, for all values of the baseline score. Under δ_6, there is zero treatment benefit for every baseline score value; this represents the global null hypothesis of no treatment effect for any stratum of the baseline score. The discretized versions of $\delta_1, \ldots, \delta_6$, based on applying (1) to each, are given in Table 5 of the Supplementary Material.

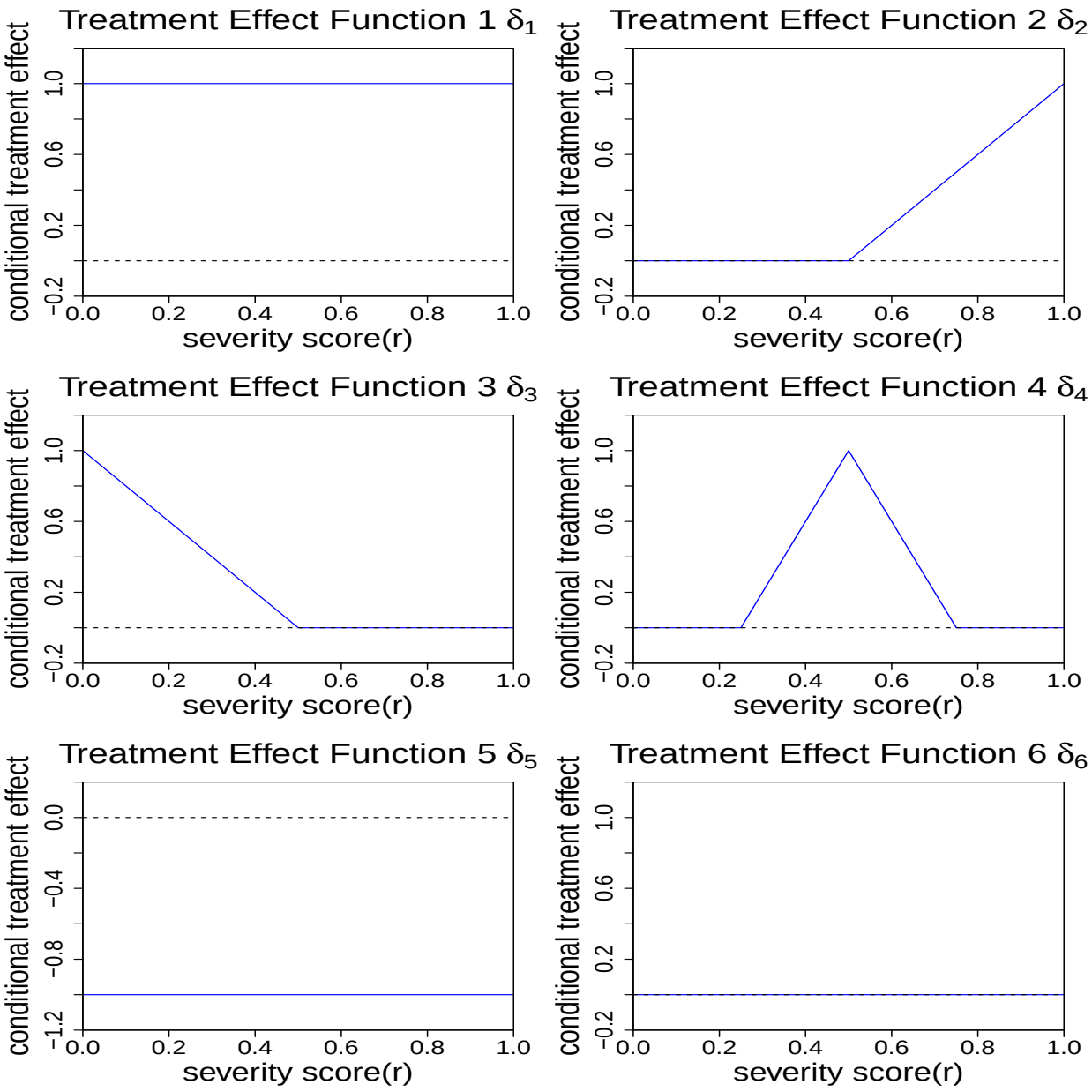


Figure 2: Six states of nature $\delta_1, \ldots, \delta_K$. The solid line represents the treatment effect function δ_k, while the dashed line is a horizontal zero line as a reference.

3 Optimization Problem and Algorithm to Solve It

3.1 Utility Function

Our utility function focuses on what happens as a consequence of the phase 2 trial. It represents a combination of the cost of conducting phase 3 trials (if they are recommended after phase 2) and the improved health of the future population minus the treatment's cost (if the phase 3 trials succeed). The utility function $U(r_l, r_u; \Delta)$ takes the function Δ and an action $(r_l, r_u) \in \mathcal{E}^{(B)}$ (which represents the interval recommended for phase 3 enrollment) as inputs, and is defined as follows: if no phase 3 trials are recommended (which occurs when $r_l = r_u$) then $U(r_l, r_u; \Delta) = 0$, and otherwise $U(r_l, r_u; \Delta)$ equals

$$P(\text{2 Phase 3 Trials Succeed}|\Delta, r_l, r_u) \times \left\{ \overbrace{\int_{r_l}^{r_u} \Delta(r)dr}^{\text{Treatment Benefit}} - \overbrace{c(r_u - r_l)}^{\text{Treatment Cost}} \right\} - \overbrace{\lambda/(r_u - r_l)}^{\text{Phase 3 Cost}}, \quad (2)$$

where $P(\text{2 Phase 3 Trials Succeed}|\Delta, r_l, r_u)$ is the conditional probability, defined below, that both phase 3 trials enrolling those with baseline scores in (r_l, r_u) succeed given Δ.

The components in curly braces in (2) represent the population health and treatment cost due to the treatment being approved (after 2 successful phase 3 trials) for use in the population with baseline scores in the range (r_l, r_u). The term $\int_{r_l}^{r_u} \Delta(r)dr$ represents the health benefit that would result from the population with baseline score in the interval (r_l, r_u) using the treatment instead of the control. This term is maximized if the interval (r_l, r_u) contains precisely those who benefit from treatment (and possibly also those who have no effect from treatment).

The second term in curly braces $c(r_u - r_l)$ represents health system costs to the future

population if both phase 3 trials succeed leading to the drug being approved and used by the recommended population (r_l, r_u). Here, c is a constant, and we assume that the health system cost is proportional to the size of the population with baseline score in (r_l, r_u). We use the term "cost" in the general sense that captures negative impacts of administering the treatment to a future population. For example, cost may reflect monetary cost and negative impacts in terms of frequency and severity of side effects caused by treatment. If the cost $c > 0$, then there is a penalty for approving the treatment for any stratum $r \in (0, 1)$ whose treatment effect $\Delta(r)$ is less than c; in particular, it penalizes for approving the treatment for strata who have zero treatment effect (unlike the first term in curly braces).

The component $P(2 \text{ Phase 3 Trials Succeed}|\Delta, r_l, r_u)$ is the probability of success of two, future phase 3 trials enrolling from the population with baseline score in the interval (r_l, r_u). Assuming that the two phase 3 trials are independent given (r_l, r_u), this probability is the squared power of a fixed (non-adaptive) design for the phase 3 trial where the average treatment effect is $(r_u - r_l)^{-1} \int_{r_l}^{r_u} \Delta(r)dr$, i.e., the average height of the treatment effect curve Δ over the interval (r_l, r_u). The prespecified sample size N for each phase 3 trial is determined by the minimum, clinically meaningful benefit $\Delta_{\min}$, type I error, type II error, and outcome variance σ^2.

The last term of the utility function, $\lambda/(r_u - r_l)$, is proportional to the duration of each phase 3 trial enrolling participants with baseline scores in (r_l, r_u), where we made the approximation that the enrollment rate is proportional to the size of the recommended population to be enrolled. This could apply, for example, to the MISTIE trial where ICH

volume is only determined after recruitment based on neuroimaging, and so excluding more participants would require longer recruitment to reach the total sample size N.

Though Type I error control in the phase 2 trial is not our focus, we do evaluate design performance when there is no treatment effect to determine how often future phase 3 trials are (erroneously) recommended.

3.2 Optimization Problem

The goal is to compute the pair of optimal decision rules at the end of Stage A and at the end of Stage B in phase 2, denoted by $d_{opt}^{(A)}$ and $d_{opt}^{(B)}$, respectively. These are the rules that maximize expected utility, where the expectation is with respect to the distribution of (Δ, X) induced by the prior π on Δ. Define

$$(d_{opt}^{(A)}, d_{opt}^{(B)}) \;=\; \underset{d^{(A)}, d^{(B)}}{\operatorname{argmax}} \mathbb{E}\left[U\left\{d^{(B)}\left(X^{(A)}, X^{(B)}[d^{(A)}\{X^{(A)}\}]\right); \Delta\right\}\right], \tag{3}$$

where the maximum is taken over all pairs $(d^{(A)}, d^{(B)})$ of decision rules as defined in Section 2.3. Throughout, expectation $\mathbb{E}$ and probability P are with respect to the prior π and a generic phase 2 adaptive design $(d^{(A)}, d^{(B)})$ unless indicated otherwise. As described in Section 3.1, the utility U depends on the action $d^{(B)}(X)$ taken at the end of Stage B (i.e., who to enroll in the future phase 3 trials) and the conditional treatment effect function Δ. The action taken at the end of Stage B depends on the data from Stages A and B, i.e., $X = (X^{(A)}, X^{(B)})$. We write $X^{(B)} = X^{(B)}[d^{(A)}\{X^{(A)}\}]$ in the display above to make explicit that the Stage B data generating distribution depends on the decision $d^{(A)}\{X^{(A)}\}$ at the end of Stage A regarding the population to be enrolled during Stage B. This highlights the

sequential nature of the decision problem.

3.3 Algorithm to Solve the Optimization Problem

Banerjee and Tsiatis (2006), Cheng and Berry (2007) and Hampson and Jennison (2015) used backward induction to provide closed form solutions to their optimal adaptive designs under a decision-theoretic framework. Since this is not possible in our problem, we instead use backward induction with Monte-Carlo forward simulation. Such backward induction has been implemented to solve problems in different contexts by, e.g., Carlin et al. (1998), Brockwell and Kadane (2003), Rossell et al. (2006), Ding et al. (2008).

Let $n^{(j)}(\tilde{r})$ denote the cumulative sample size per arm for participants with baseline score $\tilde{R} = \tilde{r}$ enrolled during or before Stage $j \in \{A, B\}$ of the phase 2 trial, for each $\tilde{r} \in \{1, \ldots, M\}$. We assume that in Stage A, n/M are enrolled from each baseline stratum $\tilde{r} \in \{1, \ldots, M\}$. For Stage B of the adaptive design, we assume an equal number are enrolled from each stratum $\tilde{r}$ contained in the selected population $d^{(A)}(X^{(A)})$, such that a total of n are enrolled in that stage. We assume that within each stage and enrolled stratum of the baseline score, an equal number are assigned to treatment and control; this can be accomplished (approximately) by stratified block randomization.

Define the sample mean difference between arms of the primary outcome using cumulative data through the end of Stage j for participants with $\tilde{R} = \tilde{r}$ as

$$\widehat{\Delta}^{(j)}(\tilde{r}) = \frac{1}{n^{(j)}(\tilde{r})} \left\{ \sum_i 1(T_i = 1, \tilde{R}_i = \tilde{r}) Y_i - \sum_i 1(T_i = 0, \tilde{R}_i = \tilde{r}) Y_i \right\},$$

where the summations are over the participants i with outcomes observed at or before the

end of Stage j, and $1(S)$ is the indicator variable taking value 1 if S is true and 0 otherwise. The statistic $\widehat{\Delta}^{(j)}(\tilde{r})$ is an estimator of $\Delta(\tilde{r})$, the average treatment effect in stratum $\tilde{r}$. Let $\widehat{\Delta}^{(j)}$ denote the vector of cumulative sample mean differences for each category of baseline score $\tilde{r} \in \{1, \ldots, M\}$ using all data up through the end of Stage $j \in \{A, B\}$.

We next define minimal sufficient statistics $\tilde{S}^{(A)}, \tilde{S}^{(B)}$ for Δ based on the data $X^{(A)}$ available at the end of Stage A and the data X available at the end of Stage B, respectively. Stage A is equivalent to a fixed design, and we define $\tilde{S}^{(A)} = \widehat{\Delta}^{(A)}$, i.e., the sample mean differences within each stratum $\tilde{r} \in \{1, \ldots, M\}$. Since Stage B involves a potential enrollment modification that is determined by the decision $d^{(A)}(X^{(A)})$, we define $\tilde{S}^{(B)} = (d^{(A)}(X^{(A)}), \widehat{\Delta}^{(B)})$.

Backward induction starts from the end of Stage B, where we have collected the overall data $X = (X^{(A)}, X^{(B)})$. The first step is to maximize the conditional expected utility over all possible Stage B decision rules given X. It follows from (3) that

$$d^{(B)}_{opt}(X) = \underset{d^{(B)}}{\operatorname{argmax}}\, \mathbb{E}[U\{d^{(B)}(X); \Delta\}|X], \tag{4}$$

where the expectation is with respect to the distribution of Δ given X.

We assume the prior distribution π on Δ consists of K point masses $\delta_1, \ldots, \delta_k$, each representing a conditional treatment effect function. For any candidate action $(r_l, r_u) \in \mathcal{E}^{(B)}$, Monte Carlo simulation is used where we draw posterior samples of Δ given the data X in order to approximate the conditional expected utility if this action is followed. i.e., $\mathbb{E}[U(r_l, r_u; \Delta)|X]$, . The posterior distribution $P(\Delta \mid X)$ depends on the data X only through the sufficient statistics $\tilde{S}^{(B)}$. Given the data X, we compute the corresponding sufficient statistics $\tilde{S}^{(B)}$ and then draw posterior samples of Δ from $P(\Delta \mid \tilde{S}^{(B)})$ as described

below.

Consider any Stage A decision rule $d^{(A)}$ and data vector $X = (X^{(A)}, X^{(B)})$ that can be generated under the phase 2 adaptive design using Stage A decision rule $d^{(A)}$. Let $E = d^{(A)}(X^{(A)})$ denote the decision at the end of Stage A; without loss of generality, we assume $d^{(A)}$ depends on the data $X^{(A)}$ only through the sufficient statistic $\hat{\Delta}^{(A)}$. We prove the following in the Supplementary Material, where $\propto$ represents proportionality with respect to functions of the data X:

$$P(\Delta = \delta_k \mid X) \propto P^E(\widehat{\Delta}^{(B)} = \widehat{\Delta}^{(B)}(X) \mid \Delta = \delta_k) P(\Delta = \delta_k), \tag{5}$$

where P^e denotes the probability density function of X under the deterministic (non-adaptive) Stage A decision rule that always enrolls population $e \in \mathcal{E}^{(A)}$ during Stage B regardless of the Stage A data. The above display reduces (up to a proportionality constant) the problem of computing $P(\Delta = \delta_k \mid X)$ to the following simpler problem: computing the conditional probability density that the cumulative sample mean differences (at the end of Stage B) equal $\widehat{\Delta}^{(B)}(X)$ given Δ under the non-adaptive Stage A decision rule that always enrolls population $E = d^{(A)}(X)$ during Stage B.

We describe how to compute $d^{(B)}_{opt}(X)$ given X. Under the probability density P^E, the statistic $\widehat{\Delta}^{(B)}$ conditional on $\Delta = \delta_k$ has a multivariate normal distribution with mean vector determined by integrating δ_k over each interval $((m-1)/M, m/M) : m = 1, \ldots, M$ using the formula on the right side of (1) and covariance matrix $\mathbf{\Sigma}$ the diagonal matrix with zeros off the main diagonal and $\mathbf{\Sigma}_{mm} = 2\sigma^2/n^{(B)}(m)$ for each $m = 1, 2, \ldots, M$. Given the observed $\widehat{\Delta}^{(B)}$, for each $k = 1, \ldots, K$ we compute the density $P^E(\widehat{\Delta}^{(B)} = \widehat{\Delta}^{(B)}(X) \mid \Delta = \delta_k)$ from this

multivariate normal distribution and multiply by the prior $P(\Delta = \delta_k)$ to obtain the right side of (12), which was shown above to be proportional to the posterior probability $P(\Delta = \delta_k \mid X)$. For each interval (r_l, r_u) in the action space $\mathcal{E}^{(B)}$, we repeatedly draw from this posterior distribution to approximate the posterior expected utility $\sum_{k \leq K} U(r_l, r_u; \delta_k) P(\Delta = \delta_k \mid X)$, which is $\mathbb{E}[U\{d^{(B)}(X); \Delta\}|X]$ under the decision $d^{(B)}(X) = (r_l, r_u)$. We set $d^{(B)}_{opt}(X)$ to be the interval (r_l, r_u) in $\mathcal{E}^{(B)}$ that gives the maximum posterior expected utility.

The derivation of $d^{(A)}_{opt}$ is analogous to that of $d^{(B)}_{opt}$, and is described in the Supplementary Material. All of the above computations were implemented in R (R Core Team, 2015).

4 Simulation Study with Six Possible Treatment Effect Curves

4.1 Simulation Setup

We implement the optimization algorithm from the previous section in a simulation study comparing the performance of the fixed versus adaptive phase 2 design. The prior π on Δ is a discrete distribution on the conditional treatment effect curves $\{\delta_k\}_{k=1}^6$ in Figure 2 that places 50% weight on δ_6 (the global null hypothesis) and equal weights on $\delta_1, \ldots, \delta_5$. We allocate greater weight to δ_6 in order to reflect the realistic possibility that an experimental treatment has no effect at all. Intuitively, the impact of having greater weight on δ_6 is that the corresponding optimal design will be more conservative in initiating phase 3 follow-up trials, thereby saving resources at the cost of lowering the chance of successful phase 3 trials.

The prior π is one possible choice; our general approach can be applied to an arbitrary set of weights and any finite set of conditional treatment effect functions, i.e., applied to any prior consisting of a finite set of point masses. The choice of prior would ideally be chosen based on prior scientific knowledge and data from earlier studies. Here we demonstrate a relatively simple case as a proof of concept.

For a given state of nature Δ, *the population who benefit from treatment* is defined as $\{r \in (0,1) : \Delta(r) > 0\}$. For example, under $\Delta = \delta_2$, the treatment benefits precisely those with $r \in (0.5, 1)$. Depending on which of $\delta_k, k = 1, \ldots, 6$, is the true state of nature, the population who benefit from treatment is an interval of the baseline score r in the set

$$\mathcal{E}^{(B)} = \{(0,1), (0.5,1), (0,0.5), (0.25,0.75), \emptyset\}. \tag{6}$$

Let

$$\mathcal{E}^{(A)} = \{(0,1), (0.5,1), (0,0.5), (0.25,0.75)\}.$$

We use the above sets $\mathcal{E}^{(A)}, \mathcal{E}^{(B)}$ as the action sets for our decision problem, representing the possible choices for who to enroll next based on the data at the end of Stages A and B, respectively.

We define $\tilde{R}$ to have $M = 4$ levels corresponding to the baseline score being in the intervals $(0, 0.25), (0.25, 0.5), (0.5, 0.75), (0.75, 1)$, respectively. We chose this discretization since any non-empty interval in $\mathcal{E}^{(A)}$ or $\mathcal{E}^{(B)}$ can be represented as a union of these intervals (ignoring the interval endpoints).

We selected the phase 2 total sample size, denoted $2n$, to be 528. This was, roughly, based on the sample size required for the fixed design (Figure 1a) to have type I error α

and power $1-\beta$, in the simple case of a constant conditional treatment effect $\Delta = \tau > 0$ and conditional variance σ^2. This sample size is $4\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2(\sigma^2/\tau^2)$; to maintain a realistic signal to noise ratio in the phase 2 trial, we set $\alpha = 0.1$, $\beta = 0.2$, $\sigma^2 = 9$ and $\tau = 0.554$ in this formula (and rounded down), which implies the total phase 2 sample size is $2n = 528$. Our choices of α, β are typical for phase 2 trials as described by Rubinstein et al. (2005). However, our choices of τ and σ were somewhat arbitrary; the value of τ was taken to be between the average treatment effect of 1 (under $\Delta = \delta_1$) and 0.25 (under each $\Delta = \delta_k, k = 2, 3, 4$). We set the prespecified sample size $N = 5564$ for each phase 3 trial based on the above formula with parameters $\tau = 0.2$ (representing the minimum, clinically meaningful benefit), type I error $\alpha = 0.05$, type II error $\beta = 0.2$, and outcome $\sigma = 3$.

Our optimization problem is invariant to rescaling in that the optimal decision rules (as functions of the sufficient statistics $\tilde{S}^{(A)}, \tilde{S}^{(B)}$) are unchanged if we multiply all of n, σ^2, N by the same positive constant. For example, the optimal decision rules would be the same if we multiply these parameters by $1/4$, i.e., setting the phase 2 total sample size $2n = 132$, outcome conditional variance $\sigma^2 = 9/4$, and phase 3 sample size $N = 1391$.

4.2 Optimal Adaptive Phase 2 Designs in Simulation Study

The optimal decision rules $d_{opt}^{(A)}$ and $d_{opt}^{(B)}$ are defined by (3), which depends on the prior π and the utility function U defined above. These rules are optimal for a pair (π, U) in terms of the Bayes criterion (3). These rules may be suboptimal for any single state of nature δ_j, since the rules find the best tradeoff in expected utility U across these different states of

nature, where the relative importance of each state of nature depends on π. The optimal decision rules were computed using the backward induction method in Section 3.3.

We explored the performance of the optimal decision rules $(d_{opt}^{(A)}, d_{opt}^{(B)})$ by conducting simulated trials with data generated from one treatment effect function δ_k at a time. For each δ_k, we simulated 200 trials under $\Delta = \delta_k$ and using the precomputed decision rules $(d_{opt}^{(A)}, d_{opt}^{(B)})$. The frequency of each population in $\mathcal{E}^{(A)}$ getting selected for Stage B enrollment was recorded; also recorded was the frequency of each population in $\mathcal{E}^{(B)}$ getting selected for the future phase 3 trials. The optimal rules $d_{opt}^{(A)}$ and $d_{opt}^{(B)}$ depend on the prespecified π and U, and are the same functions regardless of our setting Δ to be different curves $\delta_1, \ldots, \delta_k$ in the simulation study.

Table 1 shows the operating characteristics of the optimal, adaptive phase 2 design $(d_{opt}^{(A)}, d_{opt}^{(B)})$ using the utility function (2) with $\lambda = 0.01, c = 0.32$. The top half of Table 1 shows the frequency of different choices for Stage B enrollment corresponding to $d_{opt}^{(A)}$. For example, 34% of the simulated trials recommend to enroll participants from the overall population in Stage B of the phase 2 adaptive design when $\Delta = \delta_1$ is the data generating distribution. For δ_2 and δ_3, where only half of the overall population benefits from the treatment, there is a 67% chance of enrolling at least the population who benefit from treatment in Stage B.

The bottom half of Table 1 shows the frequency of different choices for the population to enroll in the two, phase 3 trials under the decision rule $d_{opt}^{(B)}$. We underline the number corresponding to the population who benefit from treatment (defined in Section 4.1) under

Table 1: Operating Characteristics of $d_{opt}^{(A)}$ (top) and $d_{opt}^{(B)}$ (bottom) for $\lambda = 0.01, c = 0.32$.

Distribution of Population Recommended for Stage B Enrollment by $d_{opt}^{(A)}$

	$d_{opt}^{(A)} = (0,1)$	$d_{opt}^{(A)} = (0.5,1)$	$d_{opt}^{(A)} = (0,0.5)$	$d_{opt}^{(A)} = (0.25,0.75)$
$\Delta = \delta_1$	0.34	0.27	0.24	0.15
$\Delta = \delta_2$	0.30	0.37	0.21	0.12
$\Delta = \delta_3$	0.34	0.22	0.33	0.11
$\Delta = \delta_4$	0.27	0.26	0.26	0.21
$\Delta = \delta_5$	0.67	0.14	0.13	0.06
$\Delta = \delta_6$	0.30	0.34	0.27	0.09

Distribution of Population Recommended for Phase 3 trials by $d_{opt}^{(B)}$

	$d_{opt}^{(B)} = (0,1)$	$d_{opt}^{(B)} = (0.5,1)$	$d_{opt}^{(B)} = (0,0.5)$	$d_{opt}^{(B)} = (0.25,0.75)$	$d_{opt}^{(B)} = \emptyset$
$\Delta = \delta_1$	**0.96**	0.01	0.01	0.02	0.00
$\Delta = \delta_2$	0.07	**0.53**	0.04	0.09	0.27
$\Delta = \delta_3$	0.11	0.02	**0.56**	0.09	0.22
$\Delta = \delta_4$	0.12	0.04	0.09	**0.40**	0.35
$\Delta = \delta_5$	0.00	0.00	0.00	0.00	**1.00**
$\Delta = \delta_6$	0.01	0.09	0.06	0.12	**0.72**

the corresponding treatment effect function δ_k. We mark in bold the largest proportion in each row, which represents the population that is recommended most frequently for the future phase 3 trials. For example, when data is generated under treatment effect function δ_2 (row 2), the proportion 0.53 is both underlined and in bold, which means that the corresponding

Table 2: Operating Characteristics of $d_{opt}^{(B)}$ for $\lambda = 0.01, c = 0.34$.

	Distribution of Population Recommended for Phase 3 trials by $d_{opt}^{(B)}$				
	$d_{opt}^{(B)} = (0,1)$	$d_{opt}^{(B)} = (0.5,1)$	$d_{opt}^{(B)} = (0,0.5)$	$d_{opt}^{(B)} = (0.25,0.75)$	$d_{opt}^{(B)} = \emptyset$
$\Delta = \delta_1$	**0.95**	0.01	0.02	0.02	0.00
$\Delta = \delta_2$	0.07	**0.48**	0.04	0.08	0.33
$\Delta = \delta_3$	0.09	0.03	**0.55**	0.08	0.25
$\Delta = \delta_4$	0.11	0.05	0.07	**0.39**	0.38
$\Delta = \delta_5$	0.00	0.00	0.00	0.00	**1.00**
$\Delta = \delta_6$	0.01	0.07	0.05	0.08	**0.79**

population $(0.5, 1)$ is the population who benefits and is the population most frequently recommended for phase 3 trials. In every row in Table 1, the bold number coincides with the underlined number, which shows that in the plurality of simulated trials the optimal design chooses the population who benefits.

The results in Table 1 are for a particular choice of utility function parameters λ and c, which encode the relative importance of the cost of conducting phase 3 trials and the treatment cost if the treatment is approved (including both financial costs and health costs such as side effects), respectively. We next show the impact of increasing c from 0.32 to 0.34, while holding all other parameters constant. Using the utility function (2) with $\lambda = 0.01, c = 0.34$, we computed the operating characteristics of the component $d_{opt}^{(B)}$ of the optimal design, summarized in Table 2. There is a higher chance of making the decision not to conduct any phase 3 trials (rightmost column), compared to Table 1, under each δ_k (except δ_5 where

no phase 3 trials are conducted with probability 1 in both tables). Under the global null hypothesis $\Delta = \delta_6$, the probability of conducting phase 3 trials (which would be a waste of resources) drops from 28% to 21% comparing $c = 0.32$ to $c = 0.34$. The tradeoff is that the optimal decision rule for $c = 0.34$ (Table 2) has lower probabilities of selecting the optimal population for enrollment in phase 3 when the overall population or a subpopulation benefits ($\Delta \in \{\delta_1, \delta_2, \delta_3, \delta_4\}$).

4.3 Optimal Adaptive versus Fixed Phase 2 Trial Design

We solved the same optimization problem as in Section 4.2 using $\lambda = 0.01, c = 0.32$, except restricting to a fixed design, i.e., setting $d^{(A)}$ to be the constant function that always selects the full population $(0, 1)$ to enroll in Stage B of phase 2; only the function $d^{(B)}$ is optimized. The resulting design is referred to as the optimal fixed design. We compare its performance to that of the optimal adaptive phase 2 design in order to determine the value added by adaptive enrichment, i.e., the value added by allowing enrollment to be restricted to a subset of the population in Stage B of phase 2.

The recommendation frequencies for phase 3 trials for the optimal fixed design are very similar to those for the optimal adaptive design in Table 1, with the maximum difference between any 2 corresponding entries being 3%. Also, we computed the expected utility for the optimal adaptive design versus the fixed design, based on 15,000 simulated trials. The expected utility is 0.07 for both designs. We also compared the contribution from each component of the utility function for these two designs, summarized in Table 3.

Table 3: Expected utility and expected value of its 4 components under the optimal fixed design and adaptive designs, respectively, based on the simulation setup in Section 4.1 using $\lambda = 0.01, c = 0.32$. Expectation is with respect to the prior π. The interval (r_l, r_u) in the formulas below represents the phase 3 enrollment decision $d^{(B)}_{opt}(X)$.

	Fixed Design	Adaptive Design
Expected Utility $\mathbb{E}[U]$	0.07	0.07
Components of Expected Utility:		
Expected Treatment Benefit $\mathbb{E}[\int_{r_l}^{r_u} \Delta(r)dr]$	0.14	0.14
Expected Treatment Cost $\mathbb{E}[c(r_u - r_l)]$	0.10	0.10
P(2 Phase 3 Trials Conducted and Both Succeed)	0.27	0.27
Expected Phase 3 Cost $\mathbb{E}[\lambda 1(r_l \neq r_u)/(r_u - r_l)]$	0.82	0.80

The probability of having 2 successful phase 3 trials is 0.27, averaged over the prior π. This is the probability of successfully demonstrating treatment efficacy when considering the combined phase 2/3 trial sequence. This may seem like a low probability of the trial sequence succeeding. However, we next show that our choice of prior π implies that regardless of what is done during phases 2 and 3 (e.g., even if both sample sizes were arbitrarily increased), there is no way to achieve a probability greater than 41% of having a successful trial sequence. Intuitively, this is because the prior selects $\Delta = \delta_5$ or $\Delta = \delta_6$ with probability 0.6, and in such cases the treatment is not beneficial for any stratum of the baseline score.

We upper bound the probability of two successful phase 3 trials under the prior π and an arbitrary rule for deciding which population to enroll in phase 3. Under π, the event

$\max_{r\in(0,1)} \Delta(r) \leq 0$, which occurs when Δ is δ_5 or δ_6, has probability 0.6. On this event, there is at most 0.05 probability of each phase 3 trial succeeding. Therefore, the probability of two successful phase 3 trials (regardless of both the decision rule for phase 3 enrollment and the phase 3 sample size) is at most $1 - 0.6(1 - 0.05^2) = 0.4015$. By comparison, the corresponding 0.27 probability for the utility-maximizing designs above (where the objective function involves terms other than just power in phase 3) is not insubstantial.

4.4 Impact of Adaptive Design on Number Assigned to Superior Treatment During Phase 2

In this section, we shift our focus to what happens to participants during phase 2, rather than after phase 2. Specifically, we measure the impact of being enrolled in the phase 2 trial compared to not being enrolled; we assume not being enrolled in the trial means that a patient would have received the standard of care, i.e., the control. We focus only on Stage B participants in the phase 2 trial since our goal is to contrast the fixed versus adaptive designs, and these designs have identical patient outcome distributions during Stage A.

We say that Stage B participant i with baseline score R_i in study arm T_i is assigned to a superior treatment (compared to control) if $\Delta(R_i) > 0$ and $T_i = 1$, that is, if the conditional treatment effect is positive in that participant's baseline stratum and she/he is assigned to the treatment arm $T_i = 1$. For example, if $\Delta = \delta_2$, each participant i with baseline score $R_i > 0.5$ in arm $T_i = 1$ is assigned to a superior treatment. We denote the proportion of Stage B participants who are assigned to a superior treatment as $f_{\text{prop}} = \frac{1}{n}\sum_i 1\{T_i = 1, \Delta(R_i) > 0\}$,

where the sum is over the n Stage B participants. Similarly, define the average benefit to Stage B participants of being enrolled in the trial compared to not being enrolled, as $f_{\text{ben}} = \frac{1}{n}\sum_i 1(T_i = 1)\Delta(R_i)$, where the sum is over the n Stage B participants.

Table 4 presents the expected proportion $\mathbb{E}(f_{\text{prop}}|\Delta)$ of Stage B participants who are assigned to a superior treatment and the expected average benefit $\mathbb{E}(f_{\text{ben}}|\Delta)$, respectively, conditional on the treatment effect function Δ. These expectations depend on the decision rule $d^{(A)}$. We evaluated the optimal rule $d^{(A)} = d^{(A)}_{opt}$ from Section 4.3, which was optimized for the utility function U in (2) using $\lambda = 0.01, c = 0.32$. We also evaluate the fixed (non-adaptive) decision rule $d^{(A)} \equiv (0, 1)$. The evaluation of these decision rules was based on the same set of simulations from Section 4.3.

For every treatment effect function δ_k where some but not all strata of r benefit from the treatment (i.e., for $\delta_k : k \in \{2, 3, 4\}$), the difference between the expected proportion assigned to a superior treatment $\mathbb{E}(f_{\text{prop}}|\Delta = \delta_2)$ for $d^{(A)}_{opt}$ versus the fixed design is 3-5%, as shown in Table 4. For the expected benefit f_{ben}, the relative improvement comparing adaptive versus fixed designs ranges from $(0.132 - 0.125)/0.125 \approx 5\%$ to $(0.151 - 0.125)/0.125 \approx 20\%$, as we consider different δ_k.

Despite not providing a higher expected utility than the optimal fixed design (as shown in Section 4.3), the adaptive design turned out to have advantages over the fixed design in the number of participants assigned to a superior treatment in the phase 2 trial (which is not reflected in the utility function from Section 3.1). That is, although the original aim of the phase 2 adaptive design was to provide more targeted information to assist in the

Table 4: Expected value of f_{prop} and f_{ben}, comparing fixed design versus adaptive design. Top half shows $\mathbb{E}(f_{\text{prop}}|\Delta = \delta_k)$ and bottom half shows $\mathbb{E}(f_{\text{ben}}|\Delta = \delta_k)$.

		Δ	Fixed Design Rule	Adaptive Design Rule
$\mathbb{E}(f_{\text{prop}}\|\Delta = \delta_k)$:	Data Generating Distributions	δ_1	50%	50%
		δ_2	25%	29%
		δ_3	25%	28%
		δ_4	25%	30%
		δ_5	0%	0%
		δ_6	0%	0%

		Δ	Fixed Design Rule	Adaptive Design Rule
$\mathbb{E}(f_{\text{ben}}\|\Delta = \delta_k)$:	Data Generating Distributions	δ_1	0.5	0.5
		δ_2	0.125	0.137
		δ_3	0.125	0.132
		δ_4	0.125	0.151
		δ_5	-0.5	-0.5
		δ_6	0.000	0.000

decision at the end of phase 2, a by-product is that the optimal adaptive design can lead to more participants assigned to a superior treatment in phase 2. From this perspective, such adaptive designs may be more ethical to conduct (under the assumptions built into our simulation study).

5 Simulation Study Mimicking Features from the MISTIE II Trial

We applied our optimized phase 2 designs in simulation studies where the data generating distribution mimics features from the completed MISTIE II trial. The baseline score R is the quantile of intracerebral hemorrhage (ICH) volume. A histogram of the pre-randomization ICH volumes for MISTIE trial participants. We define $M = 4$ baseline score categories demarcated by the 25%, 50% and 75% percentiles of ICH volume, which correspond to 33ml, 43ml, 57ml, respectively. The discrete baseline scores $\tilde{R} = 1, 2, 3, 4$ correspond to the ranges of ICH volume (in ml) $[17, 33)$, $[33, 43)$, $[43, 57)$, $[57, 120]$, respectively. The data set consists of 90 participants with complete observations $(\tilde{R}_i, T_i, Y_i) : i = 1, \ldots, 90$ in the MISTIE trial. The primary outcome is the indicator of whether the participant's functional disability score, as measured by the modified Rankin Scale, is 3 or less.

We consider adaptive phase 2 trial designs that are optimized for the problem setup in Section 4.1 using sample size $n = 64$ per stage, and utility function parameters $c = 0.1$ and each of $\lambda \in \{0.004, 0.01\}$; the resulting optimal designs are denoted $d_{opt}^{(A)}, d_{opt}^{(B)}$ (which differ

for the two different values of λ used). These optimization problems did not incorporate any features of the MISTIE data except that the baseline score is discretized into 4 categories.

We investigated the performance of the above, optimized designs in simulated phase 2 trials where the distribution of the outcome Y given study arm T and baseline score $\tilde{R}$ is the empirical distribution in the MISTIE trial. The estimated treatment effects $\tilde{\Delta}(\tilde{r})$ from this distribution are -0.07 for $\tilde{r} = (0, 0.25)$, 0.22 for $\tilde{r} = (0.25, 0.5)$, 0.12 for $\tilde{r} = (0.5, 0.75)$ and 0.19 for $\tilde{r} = (0.75, 1)$.

Data in Stage A of each simulated phase 2 trial were generated as follows: equal numbers from each baseline category were enrolled and assigned to each arm; for each participant i, a random outcome Y was drawn from the MISTIE data set empirical distribution conditioned on that participant's study arm and baseline score $(T_i, \tilde{R}_i)$. Stage B data were generated analogously, except enrollment was only from the selected population.

The above data generating distribution violates our assumption that the outcome is normally distributed conditional on the study arm and baseline score. Also, the induced treatment effect function is not in the support of the prior π. This provided an opportunity to evaluate the performance of our optimized phase 2 designs in a scenario that differs from those it was optimized for.

Using the above data generating distribution, we simulated 200 phase 2 trials and computed the proportion of trials with each population recommendation for phase 3. The results are summarized in Figures 3 and 4. In each figure, each rectangle's base represents the interval corresponding to a population recommendation in $\mathcal{E}^{(B)}$ for phase 3 enrollment, and its

height represents the observed proportion of that recommendation under the optimal design $d_{opt}^{(B)}$. In Figure 3, for example, the population $\tilde{R} \in (0.5, 1)$ (base) is recommended for phase 3 enrollment in 30% (height) of simulated trials.

Figures 3 and 4, which differ only in the phase 3 duration cost $\lambda = 0.004$ vs. $\lambda = 0.01$, show the impact of increasing this cost. This increase in λ incentivizes recommending a broader population for phase 3 enrollment (to increase the enrollment rate) or recommending no phase 3 trials. Comparing Figures 3 and 4, the overall population is recommended with frequencies 4.5% vs. 8%, and no phase 3 trials are recommended with frequencies 17.5% and 32.5%, respectively.

Since we sampled data from the empirical distribution of the MISTIE II trial as described above, the true state of nature $\tilde{\Delta}(\tilde{r})$ is $(-0.07, 0.22, 0.12, 0.19)$ for baseline ICH quartiles $\tilde{r} = 1, 2, 3, 4$, respectively. The population who benefit from treatment, in terms of baseline ICH quartiles $\tilde{R}$, consists of the last 3 categories, which represent ICH volume quantiles $(0.25, 1]$. Our phase 2 designs only allowed recommending a population to enroll in phase 3 from the action set $\mathcal{E}^{(B)}$ defined in (6). This precluded enrolling the true population who benefit from treatment $(0.25, 1]$. The allowed enrollment choices were the following: $(0, 1)$, which has the disadvantage of including the first quartile who are slightly harmed by treatment; $(0.5, 1)$ or $(0.25, 0.75)$, each of which has the disadvantage of excluding a quartile who benefit; $(0, 0.5)$, which has both types of disadvantages; and the empty set. Each optimized decision rule $d_{opt}^{(B)}$ in Figures 3 and 4 selects one of the most desirable options, i.e., one of $(0.25, 0.75)$,$(0.5, 1)$,$(0, 1)$, with total probability ranging from 60-68.5%.

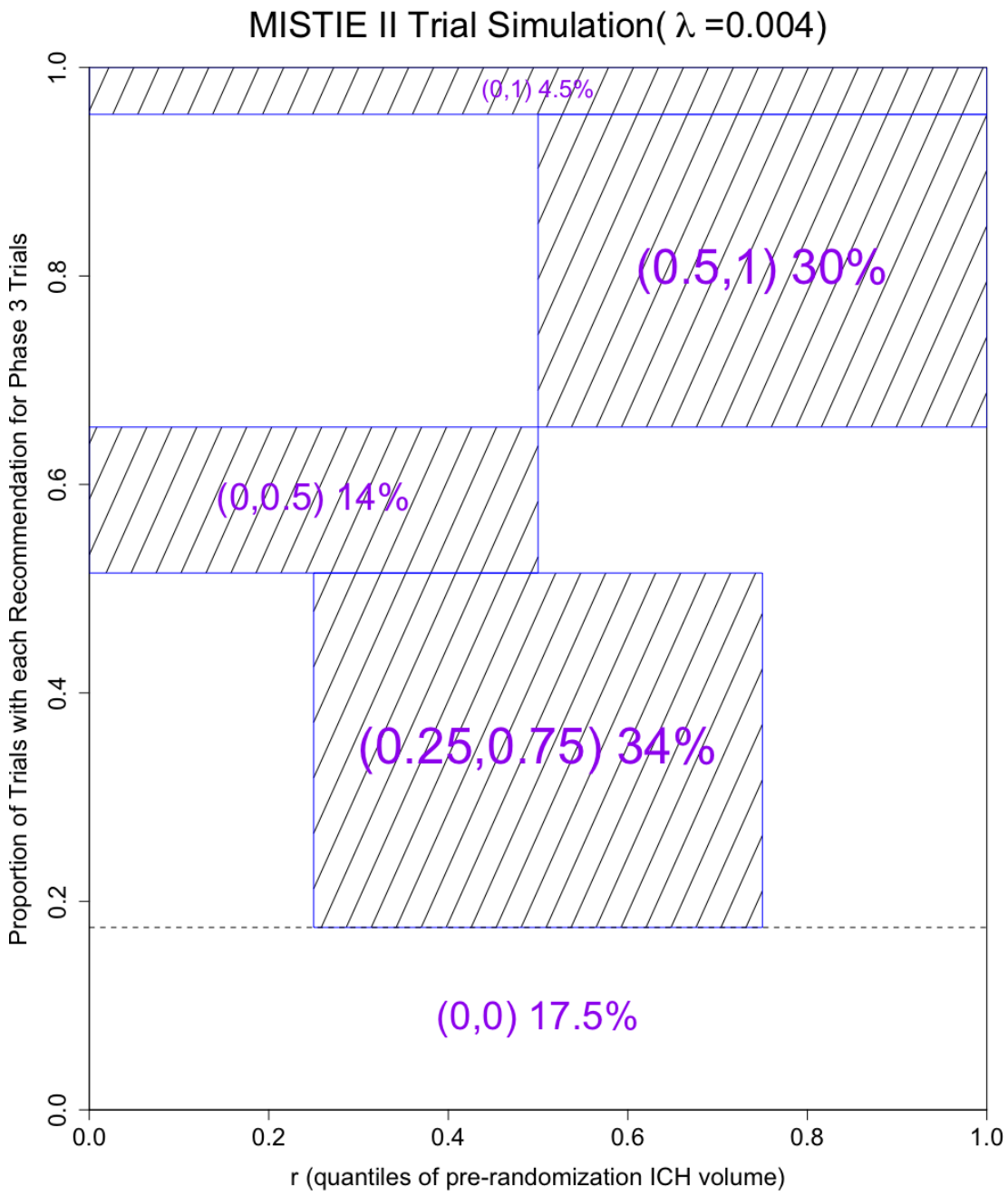


Figure 3: The plot shows the proportion of simulated trials in which the optimized adaptive phase 2 design makes each recommendation in $\mathcal{E}^{(B)}$ for phase 3 trial enrollment, at $\lambda = 0.004$.

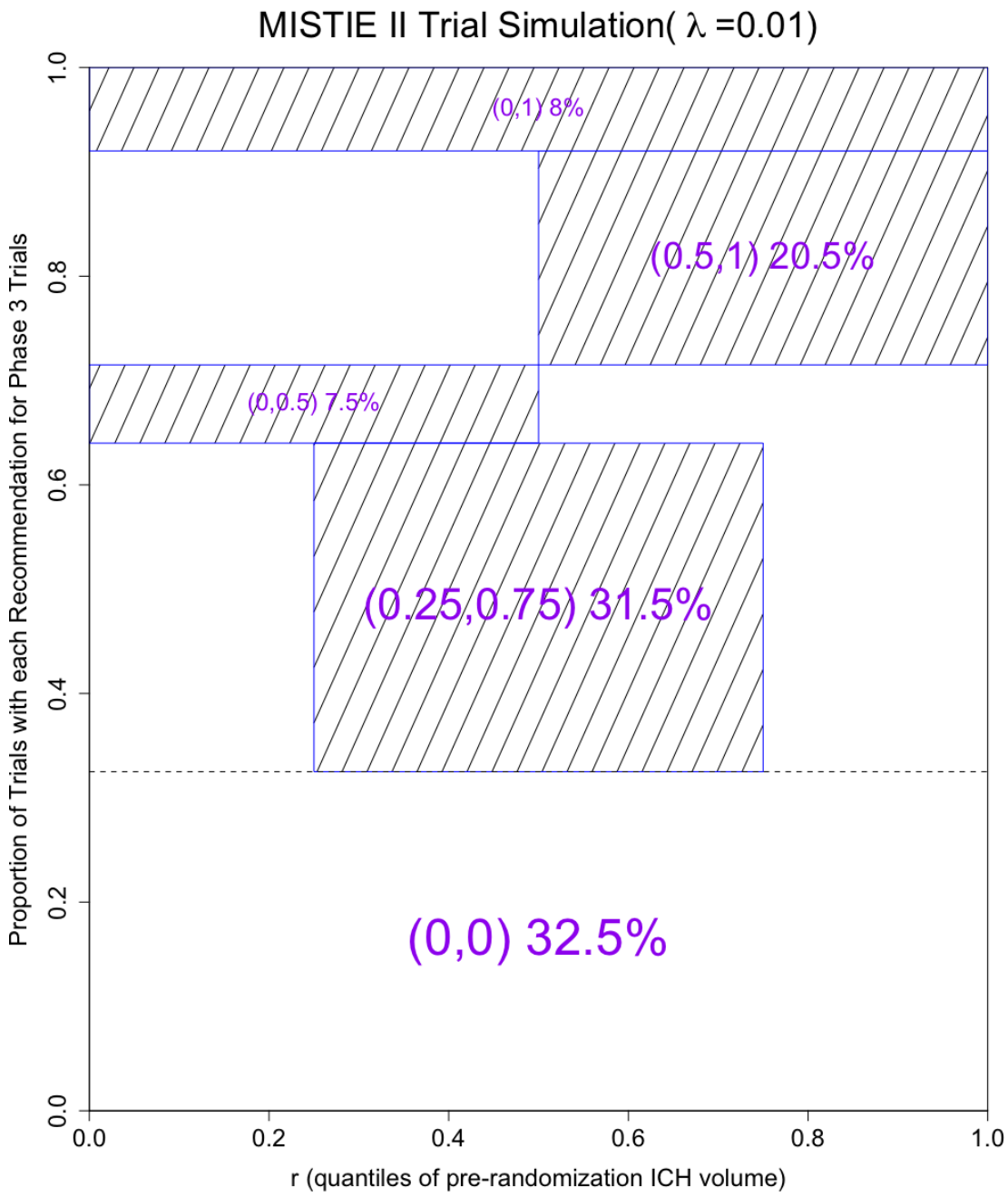


Figure 4: The plot shows the proportion of simulated trials in which the optimized adaptive phase 2 design makes each recommendation in $\mathcal{E}^{(B)}$ for phase 3 trial enrollment, at $\lambda = 0.01$.

6 Discussion

We optimized a phase 2 adaptive enrichment design where the utility function U represents a combination of the cost of conducting phase 3 trials (if they are recommended after phase 2) and the improved health of the future population due to treatment minus the treatment's cost (if the phase 3 trials succeed). Despite not providing a higher expected utility than the optimal fixed design, the adaptive design turned out to have advantages over the fixed design in the number of participants assigned to a superior treatment in the phase 2 trial. A limitation of our designs is that we require the primary outcome to be measured on each participant relatively soon after her/his enrollment, in order to avoid a long pause in enrollment between Stages A and B; this limitation is shared by many adaptive enrichment designs.

In Section 4.1, we discretized the continuous-valued baseline score R into the 4 level categorical variable $\tilde{R}$ by partitioning the support $(0, 1)$ of R into 4 equal length intervals. It may be possible to increase the expected utility of the optimal decision functions if we modify the problem by using a finer discretization of $(0, 1)$. Using a finer discretization, e.g., consecutive intervals of width 1/8, could increase the information available about the unknown value of Δ; this could be used to improve decisions and increase expected utility. A tradeoff is that increasing the fineness of the discretization leads to increased computational complexity. It is an area for future investigation to determine how much added value a finer discretization provides, and at what computational cost.

An area for future research is to incorporate the possibility of early stopping for futility

at the end of Stage A of phase 2, which could potentially save resources. We also could consider sample sizes in Stage B of phase 2 that differ from the sample size in Stage A, or that are adaptively selected based on Stage A data. Similarly, we could consider setting the sample size for the phase 3 trials based on the data from phase 2. Another area for future research is to consider a variety of different utility functions, e.g., incorporating the trial design cost function from Emerson et al. (2011).

The R code for our computations in Section 4 and 5 is available at `https://github.com/duyu8411/Phase2AdaptDesign`.

Acknowledgement

This research was supported by the Patient-Centered Outcomes Research Institute (ME-1306-03198) and the U.S. Food and Drug Administration (HHSF223201400113C). This paper's contents are solely the responsibility of the authors and do not represent the views of these agencies.

References

Anindita Banerjee and Anastasios A. Tsiatis. Adaptive two-stage designs in phase ii clinical trials. *Statistics in Medicine*, 25(19):3382–3395, 2006. ISSN 1097-0258. doi: 10.1002/sim.2501. URL `http://dx.doi.org/10.1002/sim.2501`.

AD Barker, CC Sigman, GJ Kelloff, NM Hylton, DA Berry, and LJ Esserman. I-SPY 2:

An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009.

Anthony E Brockwell and Joseph B Kadane. A gridding method for bayesian sequential decision problems. *Journal of Computational and Graphical Statistics*, 12(3):566–584, 2003.

Tianxi Cai, Lu Tian, Peggy H. Wong, and L. J. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2011. doi: 10.1093/biostatistics/kxq060. URL `http://biostatistics.oxfordjournals.org/content/12/2/270.abstract`.

Bradley P Carlin, Joseph B Kadane, and Alan E Gelfand. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, pages 964–975, 1998.

Yi Cheng and Donald A Berry. Optimal adaptive randomized designs for clinical trials. *Biometrika*, 94(3):673–689, 2007.

Theodore Colton. A model for selecting one of two medical treatments. *Journal of the American Statistical Association*, 58(302):388–400, 1963.

Theodore Colton. A two-stage model for selecting one of two treatments. *Biometrics*, 21(1): 169–180, 1965.

Meichun Ding, Gary L Rosner, and Peter Müller. Bayesian optimal design for phase ii screening trials. *Biometrics*, 64(3):886–894, 2008.

Sarah C Emerson, Kyle D Rudser, and Scott S Emerson. Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in medicine*, 30(11):1199–1217, 2011.

Boris Freidlin, Wenyu Jiang, and Richard Simon. The cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2):691–698, 2010.

Heiko Götte, Margarita Donica, and Giacomo Mordenti. Improving probabilities of correct interim decision in population enrichment designs. *Journal of biopharmaceutical statistics*, 25(5):1020–1038, 2015.

Alexandra C. Graf, Martin Posch, and Franz Koenig. Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal*, 57(1):76–89, 2015. ISSN 1521-4036. doi: 10.1002/bimj.201300257. URL `http://dx.doi.org/10.1002/bimj.201300257`.

Lisa V. Hampson and Christopher Jennison. Optimizing the data combination rule for seamless phase ii/iii clinical trials. *Statistics in Medicine*, 34(1):39–58, 2015. ISSN 1097-0258. doi: 10.1002/sim.6316. URL `http://dx.doi.org/10.1002/sim.6316`.

Daniel F Hanley, Richard E Thompson, John Muschelli, Michael Rosenblum, Nichol McBee, Karen Lane, Amanda J Bistran-Hall, Steven W Mayo, Penelope Keyl, Dheeraj Gandhi, Tim C Morgan, Natalie Ullman, W Andrew Mould, J Ricardo Carhuapoma, Carlos Kase, Wendy Ziai, Carol B Thompson, Gayane Yenokyan, Emily Huang, William C Broaddus, R Scott Graham, E Francois Aldrich, Robert Dodd, Cristanne Wijman, Jean-Louis

Caron, Judy Huang, Paul Camarata, David Mendelow, Barbara Gregson, Scott Janis, Paul Vespa, Neil Martin, Issam Awad, and Mario Zuccarello. Safety and efficacy of minimally invasive surgery plus alteplase in intracerebral haemorrhage evacuation (MISTIE): a randomised, controlled, open-label, phase 2 trial. *The Lancet Neurology*, 15(12):1228–1237, 2016. doi: 10.1016/S1474-4422(16)30234-4. URL `http://dx.doi.org/10.1016/S1474-4422(16)30234-4`.

Wenyu Jiang, Boris Freidlin, and Richard Simon. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*, 99(13):1036–1043, 2007.

Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

Johannes Krisam and Meinhard Kieser. Optimal decision rules for biomarker-based subgroup selection for a targeted therapy in oncology. *International Journal of Molecular Sciences*, 16(5):10354, 2015. ISSN 1422-0067. doi: 10.3390/ijms160510354. URL `http://www.mdpi.com/1422-0067/16/5/10354`.

Tze Leung Lai, Philip W. Lavori, and Olivia Yueh-Wen Liao. Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials*, 39(2):191 – 200, 2014. ISSN 1551-7144. doi: http://dx.doi.org/10.1016/j.cct.2014.09.001. URL `http://www.sciencedirect.com/science/article/pii/S1551714414001311`.

J Jack Lee, Xuemin Gu, and Suyu Liu. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials*, 7(5):584–596, 2010.

Qing Liu, Michael A. Proschan, and Gordon W. Pledger. A unified theory of two-stage adaptive designs. *JASA*, 97(460):1034–1041, 2002.

Shoichi Ohwada and Satoshi Morita. Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharmaceutical Statistics*, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

Michael Rosenblum, Brandon Luber, Richard E. Thompson, and Daniel Hanley. Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine*, 35(21):3776–3791, 2016. ISSN 1097-0258. doi: 10.1002/sim.6957. URL `http://dx.doi.org/10.1002/sim.6957`. sim.6957.

David Rossell, Peter Müller, and Gary L Rosner. Screening designs for drug development. *Biostatistics*, 8(3):595–608, 2006.

Lawrence V. Rubinstein, Edward L. Korn, Boris Freidlin, Sally Hunsberger, S. Percy Ivy, and Malcolm A. Smith. Design issues of randomized phase ii trials and a proposal for phase ii screening trials. *Journal of Clinical Oncology*, 23(28):7199–7206, 2005. doi: 10.1200/JCO.2005.01.149. URL `https://doi.org/10.1200/JCO.2005.01.149`. PMID: 16192604.

Amy V Spencer, Chris Harbron, Adrian Mander, James Wason, and Ian Peers. An adaptive design for updating the threshold value of a continuous biomarker. *Statistics in Medicine*, 2016.

S. J. Wang, H. Hung, and R. T. O'Neill. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, 51:358–374, 2009.

Yanxun Xu, Lorenzo Trippa, Peter Müller, and Yuan Ji. Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences*, pages 1–22, 2014.

Xian Zhou, Suyu Liu, Edward S Kim, Roy S Herbst, and J Jack Lee. Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine. *Clinical Trials*, 5(3):181–193, 2008.

Supplementary Material

Discretized versions of $\delta_1, \ldots, \delta_6$

Table 5 gives the discretized versions of $\delta_1, \ldots, \delta_6$, based on applying (1) to each.

Table 5: Average treatment effect $\tilde{\Delta}(\tilde{r})$ in each stratum $\tilde{r} \in \{1, 2, 3, 4\}$, under each possible $\Delta = \delta_1, \delta_2, \ldots, \delta_6$, as derived from Figure 2. In the row above the horizontal line, each stratum $\tilde{r}$ is followed by the interval of the baseline score that it represents.

	Average Treatment Effect $\tilde{\Delta}(\tilde{r})$ in Each Statum $\tilde{r}$ and Overall ($\mathbb{E}[\tilde{\Delta}(\tilde{R})]$)				
	$\tilde{r} = 1; (0, 0.25)$	$\tilde{r} = 2; (0.25, 0.5)$	$\tilde{r} = 3; (0.5, 0.75)$	$\tilde{r} = 4; (0.75, 1)$	Overall
$\Delta = \delta_1$	1	1	1	1	1
$\Delta = \delta_2$	0	0	0.25	0.75	0.25
$\Delta = \delta_3$	0.75	0.25	0	0	0.25
$\Delta = \delta_4$	0	0.5	0.5	0	0.25
$\Delta = \delta_5$	-1	-1	-1	-1	-1
$\Delta = \delta_6$	0	0	0	0	0

Proof of (5)

We prove (5) from Section 3.3.

$$P(\Delta = \delta_k \mid X) \propto P(X \mid \Delta = \delta_k) P(\Delta = \delta_k) \tag{7}$$

$$\propto P(\widehat{\Delta}^{(B)}, d^{(A)}(X^{(A)}), \widehat{\Delta}^{(A)} \mid \Delta = \delta_k) P(\Delta = \delta_k) \tag{8}$$

$$= P^E(\widehat{\Delta}^{(B)}, \widehat{\Delta}^{(A)} \mid \Delta = \delta_k) P(\Delta = \delta_k) \tag{9}$$

$$= P^E(\widehat{\Delta}^{(B)} \mid \Delta = \delta_k) P^E(\widehat{\Delta}^{(A)} \mid \widehat{\Delta}^{(B)}, \Delta = \delta_k) P(\Delta = \delta_k) \tag{10}$$

$$= P^E(\widehat{\Delta}^{(B)} \mid \Delta = \delta_k) P^E(\widehat{\Delta}^{(A)} \mid \widehat{\Delta}^{(B)}) P(\Delta = \delta_k) \tag{11}$$

$$\propto P^E(\widehat{\Delta}^{(B)} \mid \Delta = \delta_k) P(\Delta = \delta_k), \tag{12}$$

where (7) follows from Bayes' rule; (8) holds since $\tilde{S}^{(B)}$ is a sufficient statistic for Δ; (9) follows from the sequential structure of the data generating process; (10) follows from the definition of conditional probability; (11) follows since $\widehat{\Delta}^{(B)}$ is a sufficient statistic for Δ in the design that always enrolls population E in Stage B; (12) holds because $P^E(\widehat{\Delta}^{(A)} \mid \widehat{\Delta}^{(B)})$ does not depend on Δ, and thus is absorbed into the proportionality constant that depends only on X.

Computation of $d_{opt}^{(A)}$

At the end of Stage A, we have collected Stage A data $X^{(A)}$. The goal at this point in the backward inductive computation is to maximize the expected utility conditioned on $X^{(A)}$ over all possible Stage A enrollment choices $\mathcal{E}^{(A)}$, assuming that $d_{opt}^{(B)}$ will subsequently be used after Stage B; the maximizer is defined as $d_{opt}^{(A)}(X^{(A)})$. It follows from (3) that

$$\begin{aligned} d_{opt}^{(A)}(X^{(A)}) &= \underset{d^{(A)}}{\operatorname{argmax}}\, \mathbb{E}\left[U\left\{ d_{opt}^{(B)}(X); \Delta \right\} \middle| X^{(A)} \right] \\ &= \underset{d^{(A)}}{\operatorname{argmax}}\, \mathbb{E}\left[U\left\{ d_{opt}^{(B)}\left(X^{(A)}, X^{(B)}[d^{(A)}\{X^{(A)}\}]\right); \Delta \right\} \middle| X^{(A)} \right], \end{aligned} \tag{13}$$

where the expectation is with respect to the conditional distribution of $(\Delta, X^{(B)})$ given the Stage A data $X^{(A)}$ (which is induced by the prior π). Analogous to the computation of $d_{opt}^{(B)}$, we use Monte-Carlo simulation to approximate the posterior conditional expectation in (13) via the posterior distribution of $(\Delta, X^{(B)})$ given $X^{(A)}$,

Analogous to the computation of $d_{opt}^{(B)}$, we use Monte-Carlo simulation to approximate the posterior conditional expectation in (13) via the posterior distribution of $(\Delta, X^{(B)})$ given $X^{(A)}$, as described in the Supplementary Material. , where we first draw Δ from the pos-

terior distribution $P\left(\Delta|X^{(A)}\right)$ via (12), and generate the Stage B data, $X^{(B)}$, given Δ and a possible Stage A enrollment choice, $d^{(A)}(X^{(A)})$, according to the data generating distribution specified in Section 2.2. We thus obtain the full data, $X = (X^{(A)}, X^{(B)})$, and a value for the utility function. The average values of the utility function from multiple such draws of data approximates the posterior conditional expectation for this particular Stage A enrollment choice. We consider each possible Stage A enrollment choice in $\mathcal{E}^{(A)}$, and set the optimal Stage A enrollment decision $d^{(A)}_{opt}(X^{(A)})$ to be the maximizer over $\mathcal{E}^{(A)}$ of this posterior conditional expectation.