6-19-2013

TRIAL DESIGNS THAT SIMULTANEOUSLY OPTIMIZE THE POPULATION ENROLLED AND THE TREATMENT ALLOCATION PROBABILITIES

Brandon S. Luber
Johns Hopkins School of Medicine, Department of Oncology, Division of Biostatistics and Bioinformatics, bsluber@jhu.edu

Michael Rosenblum
Johns Hopkins Bloomberg School of Health, Department of Biostatistics

Antoine Chambaz
Modal"X, Universite' Paris Quest

Trial Designs that Simultaneously Optimize the Population Enrolled and the Treatment Allocation Probabilities

Brandon S. Luber [*], Michael Rosenblum[†], and Antoine Chambaz[‡]

June 18, 2013

Abstract

Standard randomized trials may have lower than desired power when the treatment effect is only strong in certain subpopulations. This may occur, for example, in populations with varying disease severities or when subpopulations carry distinct biomarkers and only those who are biomarker positive respond to treatment. To address such situations, we develop a new trial design that combines two types of preplanned rules for updating how the trial is conducted based on data accrued during the trial. The aim is a design with greater overall power and that can better determine subpopulation specific treatment effects, while maintaining strong control of the familywise Type I error rate. The first component of our design involves response-adaptive randomization, in which the probability of being assigned to the treatment or control arm is updated during the trial to target an optimal allocation. The second component of our design involves enrichment, where the criteria for patient enrollment may be modified to help learn which subpopulations benefit from the treatment. We do a simulation study to compare the power of our design, which we call a response-adaptive enrichment design, to three simpler designs: a standard randomized trial design, a response-adaptive design, and an enrichment design. Our simulation study compares these designs in scenarios that arise from the problem of testing the effectiveness of a hypothetical new antidepressant.

[*]Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, 550 North Broadway, Baltimore, MD 21205, USA, bsluber@jhu.edu

[†]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

[‡]Modal'X, Université Paris Ouest Nanterre 200 av. de la République, 92001 Nanterre, France

1 Introduction

In 2006, the Critical Path Opportunities List was released by the U.S. Food and Drug Administration (FDA), outlining 76 projects aimed at improving the success rate of bringing new medical discoveries from the lab to the patient. One set of projects involved improving methods for the use of adaptive designs in clinical trials (Chow and Corey, 2011). The goals of adaptive designs include designing a study that (1) is more efficient, (2) increases the success rate of the study objective, or (3) yields a better understanding of the treatment's results (FDA, 2010). We propose a new adaptive design that combines features of an enrichment design and a response-adaptive design. To the best of our knowledge, this is the first design combining these features. Our goal is to determine how these features interact, in terms of power to detect treatment benefits in different subpopulations.

Enrichment designs allow pre-planned rules for changing the population enrolled, and may be helpful when the results of a treatment may substantially vary in predefined subpopulations. An adaptive randomization, or response-adaptive design, uses the responses of part participants to adjust the probabilities of treatment assignments for future participants. This can potentially increase power through the adjustment of probabilities of treatment assignments in each subpopulation to target an optimal allocation ratio, such as the Neyman allocation (defined below). Response-adaptive designs may be useful when it is believed that a treatment or intervention not only shifts the mean of the outcomes distribution, but also alters the variance of the outcomes. The approach may only be reasonable for a trial with a short duration between enrollment and observation of the primary outcome, since the randomization probability is dependent upon the observed outcomes of the previous participants (Chow and Chang, 2008). To take advantage of the benefits that adaptive designs can potentially offer, rapid data collection and speedy application are required (Gallo *et al.*, 2006).

The motivation for combining enrichment and response-adaptive randomization into a single design is the potential for synergy. This could arise, for example, if response-adaptive randomization generates more relevant information not only for use in the final analysis, but also for use in interim analyses. If better information is available at interim analyses, this may lead to better interim decisions regarding which populations to continue enrolling from and which to stop, i.e., this may improve the enrichment component of the design. Our analysis aims to determine if such synergy is present for a particular response-adaptive enrichment design, and if so to determine how useful such a design could be in practice.

According to (FDA, 2010), both response-adaptive and enrichment designs are "less well understood, pose challenges in avoiding introduction of bias, and generally call for

statistical adjustment to avoid increasing the Type I error rate." In Section 3, with the above considerations in mind, we aim to extend the theoretical framework of a two-stage enrichment design of (Rosenblum and van der Laan, 2011) to incorporate response-adaptive randomization, while maintaining strong control of the familywise Type I error rate, defined as the probability of rejecting at least one true null hypothesis.

In addition to results for our response-adaptive enrichment design, we extend the results of (Rosenblum and van der Laan, 2011) to trial designs with fixed randomization ratios that are not 1:1, and that may even be different across subpopulations. In Section 4 we show a power simulation that compares our response-adaptive enrichment design to the enrichment design from (Rosenblum and van der Laan, 2011), a response-adaptive design, and a standard fixed design. We consider settings where the variance of the outcome under assignment to control is different than that under assignment to treatment, since it is under such scenarios that adapting the randomization probabilities has the most potential to increase power. We also explore how the different designs perform in terms of the average number of patients assigned to the superior versus inferior treatment arm.

2 Related Work

Our motivation for considering response-adaptive enrichment designs is to improve power and the ability to distinguish subpopulation treatment effects, compared to standard designs. We base our design on the enrichment design of (Rosenblum and van der Laan, 2011), and augment it with response-adaptive randomization for each subpopulation. They give a method for constructing randomized trial designs that allow changes to the population enrolled based on interim data using a prespecified decision rule, while maintaing strong control of the asymptotic, familywise Type I error rate at a specified level α. They only consider designs that have rules to potentially change which population is enrolled, and do not modify other design parameters such as the total sample size or the randomization ratios. However, they do conjecture that their general method may be extended to designs such as the one here that also involve adapting the randomization probabilities; this is the open problem we tackle.

According to (Karrison *et al.*, 2003), although there has been substantial statistical literature in the area of response-adaptive designs, there has been little use of them in practice. They point to several reasons why response-adaptive designs are not commonly used, including the logistical challenges of implementing an adaptive assignment scheme, and "the potential for bias due to selection effects, 'drift' in patient characteristics or risk factors over time, and other sources". There is also great debate as to the usefulness of response-

adaptive designs. (Korn and Freidlin, 2011) found only a small benefit in using these designs over standard fixed designs, although the context of their simulations was different from the one here. These important limitations should be kept in mind when considering designs involving response-adaptive randomization. To be useful in practice, response-adaptive designs should offer substantial benefits in order to outweigh these limitations.

Our proposed design allows for preplanned changes to the population enrolled, and is potentially useful when it is thought there may be differing treatment effects across subpopulations. Such differences by subpopulation were observed, for example, in the study from (Gunnarsdottir *et al.*, 2010), whose aim was to investigate TOP2A gene copy number changes as a means to identify groups of breast cancer patients who benefit from anthracycline treatment. In this trial, patients were randomly assigned to receive intravenous CMF (cyclophosphamide, methotrexate and fluorouracil) or CEF (cyclophosphamide, epirubicin and fluorouracil). Subgroup analyses supported that superiority of CEF over CMF may be limited to patients with TOP2A mutations, whose tumors have TOP2A ratios below 0.8 or above 2.0 (Gunnarsdottir *et al.*, 2010). In another example, which we describe below, there is suggestive evidence that the efficacy of certain antidepressants may depend on the initial severity of depression (Kirsch *et al.*, 2008).

3 Response-Adaptive Enrichment Design

3.1 Overview

There is suggestive evidence from the meta-analysis in (Kirsch *et al.*, 2008) that certain antidepressants may only be superior to a control, on average, for the population with severe pre-treatment depression. We present a response-adaptive enrichment design motivated by this work, where we consider planning a future trial of a new hypothetical antidepressant treatment. Using (Kirsch *et al.*, 2008) as a motivation, we will have a Hamilton Rating Scale of Depression (HRSD) score recorded at baseline for each participant, along with his/her corresponding HRSD score after the trial. Define each participant's improvement to be the difference between baseline HRSD score and HRSD score at end of follow-up. A negative value of this difference means the participant got worse.

Define subpopulation 1 to be those with moderate pre-treatment depression at baseline, and subpopulation 2 to be those with severe pre-treatment depression at baseline. Define the total population to be the union of these disjoint subpopulations. We assume the proportions of the total population that correspond to subpopulations 1 and 2 remain unchanged over the trial duration.

Define H_{02} to be the null hypothesis that the mean improvement corresponding to treatment is no more than that corresponding to control in subpopulation 2. In a similar manner, define the null hypothesis H_{00} corresponding to the total population. For each null hypothesis, define the alternative hypothesis to be that the mean improvement under the new anti-depressant drug is greater than under control. We focus on these null hypotheses, since the meta-analysis of (Kirsch *et al.*, 2008) suggests there will be a treatment benefit for the total population, for only subpopulation 2, or for no population. It is an area of future research to additionally consider the null hypothesis for subpopulation 1.

We now introduce our response-adaptive enrichment design, depicted in Figure 1. We first present the overall idea, and then formally define the design in Section 3.2. The response-adaptive enrichment design consists of two main stages, separated by a decision step. Each main stage has two parts. The total number of participants to be enrolled in the first and second stages are pre-specified, and cannot be modified.

Stage 1.1 (first part of stage 1) Participants are drawn from the total population. Each enrolled participant is randomly assigned to either the treatment arm or the control arm with a 50% chance. Using a stratified block randomization, we can ensure that approximately 50% of the participants in subpopulations 1 and 2 are assigned to each arm.

Stage 1.2 (second part of stage 1) Participants are drawn from the total population. Each enrolled participant is randomly assigned either to the treatment arm or to the control arm with differing probabilities, determined by targeting the Neyman allocation of treatment and control for each subpopulation separately.

The purpose of having two parts to stage 1 is to allow enough information to accrue about differences in the outcome variances under treatment versus control in stage 1.1, in order to adequately modify the randomization probabilities in stage 1.2. This also ensures that a minimum number of participants are assigned to each arm.

At the conclusion of stage 1, when all stage 1 data is available for interim analysis, we compute three z-statistics, $T_0^{(1)}, T_1^{(1)}, T_2^{(1)}$, which correspond to the total population, subpopulation 1, and subpopulation 2, respectively. The (1) superscript is a reminder that these statistics are computed at the end of stage 1. Each of these z-statistics represents the standardized difference between the mean change in outcome under treatment and under control. They are defined in Section 3.2.

Decision step We decide to keep enrolling from the total population during stage 2 if

$T_1^{(1)} > T_2^{(1)}$ or $T_1^{(1)} > 0.3$. Otherwise, we decide to enroll only from subpopulation 2 during stage 2.

In words, we continue to enroll from both subpopulations in stage 2 if we see either a greater estimated, standardized treatment effect in subpopulation 1 than in subpopulation 2 (case $T_1^{(1)} > T_2^{(1)}$) or a non-negligible positive signal for subpopulation 1 (case $T_1^{(1)} > 0.3$). If we see neither of these, then we essentially give up on subpopulation 1, and enroll in stage 2 only from subpopulation 2. This only allows for potential enrichment of subpopulation 2. Such a decision rule was used in (Rosenblum and van der Laan, 2011), where the threshold 0.3 was computed to be a value that gives a favorable power tradeoff in the simulation scenarios we consider below.

Stage 2.1 (first part of stage 2) Each enrolled participant is randomly assigned to either the treatment arm or the control arm with a 50% chance.

Stage 2.2 (second part of stage 2) Each enrolled participant is randomly assigned either to the treatment arm or the control arm with differing probabilities, determined, again, by targeting the Neyman allocation of treatment and control for each subpopulation separately.

In both stages 1.2 and 2.2, the targeting is performed to reduce the asymptotic variance of our final estimator (Chambaz and van der Laan, 2011). In estimating the Neyman allocation in stage 2, we ignore all stage 1 data. This ensures that the data generated in stage 2 is conditionally independent from the data generated in stage 1, given the enrollment decision at the end of stage 1. This conditional independence is used in our proof of strong control of the familywise Type I error rate for our design. We explore the impact of sharing information between stages in determining the Neyman allocation in stage 2, in Section 4.4.

Hypothesis Test We compute a final test statistic, T_{final}, as a weighted combination of $T_0^{(1)}$ and the stage 2 statistic $T^{(2)}$, defined below, that incorporates all stage two data. If $T_{\text{final}} > \Phi^{-1}(0.95)$ then we reject H_{0s}, where $s = 0$ if we enrolled from the total population in stage 2 and $s = 2$ otherwise.

3.2 Formal Definition of Response-Adaptive Enrichment Design

We now formally define the statistics that will be used in the above decision rule and testing procedure. For ease of comparability, we use the same notation as in (Rosenblum and van der Laan, 2011).

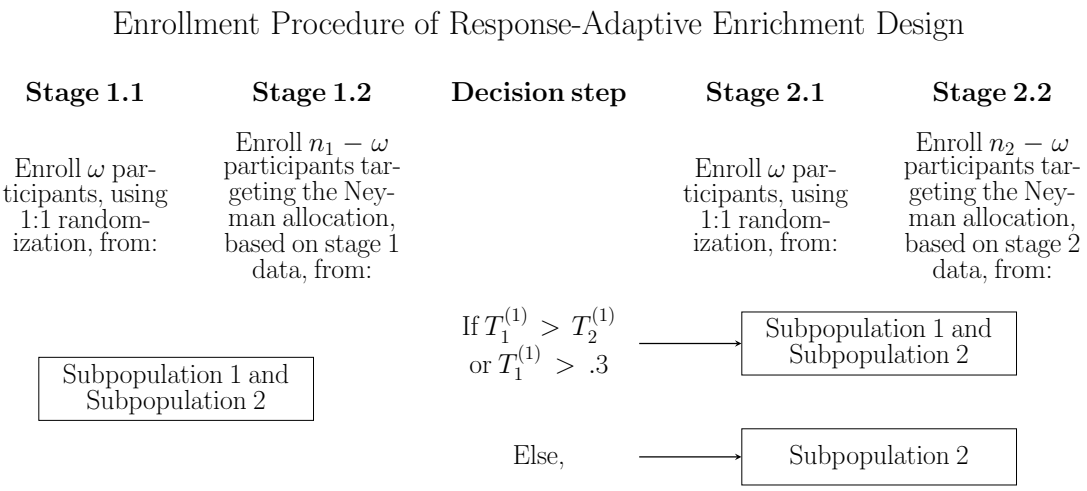

Figure 1: A flow chart of the proposed response-adaptive enrichment design, where enrollment after the first ω patients targets the Neyman allocation for each stage, and where enrollment for stage two is based on our pre-defined decision rule.

Every participant m contributes the data (I_m, S_m, A_m, Y_m) where $I_m \in \{1, 2\}$ indicates in which stage the participant entered, $S_m \in \{1, 2\}$ indicates the participant's subpopulation, $A_m \in \{0, 1\}$ indicates whether the participant was assigned to the treatment arm ($A_m = 1$) or control arm ($A_m = 0$), and $Y_m \in \mathbb{R}$ is the outcome of interest.

For fixed M_1, M_2, τ that do not depend on sample size, define $\mathcal{Q}$ to be the class of distributions Q on $\mathbb{R}$ such that $\sup_{Q \in \mathcal{Q}}\{E_Q[|Y - \mu(Q)|^3/\sigma^2(Q)^{3/2}]\} \leq M_1 < \infty$, and for each $s \in \{1, 2\}, a \in \{0, 1\}$ we have $\tau \leq \sigma^2(Q) \leq M_2$. These properties are used in our proofs of strong control of the familywise Type I error rate, which rely on uniform, multivariate central limit theorems.

Let $\mu(Q)$ and $\sigma^2(Q)$ denote the mean and variance of $Q \in \mathcal{Q}$. For each $s \in \{1, 2\}$, $a \in \{0, 1\}$, let $Q_{sa} \in \mathcal{Q}$ denote the (unknown) distribution of the outcome of interest for subpopulation s under treatment arm assignment a. The probability ϕ_s of assignment to the treatment arm that minimizes the asymptotic variance of the maximum likelihood estimator of the mean treatment effect, the so-called Neyman allocation, is given by $\phi_s = \sigma(Q_{s1})/[\sigma(Q_{s1}) + \sigma(Q_{s0})]$ for each $s \in \{1, 2\}$ (Rosenberger and Hu, 2004).

Let p_1 and p_2 denote the probabilities that a participant uniformly drawn from the total population belongs to subpopulation 1 and subpopulation 2, respectively. We assume p_1 and p_2 are known and fixed throughout the trial. Let n_1 and n_2 denote the pre-specified numbers of participants in stage 1 and stage 2. These are chosen such that $0.05 \leq n_1/(n_1+n_2) \leq 0.95$.

In each stage where participants are enrolled from the total population (i.e., for stage 1,

and possibly for stage 2, depending on the decision made after completion of stage 1), the proportions of enrolled participants from subpopulations 1 and 2 are assumed to equal p_1 and p_2, respectively; more precisely, we assume in this case that the set of participants with $S_m = 1$ enrolled in stage $i \in \{1, 2\}$ is a uniformly drawn random subset of $p_1 n_i$ out of the n_i total sample size in that stage. We assume that for each participant m, conditioned on I_m, $S_m = s$, $A_m = a$, and all the data of previously enrolled participants, the outcome Y_m is a random draw from Q_{sa}.

The two null hypotheses are defined as:

$$H_{02} : \mu(Q_{21}) - \mu(Q_{20}) \leq 0 \quad \text{and} \quad H_{00} : p_1[\mu(Q_{11}) - \mu(Q_{10})] + p_2[\mu(Q_{21}) - \mu(Q_{20})] \leq 0.$$

Stage 1.1 Enroll ω participants from the total population. Each participant is assigned to the treatment arm or to the control arm with a 50% chance.

Stage 1.2 Compute $\hat{\sigma}^2(Q_{11})$, $\hat{\sigma}^2(Q_{10})$, $\hat{\sigma}^2(Q_{21})$, $\hat{\sigma}^2(Q_{20})$, the sample variances in the ω patients enrolled in stage 1.1, which are unbiased estimators of $\sigma^2(Q_{11})$, $\sigma^2(Q_{10})$, $\sigma^2(Q_{21})$, $\sigma^2(Q_{20})$, respectively. This allows us to compute initial estimators $\hat{\phi}_1$ and $\hat{\phi}_2$ of the Neyman allocations ϕ_1 and ϕ_2.

Enroll $n_1 - \omega$ participants from the total population. The estimators $\hat{\sigma}^2(Q_{11})$, $\hat{\sigma}^2(Q_{10})$, $\hat{\sigma}^2(Q_{21})$, $\hat{\sigma}^2(Q_{20})$ are updated after each enrolled participant, with each update yielding updated estimators $\hat{\phi}_1$ and $\hat{\phi}_2$. Participants from subpopulation 1 are randomly assigned to the treatment arm with probability $\hat{\phi}_1$, the current estimator of ϕ_1, while participants from subpopulation 2 are randomly assigned to the treatment arm with probability $\hat{\phi}_2$, the current estimator of ϕ_2.

Decision step Define for each $i \in \{1, 2\}$ and $s \in \{1, 2\}$ the statistics

$$\begin{aligned} se_s^{(i)} &= \left(\frac{\hat{\sigma}^2(Q_{s1})}{\sum_{m:I_m=i,S_m=s} A_m} + \frac{\hat{\sigma}^2(Q_{s0})}{\sum_{m:I_m=i,S_m=s} (1 - A_m)} \right)^{1/2}, \\ T_s^{(i)} &= \left(\frac{\sum_{m:I_m=i,S_m=s} Y_m A_m}{\sum_{m:I_m=i,S_m=s} A_m} - \frac{\sum_{m:I_m=i,S_m=s} Y_m (1 - A_m)}{\sum_{m:I_m=i,S_m=s} (1 - A_m)} \right) / se_s^{(i)}, \end{aligned} \tag{1}$$

$$
\begin{aligned}
se_0^{(i)} &= \left(p_1^2 (se_1^{(i)})^2 + p_2^2 (se_2^{(i)})^2\right)^{1/2}, \\
T_0^{(i)} &= \left(p_1 se_1^{(i)} T_1^{(i)} + p_2 se_2^{(i)} T_2^{(i)}\right) / se_0^{(i)}. \qquad (2)
\end{aligned}
$$

The statistics $se_1^{(1)}$, $se_2^{(1)}$, and $se_0^{(1)}$ should be thought of as estimators of the standard errors of the numerators in the definitions of $T_1^{(1)}$, $T_2^{(1)}$, $T_0^{(1)}$, respectively.

Compute the above statistics at $i = 1$. We decide to keep enrolling from the total population during stage 2 if $T_1^{(1)} > T_2^{(1)}$ or $T_1^{(1)} > 0.3$. Otherwise, we decide to enroll only from subpopulation 2 during stage 2.

Stage 2.1 Enroll ω participants from the selected population. Each participant is assigned to the treatment arm or control arm with a 50% chance.

Stage 2.2 Compute the sample variances $\hat{\sigma}^2(Q_{21})$ and $\hat{\sigma}^2(Q_{20})$ based on the ω patients enrolled in stage 2.1. If the selected population includes subpopulation 1, similarly compute $\hat{\sigma}^2(Q_{11})$ and $\hat{\sigma}^2(Q_{10})$. This allows computing a new initial estimator $\hat{\phi}_2$ of the Neyman allocations ϕ_2, and if the selected population includes subpopulation 1, a new initial estimator $\hat{\phi}_1$ of the Neyman allocations ϕ_1.

Enroll $n_2 - \omega$ participants from the selected population. The estimators $\hat{\sigma}^2(Q_{21})$, $\hat{\sigma}^2(Q_{20})$, and possibly $\hat{\sigma}^2(Q_{11})$, $\hat{\sigma}^2(Q_{10})$ if the selected population includes subpopulation 1, are updated based on stage 2 data only, each update yielding updated estimators $\hat{\phi}_1$ and $\hat{\phi}_2$. Participants from subpopulation 2 are randomly assigned to the treatment arm with probability $\hat{\phi}_2$. If the selected population includes subpopulation 1, then stage 2.2 participants from subpopulation 1 are randomly assigned to the treatment arm with probability $\hat{\phi}_1$.

Hypothesis Test Define $T^{(2)} = T_0^{(2)}$ if the selected population is the total population; otherwise, set $T^{(2)} = T_2^{(2)}$. The final test statistic is the weighted combination of the test statistics from stages 1 and 2 given by

$$
T_{\text{final}} = \left(\frac{n_1}{n_1 + n_2}\right)^{1/2} T_0^{(1)} + \left(\frac{n_2}{n_1 + n_2}\right)^{1/2} T^{(2)}.
$$

If $T_{\text{final}} > \Phi^{-1}(0.95)$ then we reject the null hypothesis corresponding to the selected population enrolled from in stage 2, i.e. either H_{00} or H_{02} depending on whether the

selected population is the total population or subpopulation 2, respectively; otherwise we fail to reject any null hypothesis.

3.3 Control of Familywise Type I Error Rate for Response-Adaptive Enrichment Design

Consider the statistic $T = (T_0^{(1)}, T_1^{(1)}, T_2^{(1)}, T^{(2)})$, whose joint distribution we denote P_n with $n = n_1+n_2$. Define the class of 4-tuples of distributions $\mathcal{Q}^4 = \{(Q_{11}, Q_{10}, Q_{21}, Q_{20})$: each $Q_{sa} \in \mathcal{Q}\}$. As in (Rosenblum and van der Laan, 2011), we define strong control of the (asymptotic) familywise Type I error rate at level α to be:

$$\limsup_{n_1,n_2\to\infty} \sup_{\mathbb{Q}\in\mathcal{Q}^4} P_{\mathbb{Q},n}(\text{At least one true null hypothesis is rejected}) \leq \alpha.$$

Consider the case where the variances Q_{sa} are known. Then the true Neyman allocations ϕ_1, ϕ_2 can be computed exactly. We prove in the Supplementary Material that if the true Neyman allocations are used in place of estimated Neyman allocations throughout the response-adaptive enrichment design, it strongly controls the familywise Type I error rate at level 0.05. We conjecture that this result holds for the more realistic case of estimated variances and estimated Neyman allocations. The plausibility of this conjecture is supported by Theorem 2 from Section 4 of Chambaz and van der Laan (2011), which, roughly speaking, states that in the limit as sample size goes to infinity, the statistics in the response-adaptive design using estimated Neyman allocations behave as if the true Neyman allocations were known and used from the start.

We conducted simulations, described in Section 4.6, comparing the performance of the response-adaptive enrichment design (which uses estimated variances and estimated Neyman allocations) to the counterpart using known variances and true Neyman allocations, at $n = 488$. The resulting power and Type I error were nearly identical in all scenarios we considered. However, it remains an area of future work to prove strong control of the familywise Type I error rate for the response-adaptive enrichment design. The main difficulty in showing this is handling the dependence induced by the response-adaptive component of our design. Martingale arguments such as those in (Rosenberger and Lachin, 2002, Chapter 13) or (Chambaz and van der Laan, 2013) could potentially be used to show this. Instead of investigating this issue, we focus our energy on simulations that investigate the finite sample performance of our design. However, we note that in all our simulations (except those in Section 4.6), variances are estimated rather than assumed known.

4 Comparison of Designs in Terms of Power and Number Assigned to Superior Study Arm

4.1 Definition of Designs

Our simulations are motivated by (Kirsch *et al.*, 2008). We compare our response-adaptive enrichment design to a standard fixed design, a response-adaptive design (with no enrichment), and an enrichment design (with no response-adaptive randomization). All designs have the same total sample size $n = n_1 + n_2 = 488$.

The standard fixed design enrolls from the total population and uses 1:1 randomization to treatment and control throughout the trial. The response-adaptive design uses the response-adaptive procedure as described in Sections 3.1 and 3.2, but does not include the enrichment design component; enrollment is from the total population in both stages. For these designs, the null hypothesis H_{00} is rejected if $T_{\text{final}} > \Phi^{-1}(0.95)$ at the end of the trial. The enrichment design is the same as the response-adaptive enrichment design defined in Sections 3.1 and 3.2, except randomization is 1:1 throughout the trial.

We augment all of the above designs, including the response-adaptive enrichment design, to allow additional testing of H_{02} whenever H_{00} is rejected. This is based on the idea of a fixed sequence testing procedure as in (Maurer *et al.*, 1995). Whenever one of the above designs rejects H_{00}, the following test is carried out: if the z-statistic combining all the subpopulation 2 data from both stages exceeds a certain threshold, reject H_{02}. This threshold is set to be $\Phi^{-1}(0.95)$ for the standard fixed design and the response-adaptive design. It follows from the results of (Maurer *et al.*, 1995) that strong control of familywise Type I error rate is maintained at level 0.05 for the standard fixed design and the response-adaptive design. The threshold is slightly increased to $\Phi^{-1}(0.95) + 0.055$ for the enrichment and response-adaptive enrichment designs. It is shown in the Supplemental Material of (Rosenblum and van der Laan, 2011) that this augmented procedure controls the asymptotic, worst-case familywise Type I error at 0.05 for the enrichment design. We prove the same for the response-adaptive enrichment design in the Supplementary Material, in the case in which the variances are assumed known.

4.2 Definition of Scenarios

We compare the power of our response-adaptive enrichment design to the other designs under six scenarios, numbered 1A, 1B, 1C, 2A, 2B, 2C. Each of the six scenarios is defined exactly as in (Rosenblum and van der Laan, 2011) for ease of comparison. In particular, Q_{11}, Q_{10}, Q_{21} and Q_{20} are Gaussian distributions.

In scenarios 1A, 1B, 1C, we set $p_1 = p_2 = 1/2$ and $n_1 = n_2 = 244$. In scenarios 2A, 2B, 2C, we set $p_1 = 0.75, p_2 = 0.25$ and $n_1 = (0.3)(488) = 146, n_2 = (0.7)(488) = 342$. Each of 'A', 'B', and 'C' corresponds to a different setting of the outcome means $\mu(Q_{sa})$ under treatment and control for each subpopulation. These are defined next, motivated by scenarios observed in (Kirsch *et al.*, 2008).

In the meta-analysis of (Kirsch *et al.*, 2008), the average change in HRSD points, comparing each participant's final score to baseline score, was 7.8 in the placebo arm for those with moderate depression; it was 6.6 HRSD points in the placebo arm for those with severe depression. The point estimate for the treatment effect comparing change in HRSD between treatment and placebo was approximately 0 HRSD points for those with moderate depression (though this was based on a single study), 1.8 points in those with severe initial depression, and 3.0 points for those with very severe initial depression.

For scenarios 1A and 1B, we set the data generating distributions to mimic what was seen in (Kirsch *et al.*, 2008): zero treatment effect for those with moderate pre-treatment depression ($\mu(Q_{10}) = \mu(Q_{11}) = 7.8$) and a positive treatment effect for those with severe pre-treatment depression. We set this positive treatment effect to be 1.8 points in scenario 1A ($\mu(Q_{20}) = 7.8, \mu(Q_{21}) = 9.6$), and 3.0 points in scenario 1B ($\mu(Q_{20}) = 6.6, \mu(Q_{21}) = 9.6$). In scenario 1C, the data generating distributions are set to reflect a 1.8 point, positive treatment effect for both those with moderate pre-treatment depression and those with severe pre-treatment depression ($\mu(Q_{10}) = \mu(Q_{20}) = 7.8, \mu(Q_{11}) = \mu(Q_{21}) = 9.6$). For scenarios 2A, 2B, and 2C, we assume the same values for the means $\mu(Q_{sa})$ as in 1A, 1B, and 1C, respectively.

The results of (Kirsch *et al.*, 2008) include estimates of the total population standard deviation being approximately 8.0 HRSD points, both under treatment and under control. In our data generating distributions, we set the variance under treatment to be the same for each subpopulation; similarly we will set the variance under control (which can differ from that under treatment) to be the same for each subpopulation. We define the ratio of outcome standard deviations under treatment and control in each subpopulation to be $r = \sigma(Q_{11})/\sigma(Q_{10}) = \sigma(Q_{21})/\sigma(Q_{20})$. Below, we set r to various values, and examine the impact on the different designs. At $r = 1$, we have equal standard deviations under treatment and under control; in this case 1:1 randomization is the optimal Neyman allocation, so we do not expect any benefit of response-adaptive randomization. At values of r farther from 1, we expect more benefit from the response-adaptive randomization on power. Our goal is to see how the possible benefits of the response-adaptive component interact with enrichment.

So as to hone in on the effect of the response-adaptive component in our simulations, we

construct our data generating distributions so that the power of the standard fixed design in each scenario is unchanged as we vary the ratio r. To achieve this, it suffices that the non-centrality parameters for the z-statistics defined in Section 3.2 be invariant to r under the standard fixed design, which occurs if $\sigma^2(Q_{s0}) + \sigma^2(Q_{s1})$ is a constant; we set this constant to be 8 HRSD points, so that at $r = 1$ the variances equal those derived from (Kirsch *et al.*, 2008). This leads us to set, for each $r > 0$ and each $s \in \{1, 2\}$, the subpopulation standard deviation $\sigma(Q_{s0}) = 8\sqrt{2/(1+r^2)}$ and $\sigma(Q_{s1}) = r\sigma(Q_{s0})$. Under this definition, changing the value of r affects neither the power of the standard fixed design nor the power of the enrichment design, but does affect the power of the designs involving response-adaptive randomizaiton.

We ran simulations for each $r \in \{1, 1.5, 2, 2.5\}$. Table 1 shows the standard deviations under treatment and control for these values of r. In all six of our scenarios, the total sample size for the trial remains the same, at $n = 488$ participants. This sample size was chosen such so that the power of the standard, fixed design to reject H_{00} in scenarios 1C and 2C is 80%. For the response-adaptive enrichment design, we set $\omega = 50$. For each scenario, 100,000 simulated trials were run under each design. The `R` code used for the simulations is included in the Supplementary Material.

Table 1: Standard Deviation Values for $r \in \{1, 1.5, 2, 2.5\}$

r	$\sigma(Q_{s1})$	$\sigma(Q_{s0})$
1	8	8
1.5	9.414	6.276
2	10.119	5.060
2.5	10.505	4.202

4.3 Summary of Simulation Results

Define the overall power of a design in a given scenario to be the probability of rejecting at least one false null hypothesis. We next summarize the power of different designs under the scenarios defined above; complete details are then given in Section 4.4. All power comparisons are given as absolute differences, and values are rounded to the nearest percent.

Across all the values of r we explored, under each scenario, the response-adaptive enrichment design has at least as great overall power as any of the other three designs. When $r = 1$, across scenarios 1A, 1B, 2A, and 2B, each design with enrichment has 14 - 42% more overall power compared to the corresponding design without enrichment.

Recall that the power of the fixed design and the enrichment design do not change with r. As r is increased from 1 to 2.5, the response-adaptive design gains up to 7% more overall power and the response-adaptive enrichment design gains up to 6% more power. Over all scenarios, the gains in power comparing the response-adaptive design to the fixed design were similar to the power gains comparing the response-adaptive enrichment design to the enrichment design. Though there were small differences in the magnitudes of these gains, with the most prominent being the comparison of $r = 2.5$ to $r = 1$ in scenarios 1A and 2A, there does not appear to be a strong synergistic effect across all scenarios of response-adaptive randomization and enrichment.

In Section 4.5, we examine the impact of enrichment and response-adaptive randomization on the expected number N_{sup} of participants assigned to the superior study arm. Enrichment always improves or leaves unchanged N_{sup} compared to the analogous design without enrichment. In contrast, response-adaptive randomization, which was tailored to maximize power through the Neyman allocation, can increase or decrease N_{sup}, compared to the analogous design using 1:1 randomization. In cases where response-adaptive randomization decreases N_{sup} compared to the fixed design, this was partially mitigated by adding enrichment to the design.

4.4 Detailed Simulation Results: Power Comparison

Figures 2, 3, 4, 5 show, for each scenario, side-by-side bar plots of the proportion of simulated trial in which each null hypothesis is rejected, when $r = 1, 1.5, 2, 2.5$. In each of the six scenarios, the first bar is for the fixed design, the second bar is for the response-adaptive design, the third bar is for the enrichment design, and the fourth bar is for the response-adaptive enrichment design. The height of each bar represents overall power, and this is decomposed into the proportion of simulated trials in which only H_{02} is rejected, only H_{00} is rejected, and both are rejected.

At $r = 1$ (Figure 2), the standard deviations under treatment and control are equal. This implies that the Neyman allocations for subpopulations 1 and 2, ϕ_1 and ϕ_2, both equal 1/2. The response-adaptive design behaves almost identically as the fixed design, and the response-adaptive enrichment design behaves almost identically as the enrichment design. Therefore, at $r = 1$, we only summarize the differences between the enrichment design and the fixed design, which were also given in Section 4 of (Rosenblum and van der Laan, 2011). In scenarios 1A and 1B, the true treatment effect is zero for subpopulation 1 and positive (beneficial) for subpopulation 2. In scenario 1A, the enrichment design has 14% more overall power than the fixed design. In scenario 1B, the enrichment design has 21% more overall

power than the fixed design. In scenario 1C, where the true treatment effect is positive for both subpopulations, all designs have 80% overall power. In scenario 2A, the enrichment design has 23% more overall power than the fixed design. In scenario 2B, the enrichment design has 42% more overall power than the fixed design. In scenario 2C, the overall power is identical to scenario 1C.

At $r = 1.5$ (Figure 3), the standard deviation $\sigma(Q_{sa})$ under assignment to the treatment arm ($a = 1$) is 1.5 times that under assignment to control ($a = 0$). In all scenarios, the absolute difference in overall power between the response-adaptive design and fixed design was 1%. The same holds when comparing the response-adaptive enrichment design and enrichment design.

At $r = 2$ (Figure 4), the standard deviation $\sigma(Q_{sa})$ under assignment to the treatment arm is twice that under assignment to control. In scenarios 1A, 1B, and 1C, the response-adaptive design and the response-adaptive enrichment design have 2-3% more overall power than the fixed design and the enrichment design, respectively. In scenarios 2A and 2B, there is a gain of 1-2% in overall power comparing the response-adaptive design to the fixed design and a gain of 2% comparing the response-adaptive enrichment design to the enrichment design. In scenario 2C, the overall power is similar to scenario 1C.

At $r = 2.5$ (Figure 5), the standard deviation $\sigma(Q_{sa})$ under assignment to the treatment arm is 2.5 times that under assignment to control. In scenario 1A the response-adaptive design has 4% more overall power than the fixed design and the response-adaptive enrichment design has 6% more overall power than the enrichment design. In scenario 1B, the response-adaptive design has 7% more overall power than the fixed design and the response-adaptive enrichment design has 6% more overall power than the enrichment design. In scenario 1C, each design involving response-adaptive randomization has 6% more overall power than the corresponding design without this. In scenario 2A, the response-adaptive design has 2% more overall power than the fixed design and the response-adaptive enrichment design has 5% more overall power than the enrichment design. In scenario 2B, the response-adaptive design has 3% more overall power than the fixed design and the response-adaptive enrichment design has 4% more overall power than the enrichment design. The overall power in scenario 2C is similar scenario 1C.

We explored the impact of using both stage 1 and stage 2 data to estimate the Neyman allocations during stage 2.2. The power and Type I error were nearly identical to the results above (in which only stage 2 data is used in stage 2.2 to estimate Neyman allocations).

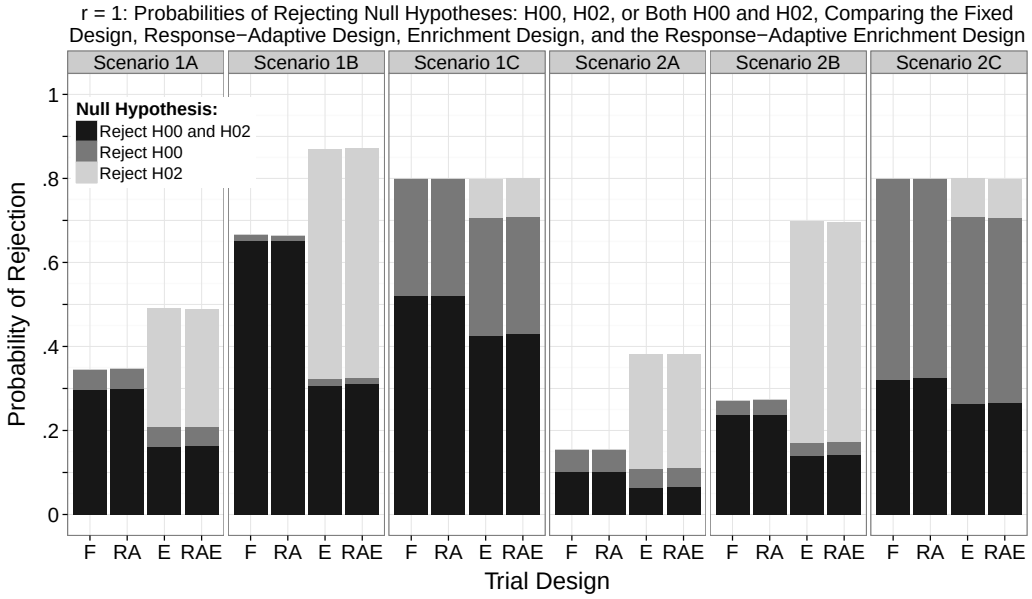

Figure 2: Power simulation with standard deviation ratio $r = 1$. Across the six scenarios, we compare the fixed design, abbreviated F, the response-adaptive design, abbreviated RA, the enrichment design, abbreviated E, and the response-adaptive enrichment design, abbreviated RAE.

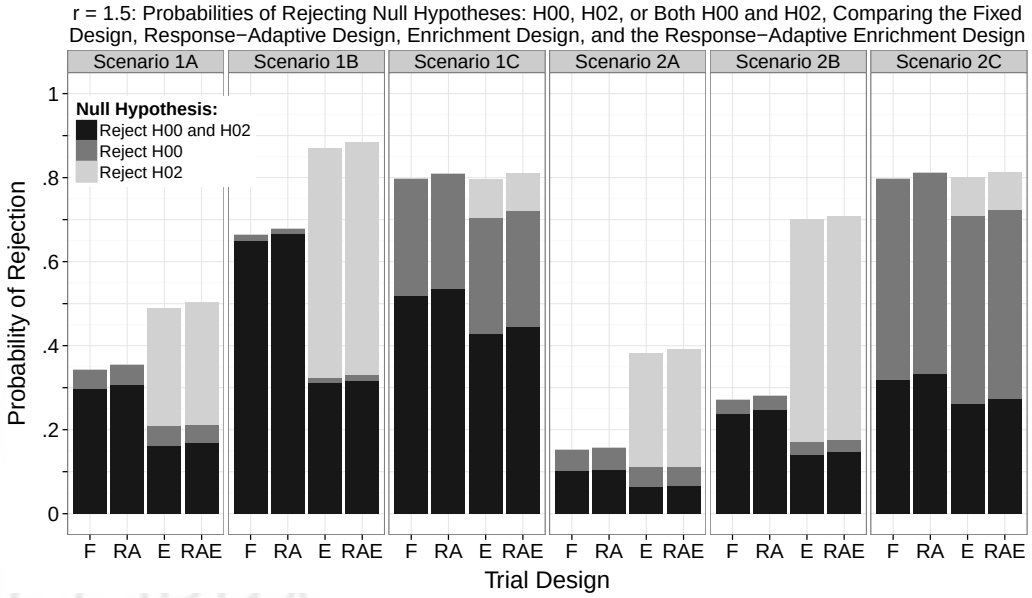

Figure 3: Power simulation with standard deviation ratio, r, set to 1.5.

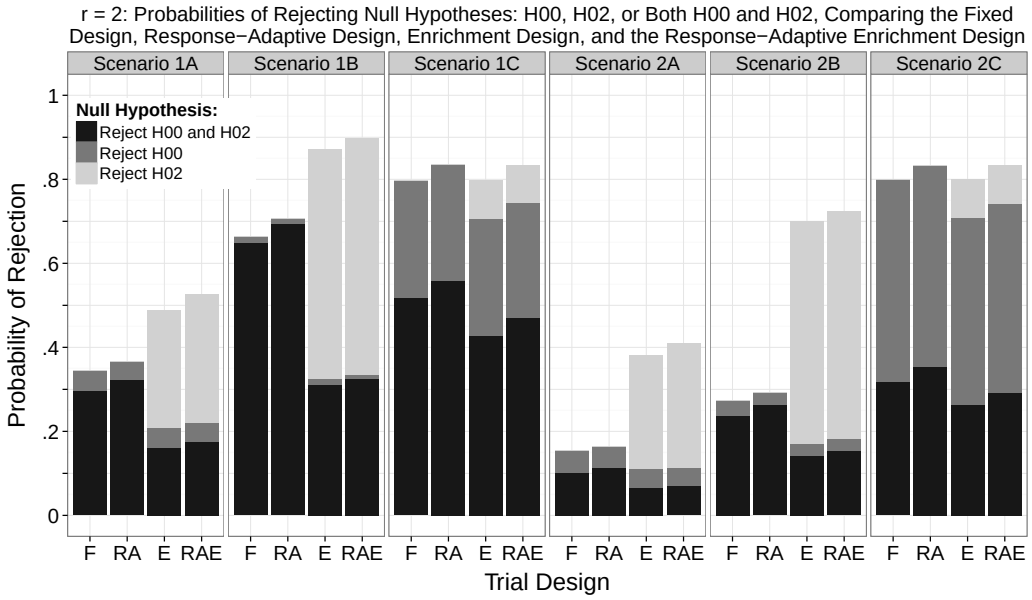

Figure 4: Power simulation with standard deviation ratio, r, set to 2.

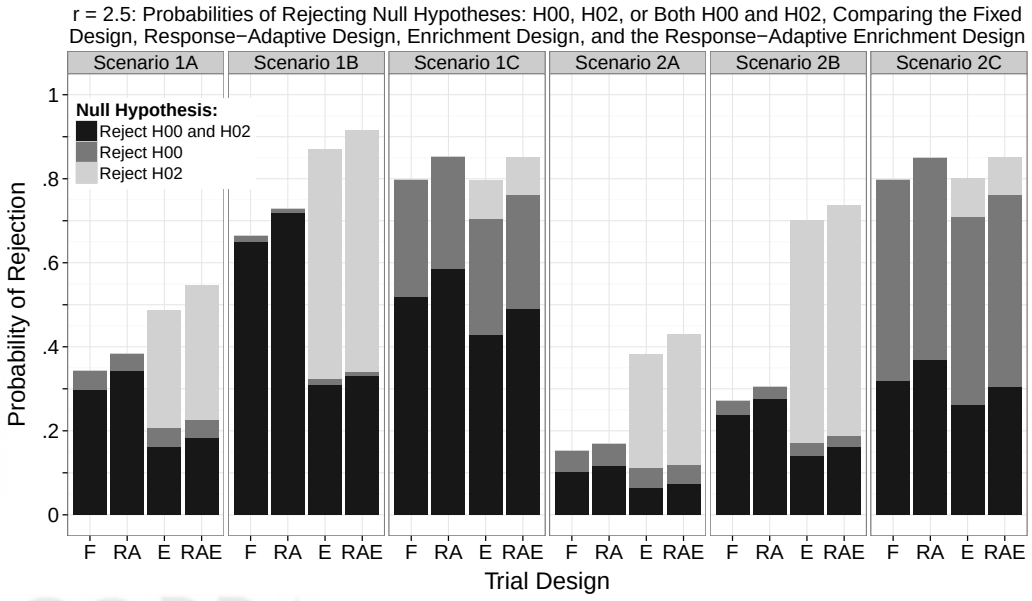

Figure 5: Power simulation with standard deviation ratio, r, set to 2.5.

4.5 Detailed Simulation Results: Assignment to Superior versus Inferior Study Arm

Our response-adaptive enrichment design targets the Neyman allocation, with the goal of maximizing power. Response-adaptive randomization also impacts patient exposure to the inferior versus superior study arm, as described by (Rosenberger and Hu, 2004).

We compare the different designs in each scenario, in terms of the expected number of participants assigned to a superior arm, i.e., an arm for which the mean outcome is strictly greater than under assignment to control for that participant's subpopulation. Because for all scenarios and subpopulations s, we have $\mu(Q_{s1}) \geq \mu(Q_{s0})$, no participant assigned to control is ever assigned to a superior arm. The expected number of participants assigned to a superior arm, which we denote by N_{sup}, is equivalent to the expected number of participants who receive a better treatment than if all participants had been assigned to control.

The top of Table 2 gives N_{sup} for the different designs, for each $r \in \{1, 1.5, 2, 2.5\}$ and each scenario. The value of N_{sup} is substantially greater in the enrichment design compared to the fixed design in scenarios 1A, 1B, 2A, 2B; this is because only one subpopulation benefits in these scenarios, and the enrichment design enrolls more of these participants (who are then assigned to the superior arm with probability 50%) with non-negligible probability. The values of N_{sup} are equal in the enrichment and fixed designs in scenarios 1C and 2C, since both subpopulations benefit from treatment in these scenarios.

The value of N_{sup} is greater in the response-adaptive design compared to the fixed design for all scenarios when $r > 1$, since the response-adaptive design generally assigns more participants to treatment in these cases.

Under scenarios 2A and 2B there is a synergistic effect of enrichment and response-adaptive components, in terms of the difference in N_{sup} due to enrichment and due to response-adaptive randomization. This is most pronounced at $r = 2.5$, where the difference in N_{sup} comparing the response-adaptive enrichment design to the enrichment design is nearly double the difference in N_{sup} comparing the response-adaptive design to the fixed design.

We also considered simulations where for each subpopulation, $\sigma^2(Q_{s1}) < \sigma^2(Q_{s0})$. The bottom of Table 2 compares the different designs, for each $r \in \{1, 1/1.5, 1/2, 1/2.5\}$ and each scenario. Similar results as above hold comparing the enrichment design to the fixed design in terms of N_{sup}. However, adding response-adaptive randomization now decreases the number of participants receiving superior treatment, because the Neyman allocation assigns more participants to control. This decrease is partially mitigated by combining enrichment (which always increases or does not change N_{sup}) with response-adaptive randomization.

Table 2: Exposure to Superior Study Arm Assignment, $r \in \{1, 1.5, 2, 2.5\}$ (top) and $r \in \{1, 1/1.5, 1/2, 1/2.5\}$ (bottom)

	r = 1						r = 1.5					
	1a	1b	1c	2a	2b	2c	1a	1b	1c	2a	2b	2c
Fixed	122	122	244	61	61	244	122	122	244	61	61	244
Response-Adaptive	123	123	244	61	61	244	145	145	288	72	72	288
Enrichment	158	159	244	129	135	244	158	159	244	129	134	244
Response-Adaptive Enrichment	158	160	244	129	134	244	184	185	283	151	157	283

	r = 2						r = 2.5					
	1a	1b	1c	2a	2b	2c	1a	1b	1c	2a	2b	2c
Fixed	122	122	244	61	61	244	122	122	244	61	61	244
Response-Adaptive	159	159	317	80	80	317	170	170	338	85	85	338
Enrichment	157	159	244	129	134	244	157	159	244	129	135	244
Response-Adaptive Enrichment	200	203	309	165	172	309	213	215	328	176	183	327

	r = 1						r = 1/1.5					
	1a	1b	1c	2a	2b	2c	1a	1b	1c	2a	2b	2c
Fixed	122	122	244	61	61	244	122	122	244	61	61	244
Response-Adaptive	123	123	244	61	61	244	100	100	200	50	50	200
Enrichment	158	159	244	129	135	244	158	159	244	129	134	244
Response-Adaptive Enrichment	158	160	244	129	134	244	133	134	205	107	112	205

	r = 1/2						r = 1/2.5					
	1a	1b	1c	2a	2b	2c	1a	1b	1c	2a	2b	2c
Fixed	122	122	244	61	61	244	122	122	244	61	61	244
Response-Adaptive	86	86	171	42	42	171	75	75	150	37	37	150
Enrichment	157	159	244	129	135	244	158	159	244	129	135	244
Response-Adaptive Enrichment	116	118	179	93	97	179	105	106	161	83	87	160

The table provides the expected number of patients (rounded to the nearest integer) assigned to a superior study arm out of 488 total patients.

4.6 Response-Adaptive Enrichment Design using True Neyman Allocations

We consider the case where the true Neyman allocations are used in the response-adaptive enrichment design. We conducted 100,000 simulations at n = 488 and compared the performance to the response-adaptive enrichment design with estimated variances. Across all scenarios, 1A-2C, and for each value $r \in \{1, 1.5, 2, 2.5\}$, the overall power was nearly identical between the two designs. The largest difference in overall power between the two designs was 0.5%, under scenario 2B, at $r = 2$.

5 Type I Error Analysis

We ran simulations where we computed the familywise Type I error rate for the four designs involved in the power analysis of Section 4, using 500,000 simulations per design per scenario. We considered sample sizes $n \in \{244, 488\}$.

We first used Gaussian distributions Q_{11}, Q_{10}, Q_{21}, Q_{20}, all with zero mean. Across all six scenarios above, 1A, 1B, 1C, 2A, 2B, 2C, the largest familywise Type I error for the fixed design was 0.053, for the response-adaptive design was 0.052, for the enrichment design was 0.053, and for the response-adaptive enrichment design was 0.053.

We also considered heavy tailed and skewed data generating distributions. To investigate the case of heavy tailed distributions, we set Q_{11}, Q_{10}, Q_{21}, Q_{20} to be identical, centered log-normal distributions, which is the distribution of $\exp(tZ) - \exp(t^2/2)$, where Z is standard normal; we considered each $t \in \{0.01, 0.1, 1, 2, 4\}$. To investigate the case of skewed distributions, we set Q_{11}, Q_{10}, Q_{21}, Q_{20} to be identical, centered negative binomial distributions with added Gaussian noise, i.e., the distribution of $Y - E(Y) + 0.01Z$, where Z is standard normal and Y is a negative binomial distribution with parameters $c \in \{0.01, 0.1, 1, 2, 4\}$ and $p = c/(c + 1)$.

Under the log-normal distributions, the familywise Type I error for the response-adaptive enrichment design was always less than for the standard fixed design, except under a few scenarios that the fixed design had type I error of zero. When $t \in \{0.01, 0.1\}$, for all $r \in \{1, 1.5, 2, 2.5\}$, the standard fixed design had type I error of zero, and under these scenarios, the largest familywise Type I error for the response-adaptive enrichment design was 0.054. Under the negative binomial simulations, the familywise Type I error for the response-adaptive enrichment design was always less than for the standard fixed design.

6 Discussion

In our simulated scenarios, relatively large differences in the variances under treatment and control are needed before a substantial power improvement (e.g. more than 5%) occurs for the response-adaptive design and response-adaptive enrichment design, compared to the standard fixed design and enrichment-only design, respectively. Due to this negative finding, the situations in which such designs using response-adaptive randomization will have a large impact on power may be limited. However, it was not known before our simulation study whether there would be a synergistic effect on power of combining response-adaptive randomization and enrichment; a contribution of our work has been to explore this possibility.

We did observe a synergistic relationship between enrichment and response-adaptive randomization, not in power, but in terms of the number of participants assigned to the superior study arm.

An open research question is to explore alternative adaptive randomization allocations and determine after what proportion of enrolled participants one should schedule an interim analysis, to optimize power and N_{sup}. A related issue is to determine how the ratios ω/n_1 and ω/n_2 affect the overall power estimates.

(Berry, 2010) argues that the advantage of response-adaptive designs is greater for comparisons of more than two treatment arms. It is an area of future work to consider response-adaptive enrichment designs for more than two treatment arms.

Acknowledgements

This research and analysis was supported by contract number HHSF2232010000072C, entitled, "Partnership in Applied Comparative Effectiveness Science," sponsored by the Food and Drug Administration, Department of Health and Human Services. This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the above agencies.

References

Berry, D. (2010). Adaptive clinical trials: The promise and the caution. *Journal of Clinical Oncology* **31**, 1423–1426.

Chambaz, A. and van der Laan, M. (2011). Estimation and testing in targeted group sequential covariate-adjusted randomized clinical trials. *U.C. Berkeley Division of Biostatistics Working Paper Series* **278**.

Chambaz, A. and van der Laan, M. (2011). Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate: Simulation study. *The International Journal of Biostatistics* **7**, 1–30.

Chambaz, A. and van der Laan, M. (2013). Inference in targeted group sequential covariate-adjusted randomized clinical trials. *Scandinavian Journal of statistics* DOI: 10.1111/sjos.12013.

Chow, S. and Chang, M. (2008). Adaptive design methods in clinical trials. *The Orphanet Journal of Rare Diseases* **3**, 1–13.

Chow, S. and Corey, R. (2011). Benefits, challenges and obstacles of adaptive clinical trial designs. *Orphanet Journal of Rare Diseases* **6**, 1–10.

Chow S., Chang, M., and Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics* **15**, 575–591.

FDA (2010). Draft guidance for industry - Adaptive design clinical trials for drugs and biologics. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development - an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics* **16**, 275–283.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability* **19**, 724–739.

Gunnarsdottir, K., Jensen, M., Zahrieh, D., Gelber, R., Knoop, A., Bonetti, M., Mouridsen, H., and Ejlertsen, B. (2010). CEF is superior to CMF for tumours with TOP2A aberrations: A subpopulation treatment effect pattern plot (STEPP) analysis on danish breast cancer cooperative group study 89d. *Breast Cancer Research and Treatment* **123**, 163–169.

Hochberg, Y. (1998). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

Karrison, T., Huo, D., and Chappell, R. (2003). A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trials* **24**, 506–522.

Kirsch, I., Deacon, B., Huedo-Medina, T., Scoboria, A., Moore, T., and Johnson, B. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* **5**, 260–266.

Korn, E. and Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology* **29**, 771–776.

Luber, B. (2012). Response-adaptive enrichment design: A simulation study. *Master's Thesis*. The Johns Hopkins University.

Mahajan, R. and Gupta, K. (2010). Adaptive design clinical trials: Methodology, challenges and prospect. *Indian J Pharmacol* **42**, 201–207.

Maurer, W., Hothorn, L. A., and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. Fischer Verlag, Stuttgart, 1995.

Rosenberger, W. and Hu, F. (2004). Maximizing power and minimizing treatment failures in clinical trials. *Clinical Trials* **1**, 141–147.

Rosenberger, W. and Lachin, J. (2002). Randomization in clinical trials: Theory and practice. Wiley-Interscience, 2002.

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

Rosenblum, M. and van der Laan, M. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* **98**, 845–860.