University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 267

Optimizing Randomized Trial Designs to Distinguish which Subpopulations Benefit from Treatment

Michael Rosenblum* Mark J. van der Laan†

*Johns Hopkins University, mrosenbl@jhsph.edu

†University of California - Berkeley, laan@berkeley.edu

Optimizing Randomized Trial Designs to Distinguish which Subpopulations Benefit from Treatment

Michael Rosenblum and Mark J. van der Laan

Abstract

It is a challenge to evaluate experimental treatments where it is suspected that the treatment effect may only be strong for certain subpopulations, such as those having a high initial severity of disease, or those having a particular gene variant. Standard randomized controlled trials can have low power in such situations. They also are not optimized to distinguish which subpopulations benefit from a treatment. With the goal of overcoming these limitations, we consider randomized trial designs in which the criteria for patient enrollment may be changed, in a preplanned manner, based on interim analyses. Since such designs allow data-dependent changes to the population sampled, care must be taken to ensure strong control of the familywise Type I error rate.

Our main contribution is a general method for constructing randomized trial designs that (1) allow changes (based on a prespecified decision rule) to the population enrolled based on interim data, (2) make no parametric model assumptions, and (3) guarantee the asymptotic, familywise Type I error rate is strongly controlled at a specified level. As a demonstration of our method, we prove new, sharp results for a simple, two-stage enrichment design. We then compare this design to a fixed design, focusing on each design's ability to determine overall and subpopulation specific treatment effects.

1 Introduction

Randomized trial designs in which participants are enrolled sequentially over time, and where interim analyses are done at prespecified points during the trial, are called "group sequential designs." We consider settings where follow up time is relatively short. At interim analyses, it is then possible to look at outcome data from the patients who have already completed the trial, and use this to optimize aspects of the trial design.

We focus on designs that allow preplanned changes to the population enrolled, based on interim analyses. Such designs, called "enrichment designs," may be useful when it is thought that the effect of a treatment may differ in certain subpopulations. For example, these could be subpopulations defined by a genomic biomarker measured at baseline such as HER2 in studies of breast cancer therapies (Baselga, 2001; Wang et al., 2007, 2009), or defined by baseline factors such as initial severity of depression in studies of antidepressants (Kirsch et al., 2008). In both examples, there is evidence that certain subpopulations (those with higher levels of HER2 overexpression, and those with high baseline severity of depression, respectively) may benefit more from specific interventions. As we show below, designs that allow preplanned changes to which subpopulations are enrolled can lead to more power, and can reveal more information about subpopulation specific treatment effects.

As a concrete example, consider designing a study to test the effectiveness of a hypothetical new antidepressant. Kirsch et al. (2008) present suggestive, though not conclusive, evidence that certain commonly prescribed antidepressants may only be effective for those with severe (rather than moderate) depression at study baseline. With this in mind, it may be advantageous to incorporate the following preplanned interim analysis in one's trial of the new antidepressant: if it is observed that the treatment effect is sufficiently low among those with moderate baseline depression, one will enroll only (or more intensely) those with severe baseline depression for the rest of the trial. We show in Section 4 that certain designs of this type can give improved power, compared to a standard fixed design, at certain alternatives of interest. Due to the multiple hypotheses potentially tested in an enrichment design, a major methodological challenge for such designs is guaranteeing strong control of the familywise Type I error rate.

Our main contribution is a general method for computing the asymptotic, worst-case, familywise Type I error rate for a wide range of enrichment designs. This is of interest, in particular, to the U.S. Food and Drug Administration, who state in a recent draft guidance (FDA, 2010) that for trials allowing certain interim modifications such as those discussed in this paper, "The chief concerns with these designs are control of the study-wide Type I error rate, minimization of the impact of any adaptation-associated statistical ... or operational bias on the estimates of treatment effects, and the interpretability of trial results." We focus throughout this paper on strong control of the familywise Type I error rate, as defined by (Hochberg & Tamhane, 1987, pp. 3, 7), and only consider designs that guarantee clear interpretability of trial results. We briefly discuss the important topics of estimation and confidence intervals in Section 6.

The familywise Type I error rate is the probability that at least one true null hypothesis is rejected. Throughout, we slightly abbreviate "familywise Type I error rate" by "familywise Type I error." The *worst-case*, familywise Type I error is the supremum over all possible data generating distributions of the familywise Type I error. The *asymptotic*, worst-case, familywise Type I error is the limit of this supremum as sample sizes in all stages of a given design go to infinity, which we formally define in Section 5.4; we also define strong control of asymptotic, familywise Type I error in Section 5.4.

Even in some relatively simple enrichment designs, such as those we present in Section 3, it is difficult or impossible to analytically determine the asymptotic, worst-case, familywise Type I error. To address this, we present a general theorem that reduces this problem for a given design to a more convenient optimization problem, that often can be exactly solved by standard software. We apply this theorem to the class of enrichment designs in Section 3, leading to new (to the best of our knowledge) results on familywise Type I error control for these designs.

Throughout this paper, we only consider designs with preplanned rules for changing the population enrolled, based on interim data. We do not consider designs that adapt the total sample size, the number of treatment arms, or the randomization probabilities, for example. However, it is possible to apply our general method to designs that involve such adaptations, and this is an area for future research.

The outline of the paper is as follows. We present related work in Section 2. Then, in Section 3, we give a class of two-stage enrichment designs that will be used as an illustration throughout the paper. In Section 4, we present a simulated application of a particular enrichment design from Section 3. The simulation is based on a meta-study of certain antidepressants (Kirsch et al., 2008), and the goal is to compare the performance of the enrichment design to a fixed design, in determining overall and subpopulation specific treatment effects. Next, in Section 5, we give the details of our general method and present our main theorem; we then apply it to prove new, sharp results for the class of enrichment designs from Section 3. We discuss limitations of our approach and open problems in Section 6.

2 Related Work

Methods have been proposed to deal with the multiple testing problem that arises from design adaptations in which the study population, study treatment, or study endpoint may be changed, e.g. (Thall et al., 1988; Schaid et al., 1990; Bauer & Köhne, 1994; Follmann et al., 1994; Follmann, 1997; Kieser et al., 1999; Hommel, 2001; Stallard & Todd, 2003; Sampson & Sill, 2005; Bischoff & Miller, 2005; Jennison & Turnbull, 2006, 2007; Wang et al., 2007, 2009). In many cases these methods lead to sharp bounds on familywise Type I error. However, for the class of designs we give in Section 3 these methods either do not apply or our method gives improved results.

Below, we describe the following related work that is especially relevant to the class of designs we present in Section 3: (Thall et al., 1988; Bauer & Köhne, 1994; Follmann,

1997; Kieser et al., 1999; Wang et al., 2007, 2009). Each of these papers incorporates a preplanned interim analysis that can result in a change in the data generating distribution for the second stage, and a method for ensuring familywise Type I error is controlled.

Thall et al. (1988) consider the problem of estimating the effect of the best treatment among a set of k prespecified treatments. They use a two stage design that drops all but the single best performing treatment after an interim analysis. Their method can be modified to apply in the setting of this paper, where rather than selecting which treatment to continue in stage two, we select which population to continue enrolling in stage two. This modification of the Thall et al. (1988) procedure to our setting, however, differs from the new class of adaptive designs we present in Section 3, in two ways. The design of Thall et al. (1988) does not allow the option of continuing to sample from both populations in stage two, and it uses a different final test statistic than we do.

Bauer & Köhne (1994) and Kieser et al. (1999) present a general multiple testing procedure that can be applied to designs where the subpopulations enrolled can be changed based on interim analyses. This procedure is an important advance, and is based on the closed testing principle (Marcus et al., 1976), as well as combining p-values from each stage using a prespecified combination rule. Application of their method to our class of designs in Section 3, however, gives a multiple testing procedure with strictly lower power than the one resulting from application of our general method from Section 5.

Follmann (1997) proves control of Type I error for a large class of useful enrichment designs. This class of designs allows a much wider range of decision rules than the ones we consider in Section 3. However, in (Follmann, 1997), the only null hypothesis tested is the global null hypothesis that neither subpopulation benefits from treatment. In contrast, we test null hypotheses corresponding to the overall population as well as specific subpopulations. The advantage of testing more than the global null hypothesis is that rejecting it does not permit one to draw conclusions about treatment effects in a specific subpopulation or in the total population. This is because the complement of the global null hypothesis is that there is a positive effect in at least one subpopulation; without further tests or assumptions, one cannot directly conclude which subpopulation this is, nor determine whether there is a net positive effect in the total population.

Wang et al. (2007, 2009) propose important enrichment designs using decision rules similar to those we give in Section 3. However, in (Wang et al., 2007), the focus is on testing the null hypothesis associated with a single subpopulation; their designs have more power than ours for detecting an effect in this single subpopulation, but less power than our method for testing the null hypothesis for the total population. The setup in Wang et al. (2009) differs from ours in that they assume if there is no net positive effect in the overall population, then there is no net positive effect in a certain, prespecified subpopulation; we do not make that assumption here.

3 A Class of Enrichment Designs

3.1 Overview

We present a class of two stage, enrichment designs tailored to the problem described in Section 1, of testing the effectiveness of a hypothetical new antidepressant. The aim of these designs is to improve power and better determine subpopulation specific treatment effects. We show how our general method gives the first (to the best of our knowledge) means of precisely computing the asymptotic, worst-case, familywise Type I error of these enrichment designs. This can be used to determine the subclass of such designs that strongly control familywise Type I error at a desired level (e.g. 0.05); we can then compare these designs to standard fixed designs in terms of overall power and ability to determine subpopulation specific treatment effects. Later, in Section 4, we present a simple enrichment design from our class that improves on a standard fixed design at certain alternatives of interest.

As described above, Kirsch et al. (2008) present suggestive, but not conclusive, evidence from a meta-analysis that a class of commonly used antidepressants may not be superior to placebo for those with moderate pretreatment depression (though this is based on only one study), while they are superior to placebo for those with severe pretreatment depression (based on 34 studies). We consider a method for designing a trial of a hypothetical new antidepressant with this in mind.

For each subject, the Hamilton Rating Scale of Depression (HRSD) score is recorded at baseline and after a set period of time. Lower HRSD scores indicate lower severity of depression. For each subject, we define his/her improvement in HRSD score to be baseline score minus end of study score. (If a subject's HRSD score increases by the end of study, his/her improvement has a negative value.)

We refer to the set of subjects with moderate depression at baseline as "subpopulation 1" and refer to the set of subjects with severe depression at baseline as "subpopulation 2." We refer to the union of subpopulations 1 and 2 (i.e. those with either moderate or severe pretreatment depression) as the "total population." Let H_{01} denote the null hypothesis that for subpopulation 1, the mean improvement in HRSD score under the new treatment is less than or equal to the mean improvement in HRSD score under the placebo. In an analogous manner, define the null hypothesis H_{02} corresponding to subpopulation 2, and the null hypothesis H_{03} corresponding to the total population. The alternative hypotheses are that the mean HRSD score improvement under treatment is greater than under placebo, for subpopulation 1, for subpopulation 2, and for the total population, respectively. Though our designs are motivated by the above application, they can be useful in many situations where the population enrolled consists of two non-overlapping subpopulations of interest (e.g., defined in terms of a genetic marker, or a risk score at baseline).

Each participant enrolled is randomly assigned to the study arm or the control arm with 50% chance of being assigned to each. For simplicity, we assume exactly 50% of participants in each subpopulation are assigned to each arm, which can be approximately

ensured using stratified block randomization. At the end of stage one, we compute three z-statistics, $T_1^{(1)}, T_2^{(1)}, T_3^{(1)}$ (defined formally below), corresponding to the standardized difference in mean HRSD improvement between treatment and control groups for subjects in subpopulation 1, subpopulation 2, and the total population, respectively. We assume that data on all subjects from stage one are available at the interim analysis.

We allow a decision to be made at the end of stage one, based on a prespecified decision rule. We denote this decision rule, which maps the data collected in stage one to a choice for how many subjects from each subpopulation to sample in stage two, by D. We consider a class of decision rules, which we denote by $\mathbf{D}$, that are Borel measurable maps from the above stage one statistics to the following two choices for the subpopulation(s) to enroll in stage two:

(i) Continue enrolling from both subpopulations in the same proportions as in stage one, or

(ii) Enroll only from the subpopulation $s \in \{1, 2\}$ corresponding to the larger of the stage one z-statistics $T_1^{(1)}$, $T_2^{(1)}$. (Any ties could be decided arbitrarily based on a prespecified rule.)

One example of a decision rule in the class $\mathbf{D}$ is to always choose option (ii), that is, to always sample in stage two only from the subpopulation with larger z-statistic from stage one. Another example of a decision rule in the class $\mathbf{D}$ is to use option (i) if the z-statistic $T_3^{(1)}$ for the total population is above a prespecified threshold, and to use option (ii) otherwise. Figure 1 gives a flow chart representing this type of decision rule. The standard fixed design, which corresponds to always choosing option (i), is also a decision rule in the class $\mathbf{D}$.

At the end of the trial, a preplanned final test statistic is computed (defined below), leading to possible rejection of one of the null hypotheses $\{H_{01}, H_{02}, H_{03}\}$. This final test statistic is a weighted combination of the z-statistic $T_3^{(1)}$ from the first stage (using all first stage data from both subpopulations) and the z-statistic from the second stage (using all second stage data). If the final test statistic exceeds a threshold c, we reject the null hypothesis corresponding to the subpopulation (or the total population) selected for enrollment in stage two.

3.2 Formal definition of statistics, assumptions, and testing procedure

We now precisely define the statistics used in the above decision rules and to test the family of null hypotheses $\{H_{01}, H_{02}, H_{03}\}$. We denote the data for subject m by (I_m, S_m, A_m, Y_m), which includes the stage the subject entered the trial ($I_m \in \{1, 2\}$), the subject's subpopulation ($S_m \in \{1, 2\}$), the subject's study arm assignment ($A_m \in \{0, 1\}$), and the subject's outcome ($Y_m \in \mathbb{R}$), respectively.

We assume in each stage, for each subpopulation sampled in that stage, that an equal number are assigned to each study arm (though it is possible to relax this assumption). We

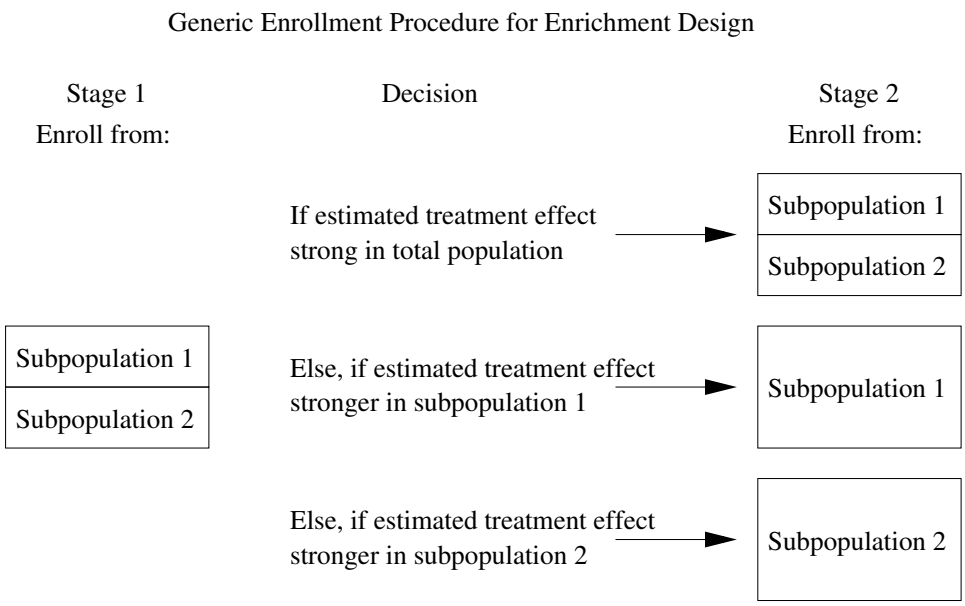

Figure 1: A flow chart depicting a type of enrichment design within our class, where enrollment in stage two can be restricted to the subpopulation with larger z-statistic based on data from stage one. Each labeled box represents a set of patients enrolled from a certain subpopulation.

assume for each subject, that conditioned on the subject's subpopulation s and study arm assignment a, his/her outcome Y is drawn from an unknown data generating distribution Q_{sa} independent of the other subjects' data. The only assumptions we make on the data generating distributions Q_{sa} are that there exists a $\gamma > 0$ and $M > 0$ that do not depend on sample size, such that the support of Q_{sa} is on $[-M, M]$ and the variance $\sigma^2(Q_{sa})$ of Q_{sa} is at least γ.

Let n_i denote the total number of subjects in stage i, which is preplanned. We require that the fraction of subjects in the first stage $n_1/(n_1 + n_2)$ be at least 0.05 and at most 0.95. We also make an assumption about the fraction of stage one subjects who are from each subpopulation $s \in \{1, 2\}$, which we denote p_s. We consider each p_s fixed (non-random) and unknown, since in practice one may not know in advance how many subjects will enroll from each subpopulation. However, we assume there is a non-negligible fraction of subjects in stage one from each subpopulation; that is, we assume for some $\tau > 0$ not depending on the sample size, that for each s, $p_s > \tau$.

Let $\mathbf{T^{(1)}} = (T_1^{(1)}, T_2^{(1)}, T_3^{(1)})$ denote the stage one z-statistics for subpopulation 1, subpopulation 2, and the total population, respectively, defined as

$$T_1^{(1)} = \sum_{m:I_m=1,S_m=1} [(Y_m A_m - Y_m(1 - A_m)]/\text{se}_1, \tag{1}$$

$$T_2^{(1)} = \sum_{m:I_m=1,S_m=2} [(Y_m A_m - Y_m(1-A_m)]/\text{se}_2, \quad (2)$$

$$T_3^{(1)} = \sum_{m:I_m=1} [(Y_m A_m - Y_m(1-A_m)]/\text{se}_3, \quad (3)$$

where the se_i are defined by

$$\text{se}_1 = \sqrt{(n_1 p_1/2)\sigma^2(Q_{10}) + (n_1 p_1/2)\sigma^2(Q_{11})}, \quad (4)$$

$$\text{se}_2 = \sqrt{(n_1 p_2/2)\sigma^2(Q_{20}) + (n_1 p_2/2)\sigma^2(Q_{21})}, \quad (5)$$

$$\text{se}_3 = \sqrt{\text{se}_1^2 + \text{se}_2^2}. \quad (6)$$

Here, se_i is the standard error for the corresponding sum in the definition of $T_i^{(1)}$, conditioned on sample sizes in each subpopulation and treatment category. For simplicity we assume these variances are known, though we allow them to differ by subpopulation and treatment category. It is possible to extend our results to the more realistic case where these variances are estimated.

We now describe the final test statistic and the rejection region for each null hypothesis. After stage two is completed, we compute the following z-statistic based on all the data from that stage:

$$T^{(2)} = \sum_{m:I_m=2} [Y_m A_m - Y_m(1-A_m)]/\text{se}_4, \quad (7)$$

where se_4 is the standard error for the corresponding sum, conditioned on the enrollment decision made just after stage one, defined analogously as (4), (5), and (6); we defer the formal definition of se_4 to Section C of the Supplementary Material.

We reject the null hypothesis H_{0j} corresponding to the subpopulation (or the total population) enrolled in stage two, if the following weighted combination of z-statistics from both stages:

$$\sqrt{n_1/(n_1+n_2)}T_3^{(1)} + \sqrt{n_2/(n_1+n_2)}T^{(2)}, \quad (8)$$

exceeds c for some prespecified threshold c. That is, if only subpopulation 1 is enrolled in stage two and (8) exceeds c, we reject H_{01}; if only subpopulation 2 is enrolled in stage two and (8) exceeds c, we reject H_{02}; if both subpopulations are enrolled in stage two and (8) exceeds c, we reject H_{03}.

We define our class of enrichment designs to be those that use a decision rule $D \in \mathbf{D}$, the z-statistics defined above, and the rejection rule defined in the previous paragraph.

3.3 Application of general method from Section 5 to above class of designs

The general method we present in Section 5 allows one to compute, for each enrichment design in the class defined above, the minimum threshold value c in (8) that guarantees strong control of the asymptotic, familywise Type I error at a given level α. This is useful

since it is quite difficult, if not impossible, to determine this minimum threshold value c analytically. Our general method gets around this obstacle by reducing this problem to an optimization problem that can be solved numerically with standard software, to any desired precision. This general method can be applied to a wide variety of designs, not only to those considered above.

We obtain, as shown in Section 5 and Sections B-E of the Supplementary Material, that under the above assumptions, for any decision rule $D \in \mathbf{D}$, that the minimum threshold value c that guarantees strong control of the asymptotic, familywise Type I error at level $\alpha = 0.05$ is $\Phi^{-1}(0.95)$, for Φ the cumulative distribution function of the standard normal. This is exactly the same threshold as would have been used in a standard fixed design, with no preplanned change. This is surprising, since allowing designs that adapt to accrued data has the potential to result in Type I error inflation. We show for the above class of enrichment designs that no such inflation can occur, at least asymptotically. Thus, the above class of designs allows flexibility at no price in terms of a stricter rejection threshold. Furthermore, in Section 4 we show one such design that provides a substantial gain in the probability of correctly detecting an effect when it exists in just one subpopulation, compared to a fixed design (though there is some price paid when there is an effect in both subpopulations).

In the above class of enrichment designs, all the first stage data (from both subpopulations) is used in the final test statistic (8), even when the final hypothesis tested concerns only one of these subpopulations. Though this may initially appear odd, since data from one subpopulation may influence whether the null hypothesis for the other subpopulation is rejected, this is a property shared by virtually all procedures for controlling familywise Type I error (except the conservative procedure of Bonferroni); for example, this property holds for the closure principle of Marcus et al. (1976), the step-down procedure of Holm (1979), and the step-up procedure of Hochberg (1988).

To explain the intuition behind why the above class of designs can be advantageous, consider the case where subpopulation 2 (those with severe baseline depression) has the larger corresponding z-statistic in stage one, and enrollment is from only this subpopulation in the second stage. Then the null hypothesis tested at the end of the study is H_{02}, the null hypothesis of no mean treatment effect in subpopulation 2. If we were to test H_{02} based only on the data from subpopulation 2 (thereby throwing out the stage one data for subpopulation 1), this would introduce a selection bias due to essentially hiding data that was unfavorable. A correction can be done to account for this, by raising the threshold that the final test statistic must exceed before the null hypothesis H_{02} is rejected, as is done in a similar situation in (Thall et al., 1988). In contrast, our testing procedure does not require any such raised threshold; the reason, intuitively, is that our test statistic already includes the penalty of having to incorporate data from the subpopulation that had a weaker signal. This intuitive argument is proved rigorously in Sections B-E of the Supplementary Material. It is an area of further research to compare our designs to ones that throw out data from discontinued subpopulations and correct for the resulting selection bias by incorporating a prespecified penalty.

4 Power Comparison between adaptive design from Section 3 and a standard fixed design

We show in Section 5 and Sections B-E of the Supplementary Material that for each decision rule in the class $\mathbf{D}$ given above, the resulting adaptive design guarantees strong control of asymptotic, familywise Type I error at level $\alpha = 0.05$, when the rejection threshold c for the final statistic (8) is set to $\Phi^{-1}(0.95)$. Having shown the validity of such enrichment designs in terms of Type I error control, we now consider their power. We present a particular design from the class $\mathbf{D}$, and compare its power to that of a standard fixed design, under several scenarios. We continue to use the example of testing a hypothetical new antidepressant. We are motivated by the meta-analysis of Kirsch et al. (2008), who present suggestive, but not conclusive, evidence that a class of antidepressants may not be superior to placebo for those with moderate pretreatment depression (subpopulation 1 in what follows), while they are superior to placebo for those with severe pretreatment depression (subpopulation 2 in what follows).

To specify a particular adaptive design from the class in the previous section, we need to choose a decision rule $D \in \mathbf{D}$. Recall that each $D \in \mathbf{D}$ is a function from the stage one statistics to two possible choices for stage two enrollment: (i) enroll from both subpopulations as in stage one or (ii) enroll only from the subpopulation corresponding to the larger of the stage one z-statistics $T_1^{(1)}$, $T_2^{(1)}$. We define our decision rule D to choose option (i) if the z-statistic $T^{(1)} > T^{(2)}$ or $T^{(1)} > 0.2$, and option (ii) otherwise.

The intuition behind our choice of decision rule is that if $T^{(1)} > T^{(2)}$ or $T^{(1)} > 0.2$, then there is at least some hope that the treatment is effective for subpopulation 1 (those with moderate pretreatment depression), and so we continue enrolling them (as well as those from subpopulation 2) in stage two. Otherwise we give up on subpopulation 1 and only enroll from subpopulation 2 for the rest of the trial. We next explain how we chose the threshold of 0.2.

Since the above design is in the class $\mathbf{D}$, the asymptotic, familywise Type I error is strongly controlled, as described above. Consider replacing the threshold 0.2 above by any other prespecified threshold. Since the resulting design remains in the class $\mathbf{D}$, the same familywise Type I error guarantee holds. It is therefore possible to examine various threshold values, and select based on which one gives the most power in scenarios of interest. We selected the threshold 0.2 to give a good tradeoff in power in the scenarios considered below.

In Section G of the Supplementary Materials we compare the familywise Type I error and power of this enrichment design to a standard fixed design, under a variety of scenarios. We present five such scenarios here, which we number 1a, 1b, 2a, 2b, and 3. In all scenarios, we assume that in stage one, equal numbers are enrolled from each subpopulation.

In scenarios 1a and 1b, we use data generating distributions for each subpopulation constructed to reflect what was seen in the meta-analysis of Kirsch et al. (2008), i.e. no effect of the treatment compared to placebo in those with moderate initial depression, but a mean improvement $r > 0$ comparing treatment vs. placebo in those with severe initial

depression. In scenario 1a, this mean improvement r is set to be 1.8 HRSD points, which is the point estimate of this quantity in (Kirsch et al., 2008); in scenario 1b, the mean improvement r is set to the more optimistic value of 3 HRSD points, which was seen in (Kirsch et al., 2008) for subjects with very severe initial depression.

In scenarios 2a and 2b, we use data generating distributions where the treatment is equally effective for both subpopulations. The mean improvement r' in treatment over placebo is set to 1.8 HRSD points in scenario 2a, and 3 HRSD points in scenario 2b. Lastly, in scenario 3, we use data generating distributions where there is no effect in any subpopulation, so we can assess Type I error of the adaptive and fixed designs.

Other parameters of the data generating distributions, such as the variance in each arm, were based on the estimates of these quantities in Kirsch et al. (2008). Total sample size was fixed in all the scenarios at 488 subjects, which was chosen to provide 80% power to the fixed design under scenario 2a. For each scenario, and for each of the two trial designs (adaptive and fixed), 100,000 simulated trials were run. See Section F of the Supplementary Materials for full details of how the data generating distributions for each scenario are defined, and for the R code used in the simulations.

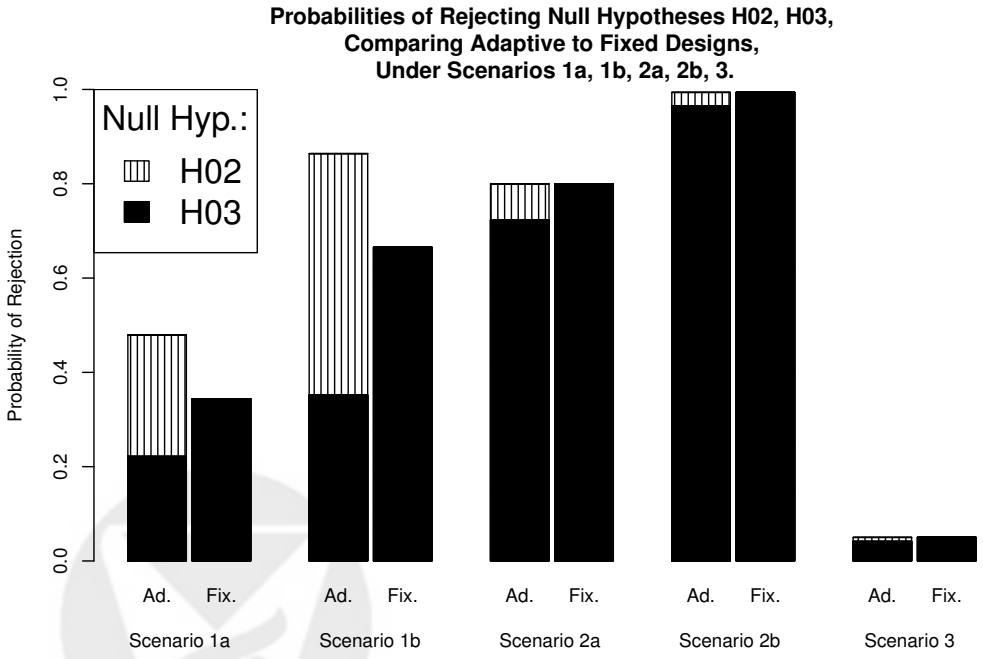

Figure 2: For each of five scenarios, we compare the performance of the adaptive (abbreviated "Ad.") to fixed (abbreviated "Fix.") designs. In each comparison, we depict the probability of rejecting each of the null hypotheses H_{02} and H_{03}, which are both false in all scenarios except scenario 3. By construction, neither of the designs ever rejects H_{01}.

In each scenario, for both the adaptive design and fixed design, the proportion of

simulated trials leading to rejection of each null hypothesis (among H_{01}, H_{02}, H_{03}) was recorded. Figure 2 shows, for each scenario, side-by-side bar plots of the probability that each null hypothesis is rejected; the bar on the left is for the adaptive design, and the bar on the right is for the fixed design. For the adaptive design we are considering, which by construction can lead to rejection of exactly one of the null hypotheses H_{02}, H_{03} (or no rejection at all), the corresponding bar is shaded to reflect the proportion of simulated trials in which each null hypothesis is rejected; the length of the striped part of the bar is the proportion of simulated trials in which H_{02} is rejected, and the length of the solid black part of the bar is the proportion of simulated trials in which H_{03} is rejected. In contrast, since the fixed design can only lead to rejection of H_{03} (or no rejection at all), the corresponding bar is always shaded solid black corresponding to rejecting H_{03}.

Consider the five scenarios depicted in Figure 2, from left to right. We call the probability of rejecting at least one of H_{02}, H_{03} when both are false (which is the case in all scenarios except scenario 3) the "overall power." In scenarios 1a and 1b, where the treatment is only effective for subpopulation 2, the adaptive design has more overall power than the fixed design, by 14% and 20%, respectively. In scenarios 2a and 2b, where the treatment is equally effective for both subpopulations, the overall power is roughly equal for both designs; however, the adaptive design fails to reject H_{03} more often, by 8% and 3%, respectively. In scenario 3, where the treatment has no effect for either subpopulation, Type I error is 0.05 for both designs.

We summarize the tradeoff involved in using the adaptive vs. fixed design, in the scenarios considered above. The adaptive design allows potential gains of 14%-20% in overall power in the scenarios where the treatment only works for those with severe pretreatment depression (subpopulation 2). In the scenarios where the treatment is equally effective for both subpopulations, the overall power of the two designs is roughly the same, but the adaptive design leads to potential losses of 3%-8% in ability to reject the null hypothesis H_{03} for the total population.

It is possible to augment both the above adaptive design and fixed design to include a subsequent test of the null hypothesis H_{02} whenever H_{03} is rejected, in a way that leads to no inflation of asymptotic, worst-case, familywise Type I error. We do not consider such designs here, though comparisons of such designs are an important area for future research.

5 General Method and Theorem

5.1 Overview

Our main theorem allows one to reduce the problem of computing the asymptotic, worst-case, familywise Type I error of a wide variety of hypothesis tests in adaptive designs to a manageable optimization problem. Before presenting our main theorem, we describe the following: the components of the design that must be prespecified; the data generating distributions and statistical model; the definition of asymptotic, worst-case, familywise Type I error; our assumptions on the statistics used; and the intuition behind the theorem.

After presenting our theorem, we show how it is applied to the class of enrichment designs from Section 3 to precisely characterize the asymptotic, worst-case, familywise Type I error for these designs.

We focus only on two stage designs below, for clarity. It is straightforward to generalize these definitions, assumptions, and our main theorem to designs with any number of stages. The theorem below can be easily generalized to apply to designs that incorporate adaptations in addition to changing the population sampled, such as changes to sample size and to randomization probabilities, but we do not present such designs here.

5.2 Items requiring prespecification

We require the following to be prespecified in any design we consider:

1. a finite set of subpopulations of interest $\mathcal{S}$, and a finite set of treatment arms $\mathcal{A}$.

2. the total number of subjects to be enrolled in each stage i, denoted by n_i. We assume there is a positive constant r, independent of sample size, such that $n_2 = \lfloor r n_1 \rfloor$. We let $n = (n_1, n_2)$.

3. for each stage i, a vector $\mathbf{T^{(i)}}$ of statistics taking values in $\mathbb{R}^{t_i}$, for some t_i. Each $\mathbf{T^{(i)}}$ is a function only of the stage i data and the enrollment decisions made before stage i.

4. a randomization procedure, specifying the probability that each subject is assigned to each treatment arm. We allow this procedure to be standard randomization, block randomization, stratified randomization, or an adaptive method of randomization.

5. a finite set $\mathcal{E}$ of potential enrollment procedures for stage two, each of which specifies the proportion of stage two subjects to enroll from every subpopulation. We assume, for each enrollment procedure in $\mathcal{E}$, that one can enroll in stage two to exactly achieve these proportions.

6. a decision function (or rule) D that maps each value of the vector of stage one statistics $\mathbf{T^{(1)}}$ to an element in the set $\mathcal{E}$ of potential enrollment proportions for each subpopulation for stage two. We restrict the complexity of the decision rule in the following sense: we require that for each possible enrollment procedure $\epsilon \in \mathcal{E}$, the set of values $D^{-1}(\epsilon)$ of the statistics $\mathbf{T^{(1)}}$ leading to this choice is a finite union of Borel measurable, convex subsets of $\mathbb{R}^{t_1}$.

7. A set of null hypotheses $\mathbf{H_0} = \{H_{0j}\}_{j \in J}$ to be tested, and for each one, a corresponding rejection region $R_j \subseteq \mathbb{R}^{t_1+t_2}$. Each null hypothesis H_{0j} represents the subset of the possible data generating distributions $\mathcal{Q}$ (defined below) for which this null hypothesis is true. We denote the set of rejection regions for the null hypotheses by $\mathbf{R} = \{R_j\}_{j \in J}$, and require that each rejection region R_j is a finite union of

Borel measurable, convex subsets of $\mathbb{R}^{t_1+t_2}$. At the end of study, we reject all null hypotheses H_{0j} for which $\left(\mathbf{T^{(1)}}, \mathbf{T^{(2)}}\right) \in R_j$.

5.3 Data generating distributions and statistical model

We assume that for each subject, conditioned on the subject's subpopulation s and study arm assignment a, his/her outcome Y is a random draw from an unknown data generating distribution Q_{sa}. We assume each such random draw is independent of the subpopulations, treatment assignments and outcomes of all the other subjects. For ease of reference, we denote the set of data generating distributions Q_{sa} by $Q' = \{Q_{sa}\}_{a\in\mathcal{A},s\in\mathcal{S}}$. We assume that Q' is an element of a statistical model $\mathcal{Q}'$. This may be a nonparametric model, such as the one specified in Section 3.2 (which puts no constraints on each Q_{sa} except for having bounded support and a minimum variance), a semiparametric model, or a parametric model.

Denote the fraction of stage one subjects who are from subpopulation $s \in \mathcal{S}$ by p_s, which we allow to be unknown before the trial starts. We denote the set of subpopulation proportions enrolled in stage one by $p = \{p_s\}_{s\in\mathcal{S}}$. Let $\mathbf{p}$ denote the set of possible p. This could be unconstrained, or constraints can be imposed based on knowledge of possible enrollment proportions.

The joint distribution of the statistics $\mathbf{T} = \left(\mathbf{T^{(1)}}, \mathbf{T^{(2)}}\right)$, which we denote by $P^{(n)}$, is a function of the data generating distributions Q', the stage one subpopulation proportions p, and the sample size n. Implicitly, $P^{(n)}$ also depends on the prespecified quantities listed in Section 5.2, including the set of subpopulations $\mathcal{S}$, the set of treatments $\mathcal{A}$, the possible enrollment procedures for stage two $\mathcal{E}$, the definition of the statistics $\mathbf{T}$, and the decision rule D.

To simplify notation in what follows, we bundle Q' and p together and denote the pair (Q', p) by Q. We let $\mathcal{Q}$ denote $\mathcal{Q}' \times \mathbf{p}$, which represents the class of all possible data generating distributions and stage one subpopulation enrollment proportions (though we refer to each $Q \in \mathcal{Q}$ below, for brevity, as a "data generating distribution").

5.4 Definition of asymptotic, worst-case, familywise type I error

The main problem addressed in this paper is to compute the asymptotic, worst-case familywise Type I error for a given design. The familywise Type I error at sample size n, and data generating distribution $Q \in \mathcal{Q}$, is

$$P^{(n)}\left(\mathbf{T} \in \bigcup_{j:Q\in H_{0j}} R_j\right). \tag{9}$$

The worst-case, familywise Type I error is defined as the supremum over $Q \in \mathcal{Q}$ of (9). The asymptotic, worst-case, familywise Type I error is defined as

$$\limsup_{n\to\infty} \sup_{Q\in\mathcal{Q}} P^{(n)} \left(\mathbf{T} \in \bigcup_{j:Q\in H_{0j}} R_j \right),$$

where $n \to \infty$ is shorthand for $n_1, n_2 \to \infty$, under the constraint given above in Section 5.2 that $n_2 = \lfloor r n_1 \rfloor$ for some positive constant r. We say the asymptotic, familywise Type I error is strongly controlled at level α if the above expression is at most α.

Having the supremum over $Q \in \mathcal{Q}$ nested inside the limit as n goes to infinity, allows for different sets of distributions to have the worst-case, familywise Type I error at each sample size n. Designs in which asymptotic, worst-case familywise Type I error is at most 0.05 guarantee control over familywise Type I error, uniformly over the data generating distributions, as sample size grows to infinity, which is a desirable property.

5.5 Assumptions on statistics

In our theorem below, we make two assumptions on the distributions of the statistics $\mathbf{T}$. These assumptions are quite general, and we expect they will hold for a wide variety of designs. The first assumption in the theorem is that the vector of stage two statistics $\mathbf{T^{(2)}}$ is independent of the stage one statistics $\mathbf{T^{(1)}}$, given the enrollment decision $D(\mathbf{T^{(1)}})$ made just after stage one. This will hold, as is the case in the class of designs from Section 3, when the decision $D(\mathbf{T^{(1)}})$ determines a subpopulation (or the total population) to obtain an independent random sample from in stage two, but otherwise the data in stage one has no effect on the enrollment procedure in stage two. This conditional independence assumption is broad enough to allow design changes such as early stopping, sample size adjustments, and adaptation of randomization probabilities, as long as these are preplanned changes based on statistics from stage one.

The second assumption in the theorem involves several parts, which require the introduction of additional notation. Recall that $P^{(n)}$, the joint distribution of the statistics $\mathbf{T}$, depends on the data generating distribution $Q \in \mathcal{Q}$ and the sample size n. For any $Q \in \mathcal{Q}$ and sample size n, denote the mean under $P^{(n)}$ of the first stage statistics $\mathbf{T^{(1)}}$ by $\mu_n^{(1)}(Q)$, and denote the distribution of the centered, stage one statistics $\mathbf{T^{(1)}} - \mu_n^{(1)}(Q)$, by $G_n^{(1)}(Q)$. We assume, for any $Q \in \mathcal{Q}$, that $G_n^{(1)}(Q)$ converges under $P^{(n)}$ to a zero mean, multivariate normal distribution $G^{(1)}(Q)$ as sample size $n \to \infty$.

We assume the vector of stage two, centered statistics, conditioned on the enrollment decision $D(\mathbf{T^{(1)}})$, converges to a multivariate normal distribution. To make this precise, for any $Q \in \mathcal{Q}$, for each potential stage two enrollment procedure $\epsilon \in \mathcal{E}$, let $\mu_n^{(2)}(Q, \epsilon)$ denote the conditional mean under $P^{(n)}$ of stage two statistics $\mathbf{T^{(2)}}$ given $D(\mathbf{T^{(1)}}) = \epsilon$. Also, let $G_n^{(2)}(Q, \epsilon)$ denote the conditional distribution under $P^{(n)}$ of the centered statistics $\mathbf{T^{(2)}} - \mu_n^{(2)}(Q, \epsilon)$ given $D(\mathbf{T^{(1)}}) = \epsilon$. We assume for any $Q \in \mathcal{Q}$ and $\epsilon \in \mathcal{E}$, that $G_n^{(2)}(Q, \epsilon)$

converges to a zero mean, multivariate normal distribution $G^{(2)}(Q, \epsilon)$ as sample size $n \to \infty$. We allow the covariance of this limit distribution to depend on the data generating distribution Q.

We furthermore assume the above convergence of centered stage one and stage two statistics is uniform over data generating distributions $Q \in \mathcal{Q}$. That is, we assume for all possible enrollment decisions $\epsilon \in \mathcal{E}$,

$$\lim_{n\to\infty} \sup_{Q\in\mathcal{Q}, C\in\mathcal{C}} \left| \int_C d(G_n^{(1)}(Q) \times G_n^{(2)}(Q,\epsilon)) - \int_C d(G^{(1)}(Q) \times G^{(2)}(Q,\epsilon)) \right| = 0, \quad (10)$$

where $\mathcal{C}$ denotes the set of all Borel measurable, convex subsets of $\mathbb{R}^{t_1+t_2}$. In many cases, such as the class of designs from Section 3, the above uniform convergence can be obtained immediately from uniform central limit theorems such as those in (Götze, 1991).

5.6 Intuition behind theorem

The theorem below reduces the problem of computing asymptotic, worst-case, familywise Type I error to a more manageable optimization problem. We present some of the intuition behind the theorem. We consider designs where the only way in which the outcomes for subjects in stage one impact the design in stage two is through the enrollment decision $D(\mathbf{T}^{(1)})$. Though this is not a requirement of the theorem below, which only requires the assumptions of the previous sections, it is easier to explain the intuition for such designs.

A familywise Type I error occurs if the vector of statistics $\mathbf{T}$ falls in the rejection region corresponding to a null hypothesis that is true. In our notation above, this event can be written as $\mathbf{T} \in \bigcup_{j:Q\in H_{0j}} R_j$. We partition this event into more manageable pieces by intersecting it, for each possible stage two enrollment procedure $\epsilon \in \mathcal{E}$, with the event that our decision rule D says to enroll using ϵ. Define each such intersection by

$$A_\epsilon = \left\{ \mathbf{T^{(1)}} \in D^{-1}(\epsilon) \qquad \& \qquad (\mathbf{T^{(1)}}, \mathbf{T^{(2)}}) \in \cup_{j:Q\in H_{0j}} R_j \right\}. \quad (11)$$

By construction, on each such event A_ϵ, the enrollment decision $D(\mathbf{T^{(1)}})$ after stage one is always ϵ. It follows that the probability of A_ϵ when $\mathbf{T}$ is generated by the adaptive design is the same as the probability of A_ϵ were $\mathbf{T}$ instead generated by the fixed design that always makes enrollment decision ϵ regardless of the stage one data. Thus, by taking the intersection with each possible enrollment decision, we reduce the problem of computing familywise Type I error for an adaptive design to that of computing it for several fixed designs, which immensely simplifies our computations.

Combining the above with the assumption from Section 5.5 that the centered statistics converge to a multivariate normal, we succeed in reducing the original problem to one of computing, for each element of a class of easy to describe regions, the probability that a zero mean, multivariate normal distribution falls in that region. These regions are the ones in braces in (11), except that we need to center them (since we will be using centered

statistics); we define the centered versions of these regions, for each possible enrollment decision $\epsilon \in \mathcal{E}$, by

$$V(Q, n, \epsilon) = \left(\left(D^{-1}(\epsilon) \times \mathbb{R}^{t_2} \right) \bigcap \left(\cup_{j:Q \in H_{0j}} R_j \right) \right) - \left(\mu_n^{(1)}(Q), \mu_n^{(2)}(Q, \epsilon) \right),$$

where for any set $B \subseteq \mathbb{R}^k$, and vector $v \in \mathbb{R}^k$, we define the shifted version of B as $B - v = \{x \in \mathbb{R}^k : x + v \in B\}$. We next present our main theorem.

Theorem 1 *(Equivalence of asymptotic, worst-case, familywise Type I error to the solution of an optimization problem) For any two stage design that adheres to the specifications in Section 5.2, under the assumptions given in Sections 5.3 and 5.5, we have*

$$\lim_{n \to \infty} \sup_{Q \in \mathcal{Q}} \left| P^{(n)} \left(\mathbf{T} \in \bigcup_{j:Q \in H_{0j}} R_j \right) - \sum_{\epsilon \in \mathcal{E}} \int_{V(Q,n,\epsilon)} d\left(G^{(1)}(Q) \times G^{(2)}(Q, \epsilon) \right) \right| = 0.$$

which implies the asymptotic, worst-case, familywise Type I error equals

$$\limsup_{n \to \infty} \sup_{Q \in \mathcal{Q}} \sum_{\epsilon \in \mathcal{E}} \int_{V(Q,n,\epsilon)} d\left(G^{(1)}(Q) \times G^{(2)}(Q, \epsilon) \right). \tag{12}$$

The expression (12) is an optimization problem, which involves finding the maximum value of the sum in the previous display, over the class of possible data generating distributions $Q \in \mathcal{Q}$, and computing its $\limsup$ as sample size n goes to infinity. A key to solving this optimization problem is that the limit distributions $G^{(1)}(Q), G^{(2)}(Q, \epsilon)$ and the set $V(Q, n, \epsilon)$ often depend on the distribution Q only through a small number of scalar parameters. This simplifies the computation of the inner supremum in (12) to finding the supremum over the set of possible values of these scalar parameters. Evaluating the integral in (12) at a particular Q, n, and ϵ is often easy, using standard functions in statistical software for computing the distribution function of a multivariate normal distribution.

For the class of designs from Section 3, we show in Section C of the Supplementary Material that the expression inside the $\limsup_{n \to \infty}$ in (12) has the same value for every sample size n; this makes evaluation of (12) even easier. We conjecture that for many designs, the value of the expression inside the $\limsup_{n \to \infty}$ in (12) will either not depend on n or will have an easy to prove limit as $n \to \infty$, though this is an open question requiring future research.

5.7 Application of theorem to class of enrichment designs from Section 3

We now give an overview of how, in Sections C-E of the Supplementary Materials, we apply the above theorem to the class of enrichment designs in Section 3, to show each of these designs strongly controls asymptotic, familywise Type I error at level 0.05. We do

this under the assumptions in Section 3.2, and where the rejection threshold in the final test statistic (8) is set to $\Phi^{-1}(0.95)$ (which is the same threshold used in the corresponding fixed design). Below we emphasize the main ideas in applying our method. Full details are given in Sections C-E of the Supplementary Material.

We first verify the conditions of the theorem hold for the class of enrichment designs in Section 3. It is straightforward to verify the class of enrichment designs meets the criteria given in Section 5.2 for which aspects of the design must be prespecified. The assumptions in Section 5.3 follow from those in Section 3.2. It remains to verify the two assumptions from Section 5.5. The first assumption follows since by construction, the stage two data only depend on the stage one data through the decision $D(\mathbf{T}^{(\mathbf{1})})$ of which subpopulation (or the total population) to enroll in stage two. The second assumption from Section 3 follows by applying the multivariate Berry-Esseen central limit theorem of (Götze, 1991), as described in Section C of the Supplementary Materials. Having verified the assumptions required by the theorem, we can now apply it.

The above theorem implies the asymptotic, worst-case, familywise Type I error for each design D in the class $\mathbf{D}$ in Section 3 equals the integral (12). (In (12), the regions $V(Q, n, \epsilon)$ and the limit distributions $G^{(1)}(Q)$ and $G^{(2)}(Q, \epsilon)$ implicitly depend on the decision rule D.) In Section C of the Supplementary Materials, using properties of multivariate normal distributions, we show that for fixed Q, n, ϵ, the integral nested inside (12) equals the probability that a certain bivariate normal distribution falls in the region $[0, \infty) \times [0, \infty)$; we show the mean and covariance of this bivariate normal distribution are simple functions of the sample size n, the enrollment decision ϵ, the first stage enrollment proportions p for each subpopulation, and the means and variances of the subpopulation and treatment specific outcome distributions in Q. This has two advantages. The first is that the inner supremum in (12) can be recast as the supremum over a finite dimensional space; the second is that for each point in this space, the integral in (12) can be computed directly from the multivariate normal distribution function. For example, the package mvtnorm can be used in R to compute this.

We have reduced the problem of evaluating (12) to finding the supremum of an easy to compute function over a finite dimensional space. We upper bound this supremum by combining a grid search over this space with an analytic upper bound on the approximation error of the grid search. The grid search involves computing the value of the summation in (12) at each point in a given grid; we do this using a simple program in R, given in Section D of the Supplementary Materials. We are able to select the size of the grid to be sufficiently large and the distance between neighboring grid points to be sufficiently small that the maximum value found in the corresponding grid search plus an upper bound on the approximation error is no more than 0.05. This implies that (12) is at most 0.05. It is straightforward to show this bound is sharp. Full details and R code are given in the Supplementary Materials.

6 Discussion

A limitation of our results is that they are asymptotic, that is, our guarantees on Type I error control are in the limit as sample sizes in both stages of the adaptive design go to infinity. However, even in fixed designs, the standard t-test comparing means in the treatment and control groups has this issue; it is only guaranteed to have correct Type I error for testing null hypotheses such as those in Section 3, in the limit as sample size goes to infinity (unless parametric or other assumptions are made on the data generating distribution).

One important issue we did not address is the potential loss of generalizability in designs that focus on particular subpopulations. There is a risk that such designs will lead to conclusions applicable to a smaller population than a fixed design would have. However, an advantage of the above adaptive designs is that when treatments are truly only effective for one subpopulation, adaptive designs can have more power to discover this.

An area of future work is deriving the bias and mean squared error of maximum likelihood estimators for the adaptive designs considered in this paper. Also, methods for constructing confidence intervals are important for such designs. Important related work in this area includes (Jennison & Turnbull, 1984, 1989; Proschan & Hunsberger, 1995; Lehmacher & Wassmer, 1999; Brannath et al., 2006; Wu et al., 2010). It is an open problem to determine the asymptotic, worst-case bias and mean squared error, as well as to construct confidence intervals that are guaranteed to be asymptotically conservative.

Acknowledgment

This research was supported by the National Institutes of Health, U.S.A.

Supplementary Material

Supplementary material is available online at

`http://people.csail.mit.edu/mrosenblum/papers/subpop_sm.pdf`

References

BASELGA, J. (2001). Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: pivotal trials. *Oncology* **61(S2)**, 14–21.

BAUER, P. & KÖHNE, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**, 1029-1041.

BISCHOFF, W. & MILLER, F. (2005). Adaptive two-stage test procedures to find the best treatment in clinical trials. *Biometrika* **92**, 197–212.

BRANNATH, W., KÖNIG, F. & BAUER, P. (2006). Estimation in flexible two stage designs. *Statistics in Medicine* **25**, 3366–3381.

FDA (2010). Draft Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics.
http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf

FOLLMANN, D. (1997). Adaptively changing subgroup proportions in clinical trials. *Statistica Sinica* **7**, 1085–1102.

FOLLMANN, D. A., PROSCHAN, M. A. & GELLER, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**, 325–336.

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

HOCHBERG, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

HOCHBERG, Y. & TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. Probability and Mathematical Statistics. Wiley–Interscience.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

HOMMEL, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* **43**, 581–589.

JENNISON, C. & TURNBULL, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5**, 33–45.

JENNISON, C. & TURNBULL, B. W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society, Series B* **51**, 305–361.

JENNISON, C. & TURNBULL, B. W. (2006). Discussion: Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Opportunities and limitations. *Biometrical Journal* **48**, 650–655.

JENNISON, C. & TURNBULL, B. W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharmaceutical Statistics* , 1135–1161, doi: 10.1080/10543400701645215.

KIESER, M., BAUER, P. & LEHMACHER, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* **41**, 261–277.

KIRSCH, I., DEACON, B., HUEDO-MEDINA, T., SCOBORIA, A., MOORE, T. & ET AL. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration. *PLoS Med* **5**, e45. doi:10.1371/journal.pmed.0050045.

LEHMACHER, W. & WASSMER, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.

MARCUS, R., PERITZ, E. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

PROSCHAN, M. & HUNSBERGER, S. (1995). Designed extension studies based on conditional power. *Biometrics* **51**, 1315–1324.

SAMPSON, A. R. & SILL, M. W. (2005). Drop-the-losers design: Normal case. *Biometrical Journal* **47**, 257–268.

SCHAID, D. J., WIEAND, S. & THERNEAU, T. M. (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika* **77**, 507–513.

STALLARD, N. & TODD, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* **22**, 689–703.

THALL, P. F., SIMON, R. & ELLENBERG, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–310.

WANG, S.-J., HUNG, H. & O'NEILL, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* **51**, 358–374.

WANG, S. J., ONEILL, R. T. & HUNG, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist.* **6**, 227–244.

WU, S. S., WANG, W. & YANG, M. C. K. (2010). Interval estimation for drop-the-losers designs. *Biometrika* **97**, 405–418.