Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment

BY M. ROSENBLUM

*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,
615 N. Wolfe St., Room E3616, Baltimore, Maryland 21205, U.S.A.
mrosenbl@jhsphe.edu*

AND M. J. VAN DER LAAN

*Division of Biostatistics, University of California, School of Public Health,
101 Haviland Hall, Berkeley, California 94720-7358, U.S.A.
laan@stat.berkeley.edu*

SUMMARY

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 enrolled, 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 allow changes to the population enrolled based on interim data using a prespecified decision rule, for which the asymptotic, familywise Type I error rate is strongly controlled at a specified level α . As a demonstration of our method, we prove new, sharp results for a simple, two-stage enrichment design. We then compare this design to fixed designs, focusing on each design's ability to determine the overall and subpopulation-specific treatment effects.

Some key words: Adaptive design; Enrichment design; Group sequential design; Optimization; Patient-oriented research; Randomized trial; Subpopulation.

1. INTRODUCTION

In many randomized trials, participants are enrolled sequentially over time, and interim analyses are done at prespecified points during the trial. We consider settings where the follow-up time is relatively short. At interim analyses, it is then possible to look at the 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 biomarker measured at baseline, such as human epidermal growth factor receptor-2 in studies of breast cancer therapies, as described by [Baselga \(2001\)](#) and [Wang et al. \(2007\)](#),

2009), or defined by baseline factors such as initial severity of depression in studies of antidepressants as described by Kirsch et al. (2008). In both examples, there is evidence that certain subpopulations may benefit more from certain treatments. In the first example, there is evidence that those with higher levels of human epidermal growth factor receptor-2 may benefit more from trastuzumab, and in the second example there is evidence that those with higher baseline severity of depression may benefit more from certain antidepressants. As we show below, designs that allow preplanned changes to which subpopulations are enrolled can lead to more power and reveal more information about subpopulation-specific treatment effects.

For 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 those with severe baseline depression for the rest of the trial. We show in § 4 that certain designs of this type can give improved power, compared with a standard fixed design, at certain alternatives of interest. By a standard fixed design, we mean a design that does not allow changes to the population enrolled, and that uses a one-sided z -test for comparing two means with a fixed sample size. We also show a power improvement for such enrichment designs compared with a fixed design using the Hochberg multiple testing procedure, at certain alternatives of interest. Due to the multiple hypotheses potentially tested in an enrichment design, a major challenge for enrichment 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 the 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 consider only designs that guarantee clear interpretability of trial results. We briefly discuss estimation and confidence intervals in § 6.

The familywise Type I error rate is the probability that at least one true null hypothesis is rejected. Throughout, we slightly abbreviate the familywise Type I error rate by the 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 § 5.4. When we say a study design strongly controls this error at level α , we mean that the corresponding asymptotic, worst-case, familywise Type I error is at most α .

Even in some relatively simple enrichment designs, such as those we present in § 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. We apply this theorem to the class of enrichment designs in § 3; by solving the corresponding optimization problem, we obtain the asymptotic, worst-case, familywise Type I error of these enrichment designs. To the best of our knowledge, our general method gives the first means of precisely computing this quantity.

Throughout this paper, we consider only 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.

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](#); [Russek-Cohen & Simon, 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 § 3, these methods either do not apply or our method gives improved results. Below, we describe related work especially relevant to the class of designs that we present in § 3.

[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 2, we select which population to continue enrolling in Stage 2. The Type I error guarantees of [Thall et al. \(1988\)](#) do not directly carry over to our setting, in which it is important to be able to enroll from multiple subpopulations in Stage 2, since this is not allowed in the analogous designs of [Thall et al. \(1988\)](#); this makes sense, since their goal was selecting the single best treatment.

[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 of [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 § 3, however, gives a multiple testing procedure with strictly lower power than the one resulting from application of our general method from § 5.

[Follmann \(1997\)](#) proves control of Type I error for a large class of useful enrichment designs. This class allows a much wider range of decision rules than the ones we consider in § 3. However, in the designs of [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.

[Russek-Cohen & Simon \(1997\)](#) present a novel two-stage adaptive design that incorporates a test for a subpopulation by treatment interaction, which is used to determine the sample size in the second stage. Their designs have good power in many scenarios, though not when using one-sided tests as done here. A difference between their work and this paper is that they do not prove strong control of familywise Type I error for their designs, which is a main contribution of this paper for the designs we present.

[Wang et al. \(2007, 2009\)](#) propose important enrichment designs using decision rules similar to those given in § 3. However, in the designs of [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 set-up in Wang et al. (2009) differs from ours in that they assume that 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.

3. A CLASS OF ENRICHMENT DESIGNS

3.1. Overview

We present a class of two-stage enrichment designs tailored to the problem described in § 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 apply our general method to precisely compute 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 fixed designs of the same total sample size, in terms of overall power and ability to determine subpopulation-specific treatment effects. In § 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, while they are superior to placebo for those with severe pretreatment depression. 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 score is recorded at baseline and after a set period of time. We refer to a subject's Hamilton Rating Scale of Depression scores simply as his/her scores. Lower scores indicate lower severity of depression. For each subject, we define his/her improvement to be baseline score minus end of study score. If a subject's 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 for subpopulation 1, that the mean improvement under the new treatment is less than or equal to the mean improvement under the placebo. In an analogous manner, define the null hypothesis H_{02} corresponding to subpopulation 2, and the null hypothesis H_{00} corresponding to the total population. We emphasize that H_{00} is not the intersection of the null hypotheses H_{01} and H_{02} ; it is the null hypothesis corresponding to the total population, which is the union of the two subpopulations. The alternative hypotheses to H_{00} , H_{01} , H_{02} are that the mean improvement under treatment is greater than under placebo, for the total population, for subpopulation 1, and for subpopulation 2, 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 treatment arm or the control arm with 50% chance of being assigned to each. For simplicity, we assume that exactly 50% of participants in each subpopulation are assigned to each arm, which can be approximately ensured using stratified block randomization. Also, for simplicity, we assume the total number of subjects to be enrolled in Stage 1 and in Stage 2 is prespecified and cannot be changed. At the end of Stage 1, we compute three z -statistics, $T_0^{(1)}$, $T_1^{(1)}$, $T_2^{(1)}$, corresponding to the standardized difference in mean improvement between treatment and control groups for subjects in the total population, in subpopulation 1 and in subpopulation 2, respectively; let $T_3^{(1)}$ denote a prespecified vector

Generic Enrollment Procedure for Enrichment Design

Fig. 1. A flow chart depicting a type of enrichment design within our class, where enrollment in Stage 2 can be restricted to the subpopulation with the larger z -statistic based on data from Stage 1. Each labelled box represents a set of patients enrolled from a certain subpopulation.

of additional Stage 1 statistics. These are defined formally below. We assume that data on all subjects from Stage 1 are available at the interim analysis.

We allow a decision to be made at the end of Stage 1, based on a prespecified decision rule. We denote this decision rule, which maps the data collected in Stage 1 to a choice for which subpopulation to enroll in Stage 2, by D . We consider the class of decision rules, denoted by $\mathcal{D}$, that are Borel measurable maps from the vector of Stage 1 statistics $(T_0^{(1)}, T_1^{(1)}, T_2^{(1)}, T_3^{(1)})$ to the following two choices for the subpopulation(s) to enroll in Stage 2:

- (i) continue enrolling from both subpopulations in the same proportions as in Stage 1 or
- (ii) enroll only from the subpopulation $s \in \{1, 2\}$ corresponding to the larger of the Stage 1 z -statistics $T_1^{(1)}, T_2^{(1)}$. Any ties can be decided by enrolling from, say, subpopulation 2.

One example of a decision rule in the class $\mathcal{D}$ is to always choose option (ii), that is, to always sample in Stage 2 only from the subpopulation with the larger z -statistic from Stage 1. Another example of a decision rule in the class $\mathcal{D}$ is to use option (i) if the z -statistic $T_0^{(1)}$ 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. Another decision rule, suggested by a reviewer, is to base the decision between options (i) and (ii) on a prespecified test for interaction; if the test rejects, one uses option (ii), and otherwise one uses option (i); this has some similarities with the design of [Russek-Cohen & Simon \(1997\)](#). The standard fixed design, which corresponds to always choosing option (i), is also a decision rule in the class $\mathcal{D}$.

At the end of the trial, a preplanned final test statistic is computed leading to a possible rejection of one of the null hypotheses $\{H_{00}, H_{01}, H_{02}\}$. This final test statistic is a weighted combination of the z -statistic $T_0^{(1)}$, which pools all Stage 1 data from both subpopulations, and the z -statistic from the second stage, which pools all second stage data. The corresponding weights are defined below. 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 2.

3.2. Formal definition of statistics, assumptions and the testing procedure

We now precisely define the statistics used in the above decision rules and to test the family of null hypotheses $\{H_{00}, H_{01}, H_{02}\}$. We denote the data for subject m by (I_m, S_m, A_m, Y_m) ; this includes the stage the subject entered the trial $I_m \in \{1, 2\}$, and the subject's subpopulation $S_m \in \{1, 2\}$, indicator of assignment to the treatment arm $A_m \in \{0, 1\}$ and outcome $Y_m \in \mathbb{R}$, respectively.

We assume in each stage, for each subpopulation sampled in that stage, that an equal number is assigned to each study arm. We 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 the sample size, such that the support of Q_{sa} is contained in $[-M, M]$ 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 Stage 1, $n_1/(n_1 + n_2)$, be between 0.05 and 0.95. We also make an assumption about the fraction of Stage 1 subjects who are from each subpopulation $s \in \{1, 2\}$, which we denote p_s . We consider each p_s nonrandom and possibly unknown, since in practice one may not know in advance how many subjects will enroll from each subpopulation. However, we assume there is a nonnegligible fraction of subjects in Stage 1 from each subpopulation; that is, we assume for some $\tau > 0$ not depending on the sample size, that for each s , $p_s > \tau$. We also assume that the proportion p_s of stage one subjects in each subpopulation s equals the corresponding population proportion.

We let $\mu(Q_{sa})$ denote the mean of the distribution Q_{sa} . The null hypotheses to be tested are:

$$\begin{aligned} H_{00} : p_1 \{\mu(Q_{11}) - \mu(Q_{10})\} + (1 - p_1) \{\mu(Q_{21}) - \mu(Q_{20})\} &\leq 0, \\ H_{01} : \mu(Q_{11}) - \mu(Q_{10}) &\leq 0, \quad H_{02} : \mu(Q_{21}) - \mu(Q_{20}) &\leq 0. \end{aligned}$$

Define the Stage 1 z -statistics for the total population, for subpopulation 1, and for subpopulation 2, respectively, as

$$T_0^{(1)} = \sum_{m:I_m=1} \{Y_m A_m - Y_m(1 - A_m)\} / \text{se}_0, \quad (1)$$

$$T_1^{(1)} = \sum_{m:I_m=1, S_m=1} \{Y_m A_m - Y_m(1 - A_m)\} / \text{se}_1, \quad (2)$$

$$T_2^{(1)} = \sum_{m:I_m=1, S_m=2} \{Y_m A_m - Y_m(1 - A_m)\} / \text{se}_2, \quad (3)$$

where

$$\text{se}_1 = \left[\frac{n_1 p_1}{2} \left\{ \sigma^2(Q_{10}) + \sigma^2(Q_{11}) \right\} \right]^{1/2}, \quad \text{se}_2 = \left[\frac{n_1 p_2}{2} \left\{ \sigma^2(Q_{20}) + \sigma^2(Q_{21}) \right\} \right]^{1/2}, \quad (4)$$

$$\text{se}_0 = \left(\text{se}_1^2 + \text{se}_2^2 \right)^{1/2}. \quad (5)$$

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 each of the variances $\sigma^2(Q_{sa})$ is 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.

Let $T_3^{(1)}$ denote an additional, prespecified vector of Stage 1 statistics. We put no constraints on $T_3^{(1)}$ except that it is a prespecified, Borel measurable map from Stage 1 data to $\mathbb{R}^k$.

We now describe the final test statistic and the rejection region for each null hypothesis. After Stage 2 is completed, we compute the following z-statistic based on all the data from that stage:

$$T^{(2)} = \sum_{m:I_m=2} \{Y_m A_m - Y_m(1 - A_m)\} / \text{se}_3, \quad (6)$$

where se_3 is the standard error for the corresponding sum, conditioned on the enrollment decision made just after Stage 1, defined analogously as (4) and (5); we give se_3 in § C of the Supplementary Material.

We reject the null hypothesis corresponding to the subpopulation, or the total population, enrolled in Stage 2, if the following weighted combination of z-statistics from both stages,

$$T_{\text{final}} = \{n_1/(n_1 + n_2)\}^{1/2} T_0^{(1)} + \{n_2/(n_1 + n_2)\}^{1/2} T^{(2)}, \quad (7)$$

exceeds some prespecified threshold c . That is, if both subpopulations are enrolled in stage two and (7) exceeds c , we reject H_{00} ; if only subpopulation 1 is enrolled in stage two and (7) exceeds c , we reject H_{01} ; if only subpopulation 2 is enrolled in stage two and (7) exceeds c , we reject H_{02} .

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

3.3. Application of the general method from § 5

The general method we present in § 5 allows one to compute, for each enrichment design in the class defined above, the minimum threshold value c for the final test statistic (7) that guarantees strong control of the asymptotic, familywise Type I error at a given level α . This is useful since even for very simple enrichment designs in this class, 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 § 5 and §§ B–E of the Supplementary Material, that under the above assumptions, for any decision rule $D \in \mathcal{D}$, 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, this class allows flexibility at no price in terms of a stricter rejection threshold. Furthermore, in § 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 with 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 Stage 1 data from both subpopulations is used in the final test statistic (7), even when the final hypothesis tested concerns only one of these subpopulations. Although 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, with an exception being the conservative Bonferroni procedure; 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 has the larger corresponding z -statistic in Stage 1, 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 1 data for subpopulation 1, this would introduce a selection bias due to essentially hiding data that was unfavourable. 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 the designs of [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 §§ 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 with a prespecified penalty.

4. POWER COMPARISON BETWEEN THE ADAPTIVE DESIGN FROM § 3 AND FIXED DESIGNS

We show in § 5 and §§ B–E of the Supplementary Material that for each decision rule in the class $\mathcal{D}$ 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 (7) 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 $\mathcal{D}$, and compare its power with that of two fixed designs under several scenarios. We continue to use the example introduced in § 3 of testing a hypothetical new antidepressant, motivated by the meta-analysis of [Kirsch et al. \(2008\)](#). We let subpopulation 1 denote those with moderate pretreatment depression, and subpopulation 2 denote those with severe pretreatment depression.

To specify a particular adaptive design from the class in the previous section, we need to choose a decision rule $D \in \mathcal{D}$. Recall that each $D \in \mathcal{D}$ is a function from the Stage 1 statistics to two possible choices for Stage 2 enrollment: (i) enroll from both subpopulations as in Stage 1 or (ii) enroll only from the subpopulation corresponding to the larger of the Stage 1 z -statistics $T_1^{(1)}$, $T_2^{(1)}$. We define our decision rule, which we denote by $\tilde{D}$, to choose option (i) if the z -statistic $T_1^{(1)} > T_2^{(1)}$ or $T_1^{(1)} > 0.3$, and option (ii) otherwise.

The intuition behind our choice of decision rule is that if $T_1^{(1)} > T_2^{(1)}$ or $T_1^{(1)} > 0.3$, then there is at least some hope that the treatment is effective for subpopulation 1, i.e., for those with moderate pretreatment depression. In this case, we continue enrolling them, as well as those from subpopulation 2, in Stage 2. 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 0.3.

Since the above design is in the class $\mathcal{D}$, the asymptotic, familywise Type I error is strongly controlled, as described above. Consider replacing the threshold 0.3 above by any other prespecified threshold. Since the resulting design remains in the class $\mathcal{D}$, the same 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, which is how we selected the threshold 0.3.

In all our simulations, we use a slightly improved version of the adaptive design just described. The improved version incorporates a fixed sequence testing procedure as in [Maurer et al. \(1995\)](#), which allows it to simultaneously reject H_{00} and H_{02} in some cases, similar to the fixed designs

we will compare it to, which we describe next. Full details are given in § F of the Supplementary Material, where we show that the asymptotic, worst-case, familywise Type I error of this improved adaptive procedure is also 0.05.

We compare the power of this adaptive design to two fixed designs, each of which uses a different multiple testing procedure. In each fixed design, the proportion of subjects enrolled from subpopulation 1 is the same in both stages. Also in each fixed design, at the end of the trial, a p -value corresponding to H_{00} is computed based on a one-sided z -test pooling all the data, comparing treatment to control; a p -value corresponding to H_{02} is computed based on a one-sided z -test using all subpopulation 2 data. The first of the multiple testing procedures is a fixed sequence procedure as in Maurer et al. (1995). It rejects H_{00} whenever the p -value corresponding to H_{00} is less than 0.05; if, in addition, the p -value corresponding to H_{02} is less than 0.05, the null hypothesis H_{02} is also rejected. We refer to this design as the fixed sequence design. The second multiple testing procedure is a step-up procedure (Hochberg, 1988). It rejects both null hypotheses H_{00} and H_{02} if both of the corresponding p -values are less than 0.05; otherwise, if the smallest of these two p -values is less than 0.025, the corresponding null hypothesis is rejected. It is straightforward to show for these two fixed designs, that the asymptotic, worst-case, familywise Type I error is at most 0.05, under the assumptions in § 3.2.

Recall we defined the standard fixed design to not allow changes to the population enrolled, and to use a one-sided z -test for H_{00} . By construction, the power of the fixed sequence design dominates the power of the standard fixed design; i.e., for any data-generating distribution, the probability the fixed sequence design rejects H_{00} equals that of the standard fixed design, while the fixed sequence design may reject H_{02} as well. Since for both these designs the asymptotic, worst-case, familywise Type I error is 0.05, we chose to use the fixed sequence design in our simulations, as it is the more powerful of the two designs.

In § H of the Supplementary Material we compare the power of the above adaptive design to the above two fixed designs, under a variety of scenarios. We present six such scenarios here, which we number 1a, 1b, 1c, 2a, 2b, 2c. In all scenarios and all designs, the total sample size is the same: 488 subjects. In scenarios 1a, 1b and 1c, we assume that half of the subjects enrolled in Stage 1 are from subpopulation 1, and the other half are from subpopulation 2; we also assume in these scenarios that the total sample size in each stage is 244 subjects. In scenarios 2a, 2b and 2c, we assume that 75% of the subjects enrolled in Stage 1 are from subpopulation 1, and the other 25% are from subpopulation 2; we also assume in scenarios 2a, 2b and 2c that the sample size in stage one is 30% of the sample size of the whole trial.

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 with placebo in those with moderate initial depression, but a positive mean improvement comparing treatment versus placebo in those with severe initial depression. In scenario 1a, this mean improvement is set to 1.8 points, which is the point estimate of this quantity in the meta-analysis of Kirsch et al. (2008); in scenario 1b, the mean improvement is set to the more optimistic value of 3 points, which was seen in the meta-analysis of Kirsch et al. (2008) for subjects with very severe initial depression. In scenario 1c, we use data-generating distributions where the treatment is equally effective for both subpopulations, with the mean improvement in treatment versus placebo set to 1.8 points. Scenarios 2a, 2b and 2c use exactly the same data-generating distributions as the corresponding scenarios 1a, 1b and 1c, respectively; the only differences between these sets of scenarios, as described above, are the proportion enrolled in Stage 1 from subpopulation 1, and the proportion of the total sample size allocated to each stage.

Table 1. *Power comparison of fixed and adaptive designs*

	Scenario 1a				Scenario 1b				Scenario 1c			
	Both	H_{00}	H_{02}	Total	Both	H_{00}	H_{02}	Total	Both	H_{00}	H_{02}	Total
Fixed sequence	30	5	0	34	65	1	0	67	52	28	0	80
Fixed Hochberg	30	2	16	48	65	1	20	86	52	21	1	74
Adaptive	16	5	28	49	31	1	55	87	43	28	9	80
	Scenario 2a				Scenario 2b				Scenario 2c			
	Both	H_{00}	H_{02}	Total	Both	H_{00}	H_{02}	Total	Both	H_{00}	H_{02}	Total
Fixed sequence	10	5	0	15	24	3	0	27	32	48	0	80
Fixed Hochberg	10	3	16	28	24	2	33	59	32	40	1	73
Adaptive	6	5	27	38	14	3	53	70	26	45	9	80

For each of six scenarios, we compare the power of the fixed sequence design, the fixed design using the Hochberg procedure and the above adaptive design. Each cell gives the approximate probability, represented as a per cent, of rejecting the null hypothesis or hypotheses indicated in the column title. The column ‘Both’ gives the probability of simultaneously rejecting H_{00} and H_{02} , the column ‘ H_{00} ’ gives the probability of rejecting only H_{00} , the column ‘ H_{02} ’ gives the probability of only rejecting H_{02} and the column ‘Total,’ which represents overall power, is the sum of the preceding three columns. Some numbers in the ‘Total’ columns appear to be off by 1, due to rounding after summing the original unrounded values.

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. \(2008\)](#). The total sample size in each design was set to be 488 subjects, in order to provide 80% power to the fixed sequence design to reject H_{00} under scenario 1c. For each scenario and design, 100 000 simulated trials were run. See § G of the Supplementary Material for full details and the R code used.

For each scenario and design, the per cent of simulated trials leading to rejection of each null hypothesis among H_{00} , H_{02} was recorded; the results are reported in Table 1, rounded to the nearest per cent. By construction, none of the designs ever rejects the null hypothesis H_{01} , so it is not considered here. We call the probability of rejecting at least one of H_{00} and H_{02} , when both are false, the overall power. Since in all the scenarios depicted in Table 1 both of these null hypotheses are false, the per cent reported in the ‘Total’ column equals the overall power of the corresponding design and scenario.

We summarize the tradeoffs in using each of the three designs in the scenarios considered above. In each of the above scenarios, the overall power of the adaptive design was at least as large as that of the other designs. Compared with the fixed sequence design, the improvement in overall power from using the adaptive design ranged from 0 to 43%, depending on the scenario. One drawback of the adaptive design, however, occurs in the scenarios where the treatment is equally effective for both subpopulations, i.e., 1c and 2c, where it is preferable to reject the null hypothesis for the total population H_{00} ; though the overall power of the adaptive and the fixed sequence design are equal in these scenarios, the adaptive design rejects H_{00} 71% of the time, while the fixed sequence design rejects H_{00} 80% of the time. These values are obtained by adding the corresponding cells in the ‘Both’ and ‘ H_{00} ’ columns of the table.

Comparing the adaptive design to the fixed design using the Hochberg procedure, the overall power is similar in scenarios 1a and 1b, but the adaptive design has more overall power in all the other scenarios, with the maximum overall power difference being 12%. Compared with the fixed sequence design, the fixed design using the Hochberg procedure has a similar drawback as the adaptive design in the scenarios where the treatment is equally effective for both subpopulations. The adaptive design and the fixed design using the Hochberg procedure have an advantage over the fixed sequence design in scenarios 1a, 1b, 2a and 2b, in that they have considerably more

power to reject the null hypothesis H_{02} ; this is arguably the most important null hypothesis to reject in these scenarios, since in these scenarios the treatment only benefits subpopulation 2.

In addition to scenarios presented in § H of the Supplementary Material, we consider situations where the treatment is harmful for one subpopulation and beneficial to the other. In such scenarios, the adaptive design and fixed design using the Hochberg procedure have more overall power than the fixed sequence design. In some of these scenarios, the fixed design using the Hochberg procedure has more overall power than the adaptive design.

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 an adaptive design to a more 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.

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 also be generalized to apply to designs that incorporate adaptations in addition to changing the population sampled, but we do not present such designs here. The main challenge in incorporating additional stages and different types of adaptations is the computational difficulty of the resulting optimization problem. It is an area of future research to determine in which cases such generalizations are computationally feasible.

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 that there is a positive constant r , independent of the 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 $T^{(i)}$ of statistics taking values in $\mathbb{R}^{t_i}$, for some t_i . Each $T^{(i)}$ 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 2, each of which specifies the proportion of Stage 2 subjects to enroll from every subpopulation. We assume, for each enrollment procedure in $\mathcal{E}$, that one can enroll in Stage 2 to exactly achieve these proportions;
- (6) a decision function D that maps each value of the vector of Stage 1 statistics $T^{(1)}$ to an element in the set $\mathcal{E}$ of potential enrollment proportions for each subpopulation for Stage 2. 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 $T^{(1)}$ leading to this choice is a finite union of Borel measurable, convex subsets of $\mathbb{R}^{t_1}$;
- (7) a set of null hypotheses $\{H_{0j}\}_{j \in J}$ 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 $\{R_j\}_{j \in J}$, 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 $(T^{(1)}, T^{(2)}) \in R_j$.

5.3. Data-generating distributions and the 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 that conditioned on each subject's subpopulation and study arm, his/her outcome Y 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 $\mathcal{Q}' = \{Q_{sa}\}_{a \in \mathcal{A}, s \in \mathcal{S}}$. We assume that $\mathcal{Q}'$ is an element of a statistical model $\mathcal{Q}'$. This may be a nonparametric model, such as the one specified in § 3.2 that puts no constraints on each Q_{sa} except for having bounded support and a minimum variance. It may also be a semiparametric model or a parametric model.

Denote the fraction of Stage 1 subjects who are from subpopulation $s \in \mathcal{S}$ by p_s , which we consider nonrandom and possibly unknown at the start of the trial. We denote the set of subpopulation proportions enrolled in Stage 1 by $p = \{p_s\}_{s \in \mathcal{S}}$. Let $\mathcal{P}$ 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 $T = (T^{(1)}, T^{(2)})$, which we denote by $P^{(n)}$, is a function of the data-generating distributions $\mathcal{Q}'$, the Stage 1 subpopulation proportions p , and the sample size n . Implicitly, $P^{(n)}$ also depends on the prespecified quantities listed in § 5.2, including the set of subpopulations $\mathcal{S}$, the set of treatments $\mathcal{A}$, the randomization procedure, the possible enrollment procedures for Stage 2 $\mathcal{E}$, the definition of the statistics T and the decision rule D .

To simplify notation in what follows, we bundle $\mathcal{Q}'$ and p together and denote the pair $(\mathcal{Q}', p)$ by $\mathcal{Q}$. We let $\mathcal{Q}$ denote $\mathcal{Q}' \times \mathcal{P}$, which represents the class of all possible data-generating distributions and Stage 1 subpopulation enrollment proportions. We refer to each $\mathcal{Q} \in \mathcal{Q}$ 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, which is defined as

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

where $n \rightarrow \infty$ is shorthand for $n_1, n_2 \rightarrow \infty$, under the constraint given in § 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 .

Designs in which the 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. Such uniform guarantees are especially important for enrichment designs, since the distributions of primary interest are those for which there is considerable uncertainty about which subpopulations benefit most from treatment. Since this set of relevant distributions changes depending on the sample size, as discussed in § C.4 of the Supplementary Material, some kind of uniform asymptotics is needed to give interpretable Type I error guarantees.

5.5. Assumptions on statistics

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

The second assumption is that centred versions of the statistics converge uniformly to a multivariate normal distribution, in a sense described below. We first introduce additional notation. For any $Q \in \mathcal{Q}$ and sample size n , denote the mean of the Stage 1 statistics $T^{(1)}$ by $\mu_n^{(1)}(Q)$, and denote the distribution of the centred Stage 1 statistics $T^{(1)} - \mu_n^{(1)}(Q)$, 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 $n \rightarrow \infty$.

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

We furthermore assume the above convergence of centred Stage 1 and Stage 2 statistics is uniform over data-generating distributions $Q \in \mathcal{Q}$. That is, we assume that for all possible enrollment decisions $\epsilon \in \mathcal{E}$,

$$\lim_{n \rightarrow \infty} \sup_{Q \in \mathcal{Q}, C \in \mathcal{C}} \left| \int_C d \left\{ G_n^{(1)}(Q) \times G_n^{(2)}(Q, \epsilon) \right\} - \int_C d \left\{ G^{(1)}(Q) \times G^{(2)}(Q, \epsilon) \right\} \right| = 0, \quad (8)$$

where $\mathcal{C}$ denotes the set of all Borel measurable, convex subsets of $\mathbb{R}^{t_1+t_2}$. For the class of designs from § 3, we prove in § C.6 of the Supplementary Material that this assumption holds, using a generalized Berry–Esseen theorem of Götze (1991).

5.6. Intuition behind the theorem

Before giving Theorem 1, we present some of the intuition that underlies it. Consider the case where data are generated according to a distribution $Q \in \mathcal{Q}$. A familywise Type I error occurs, that is, at least one true null hypothesis is rejected, if the vector of statistics T is in any of the rejection regions R_j for which the corresponding null hypothesis H_{0j} is true under Q . More concisely, a familywise Type I error occurs if and only if $T \in \bigcup_{j: Q \in H_{0j}} R_j$. We partition this event into more manageable pieces by intersecting it, for each possible Stage 2 enrollment procedure $\epsilon \in \mathcal{E}$, with the event that our decision rule D says to enroll using ϵ . Define each such intersection by

$$A_\epsilon = \left\{ T^{(1)} \in D^{-1}(\epsilon), (T^{(1)}, T^{(2)}) \in \bigcup_{j: Q \in H_{0j}} R_j \right\}. \quad (9)$$

By construction, on each event A_ϵ , the enrollment decision $D(T^{(1)})$ after Stage 1 is always ϵ . It follows that the probability of A_ϵ when T is generated by the adaptive design is the same as

the probability of A_ϵ were T instead generated by the fixed design that always makes enrollment decision ϵ regardless of the Stage 1 data. This is proved in § B of the Supplementary Material. 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 computing this for several fixed designs, which immensely simplifies our computations.

Combining the above with the assumption from § 5.5 that the centred statistics converge to a multivariate normal, we succeed in reducing the original problem to one of computing the supremum, over a class of easy to describe regions, of the probability that a zero mean, multivariate normal distribution is in each region. Each such region corresponds to that used in defining the event A_ϵ in (9), except that we centre the regions; we define the centred versions of these regions, for each $Q \in \mathcal{Q}$, sample size n , and possible enrollment decision $\epsilon \in \mathcal{E}$, by

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

where for any set $B \subseteq \mathbb{R}^m$, and vector $v \in \mathbb{R}^m$, we define the shifted version of B as $B - v = \{x \in \mathbb{R}^m : x + v \in B\}$. The sets $V(Q, n, \epsilon)$ figure prominently in the theorem below.

5.7. Equivalence of the asymptotic, worst-case, familywise Type I error and the solution of an optimization problem

THEOREM 1. *For any two stage design that adheres to the specifications in § 5.2, under the assumptions given in § 5.3 and 5.5, we have*

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

which implies that the asymptotic, worst-case, familywise Type I error defined in § 5.4 equals

$$\lim_{n \rightarrow \infty} \sup_{Q \in \mathcal{Q}} \sum_{\epsilon \in \mathcal{E}} \int_{V(Q, n, \epsilon)} d \left\{ G^{(1)}(Q) \times G^{(2)}(Q, \epsilon) \right\}. \quad (10)$$

Computing the supremum of the sum in the previous display over the class of possible data-generating distributions $\mathcal{Q}$ is an optimization problem. In the cases we consider, the supremum of the sum in the above display over $Q \in \mathcal{Q}$ is the same for all n , and so solving the corresponding optimization problem at a single n suffices; we conjecture this will be true for a wide variety of designs. 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 (10) to finding the supremum over the set of possible values of these scalar parameters. Often, evaluating the integral in (10) at a particular Q , n , and ϵ can be done with statistical software for computing the multivariate normal distribution function.

In §§ C–E of the Supplementary Material, we apply Theorem 1 to the class of enrichment designs in § 3, to show each of these designs strongly controls asymptotic, familywise Type I error at level 0.05. An outline of the main ideas behind how this theorem is applied is given in § C.3 of the Supplementary Material.

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 is only guaranteed to have correct Type I error for testing null hypotheses such as those in § 3, in the limit as sample size goes to infinity, unless parametric or other assumptions are made on the data-generating distribution.

We did simulations to determine the familywise Type I error for the adaptive design from § 4 at sample sizes ranging from 40 subjects to 2560 subjects; see § I of the Supplementary Material. In all simulations, when the nominal level was 0.05, the largest familywise Type I error for the adaptive design from § 4 was 0.0517. When the response distributions Q_{sa} in § 3 are each Gaussian, one can show the familywise Type I error for any of the adaptive designs from that section is at most 0.05 at any sample size, not just asymptotically.

Our general method depends on the ability to accurately compute values of a multivariate normal distribution function with statistical software. Our result for the class of designs in § 3 only required computing values of a bivariate normal distribution function with accuracy to four decimal places. This and similar results using the above general method are subject to the assumption that the software used for such computations has the desired accuracy. In general, the required accuracy will depend on the complexity of the design.

An important area of future work is to determine what fraction of the total sample size to enroll before making a preplanned change in the enrollment procedure, in order to maximize power under certain alternatives of interest. A related issue is that if a subpopulation makes up only a small fraction of the total population, the trial duration may be long if one enrolls only from it in Stage 2, due to difficulty in finding enough such subjects. Other important open questions are to determine how external information might be incorporated into a decision rule, and how adaptive procedures perform when outcome distributions change over time.

An area of future work related to the above adaptive designs is to determine the asymptotic, worst-case bias and mean squared error of maximum likelihood estimators, as well as to construct confidence intervals that are guaranteed to be asymptotically conservative. Important related work in this area includes, e.g., Jennison & Turnbull (1984), Proschan & Hunsberger (1995), Lehmacher & Wassmer (1999), Posch et al. (2005), Brannath et al. (2006) and Wu et al. (2010). An additional area of future work is to extend our bounds on familywise Type I error for the designs from § 3 to analogous designs involving more than two subpopulations.

ACKNOWLEDGEMENT

This research was supported by the National Institutes of Health, U.S.A., and by the U.S. Food and Drug Administration's Partnership in Applied Comparative Effectiveness Science. Opinions, ideas and statements in this paper are presented at the sole discretion of the authors and do not represent those of the National Institutes of Health or the Food and Drug Administration.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains the proof of the theorem from § 5.7, and a demonstration that the asymptotic, worst-case, familywise Type I error for the class of decision rules in § 3 is 0.05. We also give the details of the simulation study from § 4 and simulations that explore Type I error at various sample sizes.

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**, 1029–41.
- 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. *Statist. Med.* **25**, 3366–81.
- 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. *Statist. Sin.* **7**, 1085–102.
- FOLLMANN, D. A., PROSCHAN, M. A. & GELLER, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**, 325–36.
- GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Prob.* **19**, 724–39.
- HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–2.
- HOCHBERG, Y. & TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. Probability and Mathematical Statistics. New York: 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. *Biomet. J.* **43**, 581–9.
- JENNISON, C. & TURNBULL, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Contr. Clin. Trials* **5**, 33–45.
- JENNISON, C. & TURNBULL, B. W. (2006). Discussion: confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. *Biomet. J.* **48**, 650–5.
- JENNISON, C. & TURNBULL, B. W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharm. Statist.* **17**, 1135–61.
- KIESER, M., BAUER, P. & LEHMACHER, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biomet. J.* **41**, 261–77.
- 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**, 1286–90.
- MARCUS, R., PERITZ, E. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–60.
- MAURER, W., HOTHORN, L. A. & LEHMACHER, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der chemische-pharmazeutischen Industrie*, Ed. J. Vollman, vol. 6. Stuttgart: Fischer.
- POSCH, M., KOENIG, F., BRANSON, M., BRANNATH, W., DUNGER-BALDAUF, C. & BAUER, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statist. Med.* **24**, 3697–714.
- PROSCHAN, M. & HUNSBERGER, S. (1995). Designed extension studies based on conditional power. *Biometrics* **51**, 1315–24.
- RUSSEK-COHEN, E. & SIMON, R. M. (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statist. Med.* **16**, 455–64.
- SAMPSON, A. R. & SILL, M. W. (2005). Drop-the-losers design: Normal case. *Biomet. J.* **47**, 257–68.
- SCHAID, D. J., WIEAND, S. & THERNEAU, T. M. (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika* **77**, 507–13.
- STALLARD, N. & TODD, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statist. Med.* **22**, 689–703.
- THALL, P. F., SIMON, R. & ELLENBERG, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–10.
- WANG, S.-J., HUNG, H. & O'NEILL, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biomet. J.* **51**, 358–74.
- WANG, S. J., O'NEILL, R. T. & HUNG, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist.* **6**, 227–44.
- WU, S. S., WANG, W. & YANG, M. C. K. (2010). Interval estimation for drop-the-losers designs. *Biometrika* **97**, 405–18.

[Received December 2010. Revised July 2011]