9-20-2017

OPTIMIZED ADAPTIVE ENRICHMENT DESIGNS FOR MULTI-ARM TRIALS: LEARNING WHICH SUBPOPULATIONS BENEFIT FROM DIFFERENT TREATMENTS

Jon Arni Steingrimsson
Department of Biostatistics, Brown School of Public Health

Joshua Betz
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Tiachen Qian
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, mrosen@jhu.edu

Optimized Adaptive Enrichment Designs for Multi-Arm Trials: Learning which Subpopulations Benefit from Different Treatments

Jon Arni Steingrimsson, Joshua Betz, Tianchen Qian, and Michael Rosenblum

September 20, 2017

1 Abstract

We propose a class of adaptive randomized trial designs for comparing two treatments to a common control in two disjoint subpopulations. The type of adaptation, called adaptive enrichment, involves a preplanned rule for modifying enrollment and arm assignment based on accruing data in an ongoing trial. The motivation for this adaptive feature is that interim data may indicate that a subpopulation, such as those with lower disease severity at baseline, are unlikely to benefit from a particular treatment, while uncertainty remains for the other treatment and/or subpopulation. We developed a new multiple testing procedure tailored to this design problem. The procedure improves power by: leveraging the correlation between the test statistics arising from the two treatments being compared to a common control; reallocating alpha across subpopulations, and using the data only through minimally sufficient statistics. We optimize expected sample size over this class of designs, focusing on designs with 2 stages. Our approach is demonstrated in simulation studies that mimic features of a completed trial of a medical device for treating heart failure. User-friendly, open-source software that implements the trial design optimization is provided.

Keywords: Randomized Clinical Trial, Treatment Effect Heterogeneity

2 Introduction

Our trial design problem is related to the SMART-AV trial (Ellenbogen et al., 2010), a phase 4 randomized trial of patients with medically-refractive heart failure with severe left ventricular systolic dysfunction. All the participants had an implanted cardiac resynchronization therapy defibrillator. The trial aimed to investigate the effect of optimizing the atrioventricular (AV) delay in this medical device. Two methods of optimizing the atrioventricular delay (called treatments) were compared to a fixed delay of 120 milliseconds (called control). No statistically significant differences were found between the treatments and control, for the primary outcome of left ventricular end-systolic volume.

Previous scientific knowledge had indicated that participants with short QRS duration, defined as QRS $\leq$ 150 milliseconds, may be more likely to benefit from the treatments (Stein et al., 2010). This raises the question of whether a design targeted to identify treatment effects in subpopulations defined by QRS duration could have been more informative. To address this question, we develop and evaluate a new class of adaptive enrichment designs comparing two treatments to a common control in two disjoint subpopulations. Adaptive enrichment designs have a preplanned rule for modifying enrollment criteria based on accruing data in an ongoing trial (Wang et al., 2009). These designs can stop accrual of some treatment by subpopulation combinations for either efficacy or futility at the end of each stage. We compare the performance of these adaptive designs versus standard designs in determining which subpopulation by treatment combinations lead to improved outcomes.

Our proposed class of adaptive enrichment designs uses a new multiple testing procedure that combines features from (Dunnett, 1955) that involve taking the maximum of normally distributed statistics and from graphical approaches (Bretz et al., 2009, 2011; Maurer and Bretz, 2013) that reallocate alpha from null hypotheses that are rejected to the remaining null hypotheses in group sequential designs.

The proposed adaptive designs have the following properties: they leverage the correlation between the test statistics arising from that both treatments are compared to a common control; they improve efficiency by lowering the rejection threshold for the remaining null hypotheses after a null hypothesis has been rejected; they can stop arms early for efficacy or futility; they allow for continuation of remaining treatments/subpopulations after some null hypotheses are rejected in order to continue testing the remaining null hypotheses; they strongly control the familywise Type I error rate, asymptotically; the multiple testing procedure is a function of only minimally sufficient statistics and is therefore exempt from a criticism of some adaptive design approaches (Emerson, 2006) that not using minimally sufficient statistics can lead to inefficiency.

We optimize the multiple testing procedure and enrollment modification rule in order to minimize expected sample size while satisfying power and Type I error constraints. As there is no known optimization procedure that is guaranteed to converge to the global optimum, we use simulated annealing, a general purpose optimization method. Fisher and Rosenblum (2016) used simulated annealing to optimize over a class of designs evaluating the effectiveness of a single treatment in two subpopulations. Our setting differs as follows: it involves two treatments versus control, a different class of adaptive designs, a different set of null hypotheses, and a more complex set of power requirements.

Others have proposed adaptive designs for multi-arm trials for the case of a single subpopulation. Magirr et al. (2012) proposed a Dunnett test for such trials; Wason and Jaki (2012) used simulated annealing to search for the efficacy boundaries that minimize expected sample size for a multi-stage multi-arm trials. In both these references, the trial is stopped when the first null hypothesis is rejected. Several designs, i.e. Kelly et al. (2005); Thall et al. (1988); Whitehead and Jaki (2009), have been proposed that pick only one treatment at the interim analysis to continue on to the later stages; in contrast, our designs allow continuation of any number of arms. Koenig et al. (2008) proposed an adaptive Dunnett

test using the conditional error function approach; the interim analysis in their designs is used for treatment selection and no null hypothesis can be rejected for efficacy at an interim analysis. Stallard and Friede (2008) allow for rejection at an interim analysis but control of Type 1 error requires the number of treatments evaluated at each stage to be prespecified. Both Posch et al. (2005) and Bretz et al. (2010) propose flexible adaptive designs based on p-value combinations that do not use data only through minimally sufficient statistics.

Urach and Posch (2016) consider multi-arm, group sequential designs with both simultaneous and separate stopping rules. For separate stopping rules, treatment arms can be stopped for efficacy at interim analysis and the other arms are allowed to continue to the next stage. The design differs from ours in that it uses a different form of efficacy boundaries, it only considers a single population, and it is restricted to two stages and immediately observed outcomes. The search space of the optimization problem Urach and Posch (2016) consider is also substantially smaller. For example, in the case of two stage designs, Urach and Posch (2016) optimize over at most five parameters but the design proposed here optimizes over 10 parameters.

We assume non-binding futility boundaries, which is generally recommended by the U.S. Food and Drug Administration (Liu and Anderson, 2008). In addition, our method does not assume equal variances of all groups, which is assumed in several of the aforementioned references.

Section 3 describes the data structure, statistics, and the class of adaptive enrichment designs. Section 4 defines the trial design optimization problem. A simulation study that mimics features of the SMART-AV trial is used to compare performance of optimized adaptive versus standard designs, in Section 5.

3 Description of Adaptive Enrichment Designs

3.1 Data Structure, Hypotheses, and Test Statistics

We are interested in comparing two treatments versus a common control in two disjoint subpopulations. Subpopulations must be defined by measurements made before randomization, and this definition must be prespecified in the study protocol. In our motivating example, these subpopulations consist of patients with QRS $\leq$ 150ms and those with QRS $>$ 150ms.

Let π_j be the proportion of participants in subpopulation j; $\pi_1 + \pi_2 = 1$. Let $\mu_{l,j}$ denote the mean outcome under assignment to study arm $l \in \{0, 1, 2\}$ for subpopulation $j \in \{1, 2\}$. We refer to arms $l = 1, 2$ as the treatment arms and $l = 0$ as the control arm. For each subpopulation $j = 1, 2$, let $\mu_{l,j} - \mu_{0,j}$ denote the difference between the mean outcome in treatment group l and the control group. There are four null hypotheses of interest: $H_{l,j} : \mu_{l,j} - \mu_{0,j} \leq 0, l \in \{1, 2\}, j \in \{1, 2\}$, corresponding to no average treatment benefit for each subpopulation by treatment combination. Throughout, the subscript l indicates study arm and the subscript j indicates the subpopulation. Let $\sigma^2_{l,j}, l \in \{0, 1, 2\}, j \in \{1, 2\}$ denote the variance of the primary outcome in study arm by subpopulation combination (l, j).

We consider adaptive enrichment designs, i.e., designs with a preplanned rule for modify-

ing enrollment criteria based on accruing data in an ongoing trial (Wang et al., 2009). These designs may lead to stopping accrual of some treatment by subpopulation combinations for either efficacy or futility at the end of each stage.

In stage 1, both subpopulations are enrolled and each participant is assigned with probability 1/3 to a study arm $l = 0, 1, 2$. At the interim analysis after stage 1, for each subpopulation, the preplanned rule may decide to stop assigning new participants to one or both treatment arms $l = 1, 2$. The decision can differ by subpopulation, e.g., subpopulation 1 may be stopped entirely while subpopulation 2 continues enrollment and assignment to arms $l = 0, 1$.

Each participant is randomized to one of the two treatments or to the control. The arm assignment of each participant is never changed throughout the trial. For any subpopulation and stage, if neither treatment arm ($l = 1, 2$) has been stopped then the randomization ratio is 1:1:1 to each arm $l \in \{0, 1, 2\}$; if a single treatment arm ($l \in \{1, 2\}$) has been stopped, then the randomization ratio is 1:1 to the other arm and control; if both treatment arms ($l = 1, 2$) have been stopped, then the control arm is stopped as well. This randomization method can be approximately achieved by block randomization stratified by subpopulation. A reason we use 1:1 randomization ratios is that different ratios at different stages could lead to bias if the distribution of the primary outcome among subjects enrolled differs across time. In a related setting, randomizing more participants to the common control group has been shown to lead to minor efficiency improvements; Wason et al. (2012). A potential downside of higher allocation to control groups is that it may influence the willingness of subjects to participate in the trial; Halpern et al. (2003).

The following design parameters need to be specified in the study protocol: The maximum number of stages K (though we focus on $K = 2$ in our simulation study); the number of participants enrolled at stage k from subpopulation j who are assigned to arm l (denoted $n_{l,j,k}$), assuming enrollment has not been stopped for that subpopulation by arm combination before stage k; futility boundaries $f_{l,j,k}$ corresponding to stage k, subpopulation j, and treatment arm l. Define $n_k = \sum_{l=0}^{2} \sum_{j=1}^{2} n_{l,j,k}$ as the maximum number of participants that can be enrolled during stage k. By the above assumptions about randomization ratios and the assumption that enrollment is uniform over time and proportional to subpopulation size, we have $n_{l,j,k} = \pi_j n_k/3$ for each $l \in \{0, 1, 2\}, j \in \{1, 2\}, k \leq K$. Define the maximum sample size $n_{max} = \sum_{k=1}^{K} n_k$.

If enrollment is stopped early at stage k for a single treatment $l \in \{1, 2\}$ in subpopulation j, then this does not impact the number enrolled from treatment $l', l' \neq l$, i.e., there is no change in the sample size allocated to the other treatment arm or the control arm. The sample size for the control arm ($l = 0$) in a subpopulation is not impacted unless both treatment arms are stopped; in that case, assignment to the control group in that subpopulation is stopped. This means that if no treatment arm has been stopped for subpopulation j at or before the end of stage $k-1$, then $n_{l,j,k} = \pi_j n_k/3$ newly enrolled participants from subpopulation j are assigned to each arm $l = 0, 1, 2$. If exactly one treatment arm $l \in \{1, 2\}$ has been stopped for subpopulation $j \in \{1, 2\}$ at or before the end of stage $k-1$, then $n_{l,j,k} = \pi_j n_k/3$ newly enrolled participants from subpopulation j are assigned to the other treatment arm

($l' = 3 - l$) and to the control arm (for a total of $2\pi_j n_k/3$ enrolled from subpopulation j in stage k).

Let $S_{i,k}$ be a random variable taking values in $\{1, 2\}$, which indicates whether participant i at stage k belongs to subpopulation 1 or 2. Let $A_{i,k} \in \{0, 1, 2\}$ be the study arm assignment of participant i at stage k. The outcome for participant i at stage k is denoted by $Y_{i,k}$. The outcome can be of any type that ensures that the joint distribution of the test statistics given below by equation (1) follows an asymptotic multivariate normal distribution. The data on participant i at stage k in the trial consists of the vector $(A_{i,k}, S_{i,k}, Y_{i,k})$. For $l, j = 0, 1, 2, k = 1, \ldots, K$, define $\bar{Y}_{l,j,k}$ as the (cumulative) average of all observed primary outcomes from study arm and subpopulation (l, j) which are observed prior to interim analysis k. The statistic used to test null hypothesis $H_{l,j}$ as stage k is a standardized difference between $\bar{Y}_{l,j,k}$ and $\bar{Y}_{0,j,k}$. More formally, the test statistic for contrasting treatment l to the control group in subpopulation j at stage k is given by

$$Z_{l,j,k} = \left(\bar{Y}_{l,j,k} - \bar{Y}_{0,j,k}\right) \left\{ \frac{\sigma_{l,j}^2 + \sigma_{0,j}^2}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} I(S_{i,k'} = j, A_{i,k'} = l)} \right\}^{-1/2}, \tag{1}$$

where $I(X)$ is the indicator variable taking value 1 if X is true and 0 otherwise. If treatment and subpopulation combination (l, j) is not enrolled through stage k, then $Z_{l,j,k}$ is undefined. The difference in population means between treatment group l and the control group for subpopulation j is defined as $\delta_{l,j} = \mu_{l,j} - \mu_{0,j}$. Denote the vector of average treatment effects by $\boldsymbol{\delta} = (\delta_{1,1}, \delta_{2,1}, \delta_{1,2}, \delta_{2,2})$. We assume the joint distribution of the statistics $\mathbf{Z} = \{Z_{l,j,k} : l = 1, 2; j = 1, 2; k = 1, \ldots, K\}$ has the canonical form of (Jennison and Turnbull, 1999, Chapter 3.1), which holds asymptotically for many types of outcomes and statistics. This joint distribution is multivariate normal with mean and covariance matrix given in Supplementary Web Appendix 7.1.

Rosenblum et al. (2016) shows that for binary and continuous outcomes the vector of test statistics $(Z_{l,j,k})_{l=1,2,j=1,2,k=1,\ldots,K}$ depends on the data only through minimal sufficient statistics. The decision rules for the adaptive enrichment design discussed in the next subsections depend only on the data through these test statistics. It is therefore exempt from one of the criticisms of adaptive designs, i.e., that test statistics are often not a function of minimal sufficient statistics (Emerson, 2006).

When outcomes are measured with delay, the participants that are enrolled but have not yet had their outcome observed are referred to as pipeline participants. The pipeline participants do not contribute to the calculations of the test statistics but they contribute to the total sample size.

3.2 Multiple Testing Procedure and Enrollment Modification Rule

We describe our proposed class of adaptive enrichment designs, denoted by $\mathcal{D}_{ADAPT}$. Each such design consists of a multiple testing procedure for the four null hypotheses $H_{l,j}, l = 1, 2, j = 1, 2$, and an enrollment modification rule. Modifications are only made at analyses following each stage.

The enrollment modification rule has the following form: for each subpopulation, enrollment continues until both treatment arms ($l = 1, 2$) in that subpopulation have been stopped; stopping can be for efficacy or futility, based on the multiple testing procedure defined below. The multiple testing procedure involves efficacy and futility boundaries, and is designed to ensure strong control of the familywise Type I error rate, asymptotically. We next describe the construction of efficacy boundaries, followed by how they are applied to determine when each treatment by subpopulation combination is stopped for efficacy.

Efficacy boundaries are based on an error spending approach (G. Lan and DeMets, 1983). Let α denote the desired familywise Type I error rate, e.g., $\alpha = 0.05$. Let $\alpha_{j,k} > 0, j \in \{1, 2\}, 1 \leq k \leq K$ denote the alpha allocation associated with each subpopulation j at stage k, which are required to satisfy $\sum_{j=1}^{2} \sum_{k=1}^{K} \alpha_{j,k} = \alpha$.

Consider any subpopulation $j \in \{1, 2\}$. Start by calculating $(u_{j,1}, z_{j,1})$ by solving

$$P_0 \{\max(Z_{1,j,1}, Z_{2,j,1}) > u_{j,1}\} = \alpha_{j,1} \text{ and } P_0(Z_{1,j,1} > z_{j,1}) = \alpha_{j,1}.$$

where P_0 denotes the global null hypothesis of zero treatment effect for both treatments and subpopulations, which implies the mean of each $Z_{l,j,k}$ is 0.

Once $(u_{j,1}, z_{j,1})$ are calculated, the efficacy boundaries $(u_{j,k}, z_{j,k}), 1 < k \leq K$ are calculated sequentially. That is, at stage k $((u_{j,1}, z_{j,1}) \ldots, (u_{j,k-1}, z_{j,k-1}))$ have already been calculated and $z_{j,k}$ is calculated by finding the smallest value $z_{j,k}$ satisfying

$$P_0(Z_{1,j,k'} \leq z_{j,k'} \text{ for all } k' < k, \text{ and } Z_{1,j,k} > z_{j,k}) \leq \alpha_{j,k}, \tag{2}$$

and then $u_{j,k}$ is calculated by finding the minimum value $u_{j,k} \in [z_{j,k}, \infty)$ such that

$$P_0\left\{ \max(Z_{1,j,k'}, Z_{2,j,k'}) \leq u_{j,k'} \text{ for all } k' < k, \text{ and } \max(Z_{1,j,k}, Z_{2,j,k}) > u_{j,k} \right\} \leq \alpha_{j,k}. \tag{3}$$

The efficacy boundaries $z_{j,k}, k = 1, \ldots, K$ could equivalently be calculated using treatment $l = 2$ instead of treatment $l = 1$, which follows from the canonical covariance structure of the statistics $Z_{l,j,k}$ given in the Supplementary Web Appendix 7.1 and the 1:1 randomization ratio between each treatment and control arm. Equation (2) utilizes the covariance structure between test statistics for the same treatment and subpopulation but at different stages. Equation (3) utilizes the correlation between test statistics which include the same control group and test statistics from the same subpopulation but at different stages.

Now we describe how alpha reallocation can be used to improve power at the last stage K for a subpopulation if the null hypotheses corresponding to both treatments $l = 1, 2$ for the other subpopulation have been rejected. For any $j \in \{1, 2\}$, if both $H_{1,j}, H_{2,j}$ have been rejected at or before analysis K, then recompute both $z_{j',K}$ and $u_{j',K}$ for $j' \neq j$ by replacing $\alpha_{j',K}$ on the right sides of (2) and (3) by $\alpha_{j',K} + \sum_{m=1}^{K} \alpha_{j,m}$. Denote the updated values by $\tilde{z}_{j',K}$ and $\tilde{u}_{j',K}$. Each is less or equal compared to the corresponding value without the alpha reallocation.

Probabilities that involve multivariate normal distributions such as those appearing in equations (3) and (2) can quickly and reliably be calculated using the `R` package `mvtnorm`;

Genz et al. (2012). Binary search can then be used to calculate the smallest efficacy boundaries such that inequalities (3) and (2) hold.

Stopping accrual for a subpopulation by treatment combination (l, j) means that no future participants enrolled from subpopulation j are assigned to arm l; if both treatment arms $l = 1, 2$ have accrual stopped, then no future participants are enrolled from subopulation j.

We next define the enrollment modification rule and multiple testing procedure of the adaptive enrichment design $\mathcal{D}_{ADAPT}$. These involve the aforementioned efficacy boundaries and a set of futility boundaries $\mathcal{F} = (f_{l,j,k}, l = 1, 2; j = 1, 2; k \leq K)$. At the interim analysis taking place at the end of each stage $k < K$, the adaptive enrichment design $\mathcal{D}_{ADAPT}$ is defined by the following sequence of actions for each subpopulation $j \in \{1, 2\}$:

1. *If exactly one treatment arm $l \in \{1, 2\}$ previously had accrual stopped for subpopulation j:* If treatment arm $l \in \{1, 2\}$ was previously stopped for efficacy, then reject $H_{l',j}$ if $Z_{l',j,k} \geq z_{j,k}$ for $l' = 3 - l$. If treatment arm l was stopped for futility, reject $H_{l',j}$ if $Z_{l',j,k} \geq u_{j,k}$ for $l' = 3 - l$.

2. *If neither treatment arm $l = 1, 2$ previously had accrual stopped for subpopulation j:* If $\max(Z_{1,j,k}, Z_{2,j,k}) \geq u_{j,k}$, then reject the null hypothesis $H_{l,j}$ corresponding to the larger statistic. If both $\max(Z_{1,j,k}, Z_{2,j,k}) \geq u_{j,k}$ and $\min(Z_{1,j,k}, Z_{2,j,k}) \geq z_{j,k}$, then reject both subpopulation j null hypotheses $H_{1,j}, H_{2,j}$.

3. For each null hypothesis $H_{l,j}$ rejected in (1) or (2), accrual for the corresponding subpopulation by treatment combination (l, j) is stopped for efficacy. For each null hypothesis $H_{l,j}$ that has not been rejected, the corresponding subpopulation by treatment combination (l, j) has accrual stopped for futility if $Z_{l,j,k} \leq f_{l,j,k}$.

4. If accrual for both treatment arms $l \in \{1, 2\}$ in subpopulation j are stopped (either for efficacy or futility) or if $k = K$, stop all accrual of subpopulation j. Otherwise, continue subpopulation j accrual in the next stage with random assignment to the arms $l \in \{0, 1, 2\}$ that have not been stopped.

The analysis at the end of stage K is conducted according to the sequence 1-4 above, except that if both null hypotheses for subpopulation j are rejected at or before analysis K, then the efficacy thresholds $(u_{j',K}, z_{j',K})$ are replaced by $(\tilde{u}_{j',K}, \tilde{z}_{j',K})$ for the other subpopulation $j' \neq j$.

The trial continues until every treatment by subpopulation combination is stopped for efficacy/futility or the final analysis K is reached. Rejecting any null hypothesis implies that the null hypothesis is rejected at all future stages.

As stated in the introduction, we assume non-binding futility boundaries (Liu and Anderson, 2008). That is, we require the familywise Type I error rate to be controlled even if futility boundaries are ignored.

The above multiple testing procedure results in a consonant multiple testing procedure. The following theorem shows that it strongly controls the familywise Type I error rate. The proof of Theorem 3.1 is given in Supplementary Web Appendix 7.2.

Theorem 3.1. *The above adaptive enrichment design strongly controls the familywise Type I error rate, for any choice of futility boundaries $\mathcal{F}$.*

The above designs can be applied in the special case of single stage ($K = 1$), where only the multiple testing procedure is used.

We say that a multiple testing procedure $\mathcal{M}_1$ uniformly improves upon another multiple testing procedure $\mathcal{M}_2$ if all null hypothesis rejected by $\mathcal{M}_2$ are also rejected by $\mathcal{M}_1$ and there exists a data generating distribution where $\mathcal{M}_1$ has greater power than $\mathcal{M}_2$ for at least one false null hypothesis. Since $z_{j,K} \geq \tilde{z}_{j,K}$ and $u_{j,K} \geq \tilde{u}_{j,K}$ for $j \in \{1, 2\}, 1 \leq k \leq K$ and there exists a scenario such that $z_{j,K} > \tilde{z}_{j,K}$ and $u_{j,K} > \tilde{u}_{j,K}$, it follows that using the efficacy boundaries $(\tilde{u}_{j,K}, \tilde{z}_{j,K})$ uniformly improves upon using the efficacy boundaries $(u_{j,K}, z_{j,K})$.

3.3 Test for Non-inferiority Following Superiority of Both Treatments in a Subpopulation

If both treatments $l = \{1, 2\}$ in the SMART-AV trial were found to be superior to the standard of care for a subpopulation, the investigators were further interested in testing if treatment $l = 1$ (AV delay optimized with the SmartDelay electrogram-based algorithm) was non-inferior to treatment $l = 2$ (echocardiographically optimized AV delay). Within subpopulation j and for a prespecified non-inferiority margin $\tau \leq 1$, interest lies in evaluating if treatment $l = 1$ preserves more than $100 * \tau$ percent of the benefit of treatment $l = 2$ compared to the control $l = 0$. An advantage of this approach is that, since non-inferiority is tested only if superiority of both treatments to the control has already been established, the treatment effect comparing $l = 2$ versus control $l = 0$ has already been assessed within the trial, obviating the need to rely on historical data to estimate this treatment effect.

The inferiority null hypothesis for subpopulation j is defined as $H_j^{(Inf)} : \mu_{1,j} - \mu_{0,j} \leq \tau(\mu_{2,j} - \mu_{0,j})$ or equivalently $H_j^{(Inf)} : \mu_{1,j} - \tau\mu_{2,j} - (1 - \tau)\mu_{0,j} \leq 0$. For $l = 0, 1, 2, j = 1, 2$, define $N_{l,j}$ to be the total number of enrolled participants in study arm by subpopulation combination (l, j). For subpopulation j, define the standardized statistic for testing $H_j^{(Inf)}$ as

$$Z_j^{(Inf)} = \left\{\bar{Y}_{1,j,K} - \tau\bar{Y}_{2,j,K} - (1 - \tau)\bar{Y}_{0,j,K}\right\} \left\{\frac{\sigma_{1,j}^2}{N_{1,j}} + \tau^2\frac{\sigma_{2,j}^2}{N_{2,j}} + (1 - \tau)^2\frac{\sigma_{0,j}^2}{N_{0,j}}\right\}^{-1/2}.$$

Under $H_j^{(Inf)}$, the test statistic is asymptotically normally distributed with unit variance and mean at most 0 (Pigeot et al., 2003). The mean vector and correlation matrix for the family of statistics for testing superiority and then non-inferiority are given in Supplementary Web Appendix 7.3.

Implementation requires specifying the non-inferiority margin τ. That is, specifying how much treatment effect reduction for treatment $l = 1$ compared to $l = 2$ is acceptable. This choice is a clinical judgment and depends on what the benefits of treatment $l = 1$ compared to $l = 2$ are (e.g. in term of safety, side effects, or cost).

4 Optimization: Search Space, Objective Function, and Optimization Method

4.1 Optimization Problem

A generic design in the class $\mathcal{D}_{ADAPT}$ is denoted by D, which consists of the following design parameters (which are specified before the trial starts): the maximum number of stages K, the α allocations ($\alpha_{j,k} : j = 1, 2; k = 1, \ldots, K$), the futility boundaries ($f_{l,j,k} : l = 1, 2; j = 1, 2; k = 1, \ldots, K$), and the sample sizes ($n_{l,j,k} : l = 0, 1, 2; j = 1, 2; k = 1, \ldots, K$).

Let $M = (\mu_{l,j}, \sigma^2_{l,j}, \pi_1 : l = 0, 1, 2; j = 1, 2)$ denote the population parameters. The joint distribution of statistics $\mathbf{Z}$ is determined by the population parameters M and design parameters D. For given M, D, let $ESS(M, D)$ denote the expected sample size for design D under population parameters M. The expectation is with respect to the distribution $\mathbf{Z}$.

We next define our optimization goal, called the objective function, which maps each design D to a real value. It is defined as $ESS^{\Lambda}(D) = \int_M ESS(M, D)d\Lambda(M)$, where Λ is a distribution on the population parameters M. An example of Λ is given in Section 5. Our optimization problem is formulated in the decision theory framework and the only role of Λ is in defining the objective function. Calculating the objective function requires evaluating multivariate integrals. All the results presented in Section 5 approximate the integral by simulating from the multivariate normal distribution.

The optimization problem is to search for the design D that minimizes expected sample size $ESS^{\Lambda}(D)$ while strongly controlling the asymptotic, familywise Type I error rate at level α and satisfying prespecified power constraints. The search space for the optimization problem, denoted by $\mathcal{S}$, consists of a subclass of the designs $\mathcal{D}_{ADAPT}$, e.g., all such designs with $K = 2$.

4.2 Optimization Method

To search for the optimal design, we use simulated annealing. This general purpose optimization algorithm is not guaranteed to find the global optimum solution (which is an open problem), but still can be useful if it returns adaptive designs with improved performance compared to standard designs. We use the `optim` function in `R` which implements the simluated annealing algorithm described in Bélisle (1992). For a more detailed description we refer to Bélisle (1992) and Fisher and Rosenblum (2016).

5 Application to the SMART-AV Trial

5.1 Optimization Problem Definition

The primary outcome in the SMART-AV trial was the six month change in left ventricular end-systolic volume, which is measured on a continuous scale with a six month delay. We set the accrual rate to 20 participants per month. The familywise Type I error rate is set to be

$\alpha = 0.05$ for all designs. As the proportion of the population with narrow QRS is estimated to be 49%, we set $\pi_1 = 0.49$. In the design of the SMART-AV trial, the standard deviation was assumed to be 60ml, and therefore we set $\sigma_{l,j} = 60, l \in \{0,1,2\}, j \in \{1,2\}$ throughout.

Given fixed values of the variances $\sigma^2_{l,j} = 60$ and $\pi_1 = 0.49$, the joint distribution of statistics $\mathbf{Z}$ depends on the population parameters M only through the average treatment effects $\boldsymbol{\delta}$. Therefore, it suffices to define the distribution Λ on $\boldsymbol{\delta}$. The minimum clinically meaningful treatment effect used for powering the SMART-AV trial was 15ml. We use this in our definition of Λ, which is defined to be the equally weighted mixture of the following six scenarios (each with a point mass at a specific value of $\boldsymbol{\delta}$): a. $\boldsymbol{\delta} = (0,0,0,0)$; b. $\boldsymbol{\delta} = (15,0,0,0)$; c. $\boldsymbol{\delta} = (15,15,0,0)$; d. $\boldsymbol{\delta} = (15,0,15,0)$; e. $\boldsymbol{\delta} = (15,15,15,0)$; and f. $\boldsymbol{\delta} = (15,15,15,15)$. Scenario (a) represents the global null hypothesis of no average effect for either treatment in any subpopulation. Scenario (f) represents a benefit of 15ml for each treatment arm by subpopulation combination. The other scenarios involve benefits of some treatments for some subpopulations.

The prior distribution is asymmetric in that a positive treatment effect is only expected in subpopulation two (consisting of participants with wide QRS) if there is also a positive treatment effect in subpopulation one (consisting of participants with narrow QRS). This allows us to incorporate the prior scientific knowledge that the subpopulation consisting of patients with narrow QRS is more likely to benefit from treatment.

The power constraints in each scenario (a)-(f) are that for each treatment arm $l \in \{1,2\}$ and subpopulation $j \in \{1,2\}$ for which the corresponding treatment effect $\delta_{l,j}$ is at least the minimum, clinically meaningful level 15ml, there must be at least 80% power to reject $H_{l,j}$. For example, for the null hypothesis $H_{1,1}$, the power constraints are at least 80% power to reject $H_{1,1}$ in each of scenarios (b)-(f).

5.2 Classes of Designs Compared

We will evaluate the performance of the different designs from the class $\mathcal{D}_{ADAPT}$, with design parameters selected by minimizing expected sample size under power and Type I error constraints as described in Section 4. For computational simplicity, we restrict the maximum number of stages to be two. We next describe four subclasses of designs from $\mathcal{D}_{ADAPT}$, in increasing order of complexity. We will solve the optimization problem for each class and then compare the resulting four designs in terms of expected and maximum sample size.

The first (and simplest) design class has a single stage $K = 1$ with equal α allocation between the two subpopulations, i.e. each $\alpha_{j,1} = \alpha/2$; the sample size is set to be the smallest such that the power and Type I error constraints are all satisfied. The second design class has a single stage design where the α allocation between the two subpopulations is optimized; that is, simulated annealing is used to find the smallest sample size such that there exists a pair $(\alpha_{1,1}, \alpha_{2,1})$ for which the power and Type I error constraints are all satisfied. The third design class from $\mathcal{D}_{ADAPT}$ has $K = 2$ stages and sets equal α allocation across the two subpopulations and stages (i.e. $\alpha_{j,k} = \alpha/4$ for each $j,k = 1,2$) with an interim analysis performed when half of the participants have the primary outcome observed, and all futility

	ESS	MSS
Simple 1-Stage Design	1818	1818
Optimized-alpha 1-Stage Design	1797	1797
Simple 2-Stage Design	1610	1962
Optimized-alpha 2-Stage Design	1581	2320

Table 1: Expected sample size (ESS^{Λ}) and maximum sample size (MSS) for the four designs for the SMART-AV trial. The expected sample size is calculated using the distribution Λ described in Section 5.

boundaries are set to zero. The fourth design class is just as the third except the α allocation, futility boundaries, and interim analysis timing are optimized. We refer to the above four comparison designs (consisting of the optimized design from each class) as the simple 1-stage design, optimized-alpha 1-stage design, simple 2-stage design, and optimized-alpha two-stage design, respectively.

5.3 Results

Table 1 shows the expected and maximum sample size for the four designs. The results show that there is potential for substantial savings in expected sample size by using a 2 stage, adaptive enrichment design compared to a single stage design. The price paid for the decrease in expected sample size is an increase in the maximum sample size. The optimized, 2 stage design has slightly lower expected sample size compared to the non-optimized, 2 stage design, but at the cost of a substantial increase in maximum sample size. This may be due to that the setup of the optimization does not constrain the maximum sample size, so that (intuitively) there was no incentive for the optimizer to search for designs with an optimal trade-off in terms of expected versus maximum sample size. This is an area of future research.

For the optimized-alpha 1-stage design, the proportion of alpha allocated to subpopulation one is 78%.

For the optimized-alpha 2-stage design, an interim analysis was conducted at 43.7% accrual of the primary outcome. 41.1% of the total α was spent on subpopulation 1 in stage 1, 37.0% of the total α was spent on subpopulation 2 in stage 1. 13.5% of the total α was spent on subpopulation 1 in stage 2, and 8.4% of the total α was spent on subpopulation 2 in stage 2. The futility boundaries are $(f_{1,1,1}, f_{2,1,1}, f_{1,2,1}, f_{2,2,1}) = (-2.47, 1.00, 1.05, 1.07)$. The futility boundaries are much more aggressive in stopping the population that is less likely to benefit.

As testing for non-inferiority is a secondary objective, the trial is only powered for superiority testing. Recall that a test for non-inferiority is only performed when both treatments are found superior to the control (i.e., both corresponding null hypotheses are rejected). We calculate the power for the test for non-inferiority in the setting when $\boldsymbol{\delta} = (15, 15, 15, 15)$ using a non-inferiority margin $\tau = 0.7$. For each subpopulation, the test for non-inferiority is set to control the Type I error rate at a 0.025 level, which ensures that the familywise

Type I error rate is at most 0.05. The power to reject the inferiority null hypothesis in each subpopulation is only 12%.

Publicly available, open-source, user-friendly software implementing the four designs compared in this section is available from the website http://rosenblum.jhu.edu

6 Discussion

We developed a new class of adaptive enrichment designs comparing two treatments to a common control in two disjoint subpopulations. The design parameters are selected by minimizing expected sample size subject to Type I error and power constraints. The empirical results show that the proposed design has the potential to substantially reduce the expected sample size compared to using a single stage design. The trade-off is that the maximum sample size is larger for the adaptive enrichment designs.

The design uses simulated annealing to search for the optimal design parameters. The algorithm is not guaranteed to find the global optimum. An interesting future research direction would be to investigate how the starting values for the optimization as well as the selection of temperature parameter affect the performance of the optimization.

Utilizing information from prognostic baseline variables can result in improved estimators compared to using the difference of sample mean estimators given by equation (1). Another interesting extension would be to use these estimators in all analyses.

Acknowledgments

This work was supported by the Patient-Centered Outcomes Research Institute (ME-1306-03198) and the U.S. Food and Drug Administration (HHSF223201400113C). This publication's contents are solely the responsibility of the author and do not necessarily represent the official views of the above agency.

References

Bélisle, C. J. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability 29*(04), 885–895.

Bittman, R. M., J. P. Romano, C. Vallarino, and M. Wolf (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika 96*(2), 399–410.

Bretz, F., T. Hothorn, and P. Westfall (2010). *Multiple comparisons using R*. CRC Press.

Bretz, F., F. Koenig, W. Brannath, E. Glimm, and M. Posch (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine 28*(8), 1181.

Bretz, F., W. Maurer, W. Brannath, and M. Posch (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine 28*(4), 586–604.

Bretz, F., M. Posch, E. Glimm, F. Klinglmueller, W. Maurer, and K. Rohmeyer (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal 53*(6), 894–913.

Bristol, D. R. (1989). Designing clinical trials for two-sided multiple comparisons with a control. *Controlled Clinical Trials 10*(2), 142–152.

Di Scala, L. and E. Glimm (2011). Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine 30*(26), 3067–3081.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*(272), 1096–1121.

Dunnett, C. W., M. Horn, and R. Vollandt (2001). Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. *Journal of Statistical Planning and Inference 97*(2), 367–384.

Dunnett, C. W. and A. C. Tamhane (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine 10*(6), 939–947.

Ellenbogen, K. A., M. R. Gold, T. E. Meyer, I. F. Lozano, S. Mittal, A. D. Waggoner, B. Lemke, J. P. Singh, F. G. Spinale, J. E. Van Eyk, et al. (2010). Primary Results From the SmartDelay Determined AV Optimization: A Comparison to Other AV Delay Methods Used in Cardiac Resynchronization Therapy (SMART-AV) Trial Clinical Perspective. *Circulation 122*(25), 2660–2668.

Emerson, S. S. (2006). Issues in the use of adaptive clinical trial designs. *Statistics in Medicine 25*(19), 3270–3296.

Fisher, A. and M. Rosenblum (2016). Stochastic optimization of adaptive enrichment designs for two subpopulations. *Johns Hopkins University, Dept. of Biostatistics Working Papers. http://biostats.bepress.com/jhubiostat/paper279*.

Food, D. Administration, et al. (1998). E9: statistical principles for clinical trials. *Federal Register 63*(179), 49583–49598.

Friede, T., N. Parsons, and N. Stallard (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine 31*(30), 4309–4320.

G. Lan, K. K. and D. L. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika 70*(3), 659–663.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2012). mvtnorm: Multivariate normal and t distributions. 2012. *R package version 0.9-9992*.

Halpern, S. D., J. H. Karlawish, D. Casarett, J. A. Berlin, R. R. Townsend, and D. A. Asch (2003). Hypertensive patients' willingness to participate in placebo-controlled trials: implications for recruitment efficiency. *American Heart Journal 146*(6), 985 – 992.

Hayter, A. J. and A. C. Tamhane (1991). Sample size determination for step-down multiple test procedures: orthogonal contrasts and comparisons with a control. *Journal of Statistical Planning and Inference 27*(3), 271–290.

Hida, E. and T. Tango (2011). On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Statistics in Medicine 30*(3), 224–231.

Jennison, C. and B. W. Turnbull (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

Kelly, P. J., N. Stallard, and S. Todd (2005). An adaptive group sequential design for phase ii/iii clinical trials that select a single treatment from several. *Journal of Biopharmaceutical statistics 15*(4), 641–658.

Koenig, F., W. Brannath, F. Bretz, and M. Posch (2008). Adaptive dunnett tests for treatment selection. *Statistics in Medicine 27*(10), 1612–1625.

Liu, Q. and K. M. Anderson (2008). On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association 103*(484).

Magirr, D., T. Jaki, and J. Whitehead (2012). A generalized dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika 99*(2), 494–501.

Marcus, R., P. Eric, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*(3), 655–660.

Maurer, W. and F. Bretz (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research 5*(4), 311–320.

O'Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics*, 549–556.

Pigeot, I., J. Schäfer, J. Röhmel, and D. Hauschke (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine 22*(6), 883–899.

Posch, M., F. Koenig, M. Branson, W. Brannath, C. Dunger-Baldauf, and P. Bauer (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine 24*, 36973714.

Posch, M., W. Maurer, and F. Bretz (2011). Type i error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharmaceutical statistics 10*(2), 96–104.

Proschan, M. A. and S. A. Hunsberger (1995). Designed extension of studies based on conditional power. *Biometrics*, 1315–1324.

Rosenblum, M., X. Fang, and H. Liu (2014). Optimal, two stage, adaptive enrichment designs for randomized trials using sparse linear programming. *Johns Hopkins University, Dept. of Biostatistics Working Papers. http://biostats.bepress.com/jhubiostat/paper273*.

Rosenblum, M., H. Liu, and E.-H. Yen (2014). Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *Journal of the American Statistical Association 109*(507), 1216–1228.

Rosenblum, M., B. Luber, R. E. Thompson, and D. Hanley (2016). Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine*.

Spiessens, B. and M. Debois (2010). Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary clinical trials 31*(6), 647–656.

Stallard, N. and T. Friede (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine 27*(29), 6209–6227.

Stein, K. M., K. A. Ellenbogen, M. R. Gold, B. Lemke, I. F. Lozano, S. Mittal, F. G. Spinale, J. E. Van Eyk, A. D. Waggoner, and T. E. Meyer (2010). SmartDelay Determined AV Optimization: A Comparison of AV Delay Methods Used in Cardiac Resynchronization Therapy (SMART-AV): Rationale and Design. *Pacing and Clinical Electrophysiology 33*(1), 54–63.

Thall, P. F., R. Simon, and S. S. Ellenberg (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika 75*(2), 303–310.

Urach, S. and M. Posch (2016). Multi-arm group sequential designs with a simultaneous stopping rule. *Statistics in Medicine*.

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

Wason, J. and T. Jaki (2012). Optimal design of multi-arm multi-stage trials. *Statistics in Medicine 31*(30), 4269–4279.

Wason, J., D. Magirr, M. Law, and T. Jaki (2012). Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 0962280212465498.

Whitehead, J. and T. Jaki (2009). One-and two-stage design proposals for a phase ii trial comparing three active treatments with control using an ordered categorical endpoint. *Statistics in Medicine 28*(5), 828–847.

7 Appendix

7.1 Distribution of Test Statistics

We derive the mean and the covariance matrix of the asymptotic joint distribution of the test statistics.

As before we assume that the randomization ratio within each subpopulation is $1:1:1$. That is, within each subpopulation we have equal allocation to each treatment and control group that have not already been stopped for efficacy or futility. This can approximately be achieved using block randomization.

Theorem 7.1. *The joint distribution of* $(Z_{1,1,1}, Z_{1,2,1}, Z_{2,1,1}, Z_{2,2,1}, \ldots, Z_{1,1,K}, Z_{1,2,K}, Z_{2,1,K}, Z_{2,2,K})$ *is asymptotically normal with mean*

$$E[Z_{l,j,k}] = \frac{\delta_{l,j}}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^{k}\sum_{i=1}^{n_s} I(S_{i,s}=j)I(A_{i,s}=l)}}} = \frac{\mu_{l,j}-\mu_{0,j}}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^{k}\sum_{i=1}^{n_s} I(S_{i,s}=j)I(A_{i,s}=l)}}}.$$

and covariance matrix with

$$\begin{aligned}
Cov(Z_{l,j,k}, Z_{l,j,k}) &= 1 \\
Cov(Z_{l,j,k}, Z_{l,j,k'}) &= \sqrt{\frac{\sum_{s=1}^{\min(k,k')} n_{j,s}}{\sum_{s=1}^{\max(k,k')} n_{j,s}}} \\
Cov(Z_{l,j,k}, Z_{l',j,k}) &= \frac{\sigma^2_{0,j}}{\sqrt{(\sigma^2_{l,j}+\sigma^2_{0,j})(\sigma^2_{l',j}+\sigma^2_{0,j})}} \\
Cov(Z_{l,j,k}, Z_{l',j,k'}) &= \frac{\sigma^2_{0,j}}{\sqrt{(\sigma^2_{l,j}+\sigma^2_{0,j})(\sigma^2_{l',j}+\sigma^2_{0,j})}} \sqrt{\frac{\sum_{s=1}^{\min(k,k')} n_{j,s}}{\sum_{s=1}^{\max(k,k')} n_{j,s}}} \\
Cov(Z_{l,1,k}, Z_{l',2,k'}) &= 0 \quad \textit{for all other combinations of } (l,k,l',k')
\end{aligned}$$

Proof. Basic calculations show that $E[Z_{l,j,k}]$ has the desired form. When deriving the covariance matrix, we will repeatedly use the property that for a given $l \in \{0,1,2\}$

$$\begin{aligned}
&Cov\left(\frac{1}{\sum_{s=1}^{k} n_{j,s}} \sum_{s=1}^{k}\sum_{i=1}^{n_s} I(A_{i,s}=l)I(S_{i,s}=j)Y_{i,s}, \frac{1}{\sum_{s=1}^{k'} n_{j,s}} \sum_{s=1}^{k'}\sum_{i=1}^{n_s} I(A_{i,s}=l)I(S_{i,s}=j)Y_{i,s}\right) \\
&= \frac{1}{\sum_{s=1}^{k} n_{j,s}} \frac{1}{\sum_{l=1}^{k'} n_{j,s}} \sum_{s=1}^{\min(k,k')}\sum_{i=1}^{n_s} I(A_{i,s}=l)I(S_{i,s}=j)\sigma^2_{l,j} \\
&= \frac{\sigma^2_{l,j}}{\max(\sum_{s=1}^{k} n_{j,s}, \sum_{s=1}^{k'} n_{j,s})}
\end{aligned}$$

We have

$$
\begin{aligned}
&Cov(Z_{l,j,k}, Z_{l,j,k}) \\
&= \frac{1}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}} \frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{l=1}^k n_{j,s}}}} \frac{1}{\sum_{s=1}^k n_{j,s}} \frac{1}{\sum_{s=1}^k n_{j,s}} \sum_{s=1}^{k} \sum_{i=1}^{n_s} I(S_{i,s}=j)[I(A_{i,s}=l)\sigma^2_{l,j} + I(A_{i,s}=0)\sigma^2_{0,j}] \\
&= 1.
\end{aligned}
$$

If $k = k'$, $j = j'$ and $l \neq l'$

$$
\begin{aligned}
&Cov(Z_{l,j,k}, Z_{l',j,k}) \\
&= \frac{1}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}} \frac{\sigma^2_{l',j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}}}} \frac{1}{\sum_{s=1}^k n_{j,s}} \frac{1}{\sum_{s=1}^k n_{j,s}} \sum_{s=1}^{k} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=0)\sigma^2_{0,j} \\
&= \frac{1}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}} \frac{\sigma^2_{l',j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}}}} \frac{\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}} \\
&= \frac{\sigma^2_{0,j}}{\sqrt{(\sigma^2_{l,j}+\sigma^2_{0,j})(\sigma^2_{l',j}+\sigma^2_{0,j})}}.
\end{aligned}
$$

If $k \neq k'$, $j = j'$, and $l = l'$

$$
\begin{aligned}
&Cov(Z_{l,j,k}, Z_{l,j,k'}) \\
&= \frac{1}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}} \frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^{k'} n_{j,s}}}} \frac{1}{\sum_{s=1}^k n_{j,s}} \frac{1}{\sum_{s=1}^{k'} n_{j,s}} \\
&\sum_{s=1}^{\min(k,k')} \sum_{i=1}^{n_s} I(S_{i,s}=j)[I(A_{i,s}=l)\sigma^2_{l,j} + I(A_{i,s}=0)\sigma^2_{0,j}] \\
&= \frac{1}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^k n_{j,s}} \frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^{k'} n_{j,s}}}} \frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\max(\sum_{s=1}^k n_{j,s}, \sum_{s=1}^{k'} n_{j,s})} \\
&= \sqrt{\frac{\min(\sum_{s=1}^k n_{j,s}, \sum_{s=1}^{k'} n_{j,s})}{\max(\sum_{s=1}^k n_{j,s}, \sum_{s=1}^{k'} n_{j,s})}}.
\end{aligned}
$$

If $k \neq k'$, $l \neq l'$ and $j = j'$

$$
\begin{aligned}
&Cov(Z_{l,j,k}, Z_{l',j,k'}) \\
&= \frac{1}{\sqrt{\frac{\sigma^2_{l,j}+\sigma^2_{0,j}}{\sum_{s=1}^{k} n_{j,s}} \frac{\sigma^2_{l',j}+\sigma^2_{0,j}}{\sum_{s=1}^{k'} n_{j,s}}}} \frac{1}{\sum_{s=1}^{k} n_{j,s}} \frac{1}{\sum_{s=1}^{k'} n_{j,s}} \sum_{l=1}^{\min(k,k')} \sum_{i=1}^{n_s} I(S_{i,s} = j) I(A_{i,s} = 0) \sigma^2_{0,j} \\
&= \sqrt{\frac{\min(\sum_{s=1}^{k} n_{j,s}, \sum_{s=1}^{k'} n_{j,s})}{\max(\sum_{s=1}^{k} n_{j,s}, \sum_{s=1}^{k'} n_{j,s})}} \frac{\sigma^2_{0,j}}{\sqrt{(\sigma^2_{0,j} + \sigma^2_{l,j})(\sigma^2_{0,j} + \sigma^2_{l',j})}}.
\end{aligned}
$$

□

Note that if all variances are assumed equal $\frac{\sigma^2_{0,j}}{\sqrt{(\sigma^2_{0,j}+\sigma^2_{l,j})(\sigma^2_{0,j}+\sigma^2_{l',j})}} = \frac{1}{2}$.

7.2 Proof of Theorem 3.1

For convenience of notation below, define $\tilde{z}_{j,k}$ and $\tilde{u}_{j,k}$ to equal $z_{j,k}$ and $u_{j,k}$, respectively, for all $j \in \{1, 2\}$ and $k < K$.

Define a closed testing procedure using the following local tests:

- Test of elementary null hypothesis $H_{l,j}$: reject if $Z_{l,j,k} > \tilde{z}_{j,k}$ for at least one $k \in \{1, \ldots, K\}$.
- Intersection test of $H_{1,j} \cap H_{2,j}, j \in \{1, 2\}$: reject if $Z_{l,j,k} > \tilde{u}_{j,k}$ for at least one pair (l, k), $l \in \{1, 2\}, k \in \{1, \ldots, K\}$.
- For $j \neq j', j, j' \in \{1, 2\}$, any $l, l' \in \{1, 2\}$, intersection test of $H_{l,j} \cap H_{l',j'}$: reject if $Z_{l,j,k} > z_{j,k}$ or $Z_{l',j',k} > z_{j',k}$ for at least one $k \in \{1, \ldots, K\}$.
- For $j \neq j', j, j' \in \{1, 2\}$, any $l' \in \{1, 2\}$, intersection test of $H_{1,j} \cap H_{2,j} \cap H_{l',j'}$: reject if $Z_{l,j,k} > u_{j,k}$ for at least one pair $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$ or if $Z_{l',j',k} > z_{j',k}$ for some $k \in \{1, \ldots, K\}$.
- Intersection test of all 4 null hypotheses: reject if $Z_{l,j,k} > u_{j,k}$ for at least one triple $(l, j, k), l, j \in \{1, 2\}, k \in \{1, \ldots, K\}$.

Now we show that for a fixed $l, j \in \{1, 2\}$ $\mathcal{D}_{ADAPT}$ rejects $H_{l,j}$ if and only if the closed testing procedure rejects every intersection hypothesis involving $H_{l,j}$. Without loss of generality we assume $(l, j) = (1, 1)$. We start by showing that if $\mathcal{D}_{ADAPT}$ rejects $H_{1,1}$, every intersection test involving $H_{1,1}$ is rejected.

If $\mathcal{D}_{ADAPT}$ rejects $H_{1,1}$, then at least one of the following two statements is true: 1) there exits a $k \in \{1, \ldots, K\}$ s.t. $Z_{1,1,k} > \tilde{u}_{1,k}$; 2) there exits a $k \in \{1, \ldots, K\}$ such that $Z_{1,1,k} > \tilde{z}_{1,k}$ and $Z_{2,1,m} > \tilde{u}_{1,m}$ for some $m \leq k$. We now go through the five different types of intersection tests.

- The intersection test $H_{1,1}$: As $\tilde{u}_{1,k} \geq \tilde{z}_{1,k}$ for all $k \in \{1, \ldots, K\}$, both conditions 1) and 2) imply that $Z_{1,1,k} > \tilde{z}_{1,k}$ for at least one $k \in \{1, \ldots, K\}$. Hence, the test of the intersection test only involving $H_{1,1}$ is rejected.

- The intersection test $H_{1,1} \cap H_{2,1}$: In order for at least one of 1) or 2) to hold, at least one of $Z_{1,1,k} > \tilde{u}_{1,k}$ for at least one $k \in \{1, \ldots, K\}$, or $Z_{2,1,k} > \tilde{u}_{1,k}$ for at least one $k \in \{1, \ldots, K\}$ needs to be true. That implies that $H_{1,1} \cap H_{2,1}$ is rejected.

- The intersection test $H_{1,1} \cap H_{l',2}$ for $l' \in \{1, 2\}$: If alpha re-allocation is done from population 1 to population 2, $Z_{1,1,k} > z_{1,k}$ for at least one $k \in \{1, \ldots, K\}$. If alpha re-allocation is done from population 2 to population 1, $Z_{l',2,k} > z_{2,k}$ for at least one $k \in \{1, \ldots, K\}$. If no reallocation is done, then $\tilde{z}_{j,k} = z_{j,k}$ and $\tilde{u}_{j,k} = u_{j,k}$ for all combinations of $(j, k), j \in \{1, 2\}, k \in \{1, \ldots, K\}$. As $\tilde{u}_{1,k} \geq \tilde{z}_{1,k}$ for all $k \in \{1, \ldots, K\}$, it follows that if 1) and 2) hold then there exists a $k \in \{1, \ldots, K\}$ s.t. $Z_{1,1,k} > \tilde{z}_{1,k} = z_{1,k}$. So the intersection test is rejected for all three cases.

- The intersection test $H_{1,2} \cap H_{2,2} \cap H_{1,1}$: If alpha re-allocation is done from population 1 to population 2, then $Z_{1,1,k} > z_{1,k}$ for at least one $k \in \{1, \ldots, K\}$. If alpha re-allocation is done from population 2 to population 1, then $Z_{l,2,k} > u_{2,k}$ for at least one $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$ combination. If no reallocation is done, then $\tilde{z}_{j,k} = z_{j,k}$ and $\tilde{u}_{j,k} = u_{j,k}$ for all combinations of $(j, k), j \in \{1, 2\}, k \in \{1, \ldots, K\}$. As $\tilde{u}_{1,k} \geq \tilde{z}_{1,k}$ for all $k \in \{1, \ldots, K\}$, it follows that if 1) and 2) hold then there exists a $k \in \{1, \ldots, K\}$ s.t. $Z_{1,1,k} > \tilde{z}_{1,k} = z_{1,k}$. So the intersection test is rejected for all three cases.

- The intersection test $H_{1,1} \cap H_{2,1} \cap H_{l',2}$ for $l' = 1, 2$. If alpha re-allocation is done from population 1 to population 2, then $Z_{l,1,k} > u_{1,k}$ for at least one $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$ pair. If alpha re-allocation is done from population 2 to population 1, then $Z_{l',2,k} > z_{2,k}$ for at least one $k \in \{1, \ldots, K\}$. If no reallocation is done, then $\tilde{z}_{j,k} = z_{j,k}$ and $\tilde{u}_{j,k} = u_{j,k}$ for all combinations of $(j, k) j \in \{1, 2\}, k \in \{1, \ldots, K\}$. It follows that if 1) and 2) hold then there exists a pair $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$ s.t. $Z_{l,1,k} > \tilde{u}_{1,k} = u_{1,k}$. So the intersection test is rejected for all three cases.

- The intersection test $H_{1,1} \cap H_{2,1} \cap H_{1,2} \cap H_{2,2}$: If alpha re-allocation is done from population 1 to population 2, then $Z_{l,1,k} > u_{1,k}$ for at least one $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$ pair. If alpha re-allocation is done from population 2 to population 1, then $Z_{l,2,k} > u_{2,k}$ for at least one pair $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$. If no reallocation is done, then $\tilde{z}_{j,k} = z_{j,k}$ and $\tilde{u}_{j,k} = u_{j,k}$ for all combinations of $(j, k), j \in \{1, 2\}, k \in \{1, \ldots, K\}$. It follows that if 1) and 2) hold there exists a pair $(l, k), l \in \{1, 2\}, k \in \{1, \ldots, K\}$ s.t. $Z_{l,1,k} > \tilde{u}_{1,k} = u_{2,k}$. So the intersection test is rejected for all three cases.

This shows that if $\mathcal{D}_{ADAPT}$ rejects $H_{1,1}$ then all intersection tests involving $H_{1,1}$ are also rejected.

If all intersection tests involving $H_{1,1}$ are rejected, it follows that the elementary test $\cap_{\{(l,j)=(1,1)\}} H_{l,j}$ is rejected. Hence, $Z_{1,1,k} > \tilde{z}_{1,k}$ for at least one $k \in \{1, \ldots, K\}$, which implies that both conditions 1) and 2) hold and therefore the multiple testing procedure $\mathcal{D}_{ADAPT}$ rejects $H_{1,1}$. This completes the proof that for a fixed $l, j \in \{1, 2\}$ $\mathcal{D}_{ADAPT}$ rejects $H_{l,j}$ if and only if the closed testing procedure rejects every intersection hypothesis involving $H_{l,j}$.

Now we want to show that all intersection tests control the familywise Type I error rate.

- Elementary null hypothesis $H_{l,j}$: For a given $l, j \in \{1, 2\}$ under the null $H_{l,j}$, the probability of making a Type I error is

$$\sum_{k=1}^{K} P(Z_{l,j,k'} \leq \tilde{z}_{j,k'} \text{ for all } k' < k, Z_{l,j,k} > \tilde{z}_{j,k}) \leq \sum_{k=1}^{K} \alpha_{1,k} + \sum_{k=1}^{K} \alpha_{2,k} = \alpha$$

  where the inequality follows from the construction of the efficacy boundaries $\tilde{z}_{j,k}$.

- Intersection of $H_{1,j} \cap H_{2,j}$: Here the null hypothesis is that $H_{1,j}$ and $H_{2,j}$ are both true. The probability of a Type I error in subpopulation $j \in \{1, 2\}$ is

$$\begin{aligned}\sum_{k=1}^{K} & P(\max(Z_{1,j,k'}, Z_{2,j,k'}) \leq \tilde{u}_{j,k'} \text{ for all } k' < k, \max(Z_{1,j,k}, Z_{2,j,k}) > \tilde{u}_{j,k}) \\ & \leq \sum_{k=1}^{K} \alpha_{1,k} + \sum_{k=1}^{K} \alpha_{2,k} = \alpha.\end{aligned}$$

  Here, the inequality follows from the construction of the efficacy boundaries $\tilde{u}_{j,k}$.

- Let $j \neq j'$, with $j, j' \in \{1, 2\}$ and $l, l' \in \{1, 2\}$. We look at the intersection test of $H_{l,j} \cap H_{l',j'}$: Under the null of $H_{l,j}$ and $H_{l',j'}$ both being true, the Type I error is bounded above by

$$\begin{aligned}& \sum_{k=1}^{K} P(Z_{l,j,k'} \leq z_{j,k'} \text{ for all } k' < k, Z_{l,j,k} > z_{j,k}) \\ & + \sum_{k=1}^{K} P(Z_{l',j',k'} \leq z_{j',k'} \text{ for all } k' < k, Z_{l',j',k} > z_{j',k}) \\ & \leq \sum_{k=1}^{K} \alpha_{j,k} + \sum_{k=1}^{K} \alpha_{j',k} = \alpha,\end{aligned}$$

  where the inequality follows from the construction of the efficacy boundaries $z_{j,k}$

- For $j \neq j'$ with $j, j' \in \{1, 2\}$ and $l' \in \{1, 2\}$, intersection test of $H_{1,j} \cap H_{2,j} \cap H_{l',j'}$: Under the null of all three hypothesis in the test being true, the Type I error is bounded

from above by

$$\sum_{k=1}^{K} P(\max(Z_{1,j,k'}, Z_{2,j,k'}) \leq u_{j,k'} \text{ for all } k' < k, \max(Z_{1,j,k}, Z_{2,j,k}) > u_{j,k})$$
$$+ \sum_{k=1}^{K} P(Z_{l',j',k'} \leq z_{j',k'} \text{ for all } k' < k, Z_{l',j',k} > z_{j',k})$$
$$\leq \sum_{j=1}^{2} \sum_{k=1}^{K} \alpha_{j,k} = \alpha,$$

where the inequality follows from the construction of the efficacy boundaries $(z_{j,k}, u_{j,k})$.

- Intersection test of all 4 null hypotheses: Under the null of no treatment effect in any subpopulation and treatment combination, the Type I error is bounded above by

$$\sum_{j=1}^{2} \sum_{k=1}^{K} P(\max(Z_{1,j,k'}, Z_{2,j,k'}) \leq u_{j,k'} \text{ for all } k' < k, \max(Z_{1,j,k}, Z_{2,j,k}) > u_{j,k})$$
$$\leq \sum_{j=1}^{2} \sum_{k=1}^{K} \alpha_{j,k} = \alpha,$$

where the inequality follows from the construction of the efficacy boundaries $u_{j,k}$.

The closed testing principle implies that $\mathcal{D}_{ADAPT}$ controls the familywise Type I error rate at a level α.

7.3 Mean Structure and Correlation for Test of Non-Inferiority

The correlation between the test statistics for non-inferiority and test for superiority is given by

$$Cov(Z_j^{(Inf)}, Z_{1,j,k}) = \frac{\frac{\sigma_{1,j}^2}{\sum_{s=1}^{\max(K_j^{(1)},k)} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=1)} + \frac{(1-\tau)\sigma_{0,j}^2}{\sum_{s=1}^{\max(K_j^{(0)},k)} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=1)}}{\sqrt{\frac{\sigma_{1,j}^2+\sigma_{0,j}^2}{\sum_{s=1}^{k} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=1)}} M_{Den,j}}$$

$$Cov(Z_j^{(Inf)}, Z_{2,j,k}) = \frac{-\frac{\tau\sigma_{2,j}^2}{\sum_{s=1}^{\max(K_j^{(2)},k)} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=2)} + \frac{(1-\tau)\sigma_{0,j}^2}{\sum_{s=1}^{\max(K_j^{(0)},k)} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=2)}}{\sqrt{\frac{\sigma_{2,j}^2+\sigma_{0,j}^2}{\sum_{s=1}^{k} \sum_{i=1}^{n_s} I(S_{i,s}=j) I(A_{i,s}=2)}} M_{Den,j}}$$

The mean vector is given by

$$E[Z_j^{(Inf)}] = \frac{\mu_{1,j} - \tau\mu_{2,j} - (1-\tau)\mu_{0,j}}{\sqrt{\frac{\sigma_{1,j}^2}{N_{1,j}} + \tau^2 \frac{\sigma_{2,j}^2}{N_{2,j}} + (1-\tau)^2 \frac{\sigma_{0,j}^2}{N_{0,j}}}},$$

where $N_{l,j}, l = 0, 1, 2$ is the sample size in group l and subpopulation j.