Johns Hopkins University, Dept. of Biostatistics Working Papers

1-29-2018

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

January 27, 2018

1 Abstract

We consider the problem of designing a randomized trial for comparing two treatments versus a common control in two disjoint subpopulations. The subpopulations could be defined in terms of a biomarker or disease severity measured at baseline. The goal is to determine which treatments benefit which subpopulations. We develop a new class of adaptive enrichment designs tailored to solving this problem. Adaptive enrichment designs involve a preplanned rule for modifying enrollment based on accruing data in an ongoing trial. The proposed designs have preplanned rules for stopping accrual of treatment by subpopulation combinations, either for efficacy or futility. The motivation for this adaptive feature is that interim data may indicate that a subpopulation, such as those with lower disease severity at baseline, is unlikely to benefit from a particular treatment while uncertainty remains for the other treatment and/or subpopulation. We optimize these adaptive designs to have the minimum expected sample size under power and Type I error constraints. We compare

the performance of the optimized adaptive design versus an optimized non-adaptive (single stage) design. Our approach is demonstrated in simulation studies that mimic features of a completed trial of a medical device for treating heart failure. The optimized adaptive design has 25% smaller expected sample size compared to the optimized non-adaptive design; however, the cost is that the optimized adaptive design has 8% greater maximum sample size. Open-source software that implements the trial design optimization is provided, allowing users to investigate the tradeoffs in using the proposed adaptive versus standard designs.

Keywords: Randomized Clinical Trial, Treatment Effect Heterogeneity

2 Introduction

Our trial design problem is motivated by 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). The proposed designs can stop accrual of 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 treatment by subpopulation combinations lead to improved outcomes. Our proposed class of adaptive enrichment designs uses a multiple testing procedure that combines advantageous features from the methods of Dunnett (1955), Maurer and Bretz (2013), and Rosenblum et al. (2016).

The proposed adaptive designs have the following properties: they leverage correlations between treatment effect estimators that share a common control; they allow for continued accrual of remaining treatments/subpopulations after some null hypotheses are rejected in order to continue testing the remaining null hypotheses; they improve power by lowering the rejection threshold for the remaining null hypotheses after a null hypothesis has been rejected; they strongly control the familywise Type I error rate, asymptotically; the multiple testing procedure is a function of only minimal sufficient statistics, and therefore avoids power losses that can affect some adaptive designs as described by Emerson (2006). We use non-binding futility boundaries, which are generally preferred by the U.S. Food and Drug Administration (Liu and Anderson, 2008).

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 solution, we use simulated annealing, a general purpose optimization method. Fisher and Rosenblum (ress) used simulated annealing to optimize over a class of designs evaluating the effectiveness of a single treatment in two subpopulations. Our setting differs since 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 overall population (but

not considering subpopulations). Magirr et al. (2012) generalized the method of Dunnett (1955) to such trials and show how to compute sample sizes under power constraints for the least favorable configuration of treatment effects. Wason and Jaki (2012) used simulated annealing to search for the efficacy boundaries that minimize expected sample size for trials with multiple arms and stages. In both of the aforementioned references, the trial is stopped when the first null hypothesis is rejected, unlike our approach. Several designs, e.g., Thall et al. (1988); Kelly et al. (2005), have been proposed that pick only one treatment at the interim analysis to continue to the later stages. Stallard and Friede (2008) proposed an adaptive design with treatment selection at an interim analysis, but the number of treatments allowed to continue after each stage must be prespecified. In contrast, our designs do not a priori restrict how many treatment by subpopulation combinations will continue to later stages. Posch et al. (2005), Koenig et al. (2008), and Bretz et al. (2010) propose adaptive designs based on the p-value combination or conditional error function approaches; these approaches allow more flexibility in the adaptation rule than those considered here, but at the cost of not using data only through minimal sufficient statistics, which can lead to power loss.

Urach and Posch (2016) optimize multi-arm, group sequential designs. Their designs differ from ours in that they only consider the overall population, use a different form of efficacy boundaries, and do not reallocate alpha between null hypotheses to improve power. The search space of the optimization problem of Urach and Posch (2016) is substantially smaller than ours. For example, Urach and Posch (2016) optimize over at most five parameters but the optimized design presented in Section 6.3 involves over 24 parameters.

Section 3 describes the data structure, null hypotheses, and statistics. The proposed class of adaptive enrichment designs is defined in Section 4. Section 5 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 6.

The open-source software implementing our trial design optimization, which has a graphical user-interface that runs on a web-browser, is described in Section 7. Directions for future research are discussed in Section 8.

3 Problem Setup

3.1 Data Structure and Null Hypotheses

We are interested in comparing two treatments versus a common control in two disjoint subpopulations that partition the overall population. 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 (short QRS) and those with QRS $>$ 150ms (long QRS).

Throughout, the subscript $a \in \{0, 1, 2\}$ denotes the study **a**rm, $s \in \{1, 2\}$ denotes the **s**ubpopulation, and $k \leq K$ denotes the stage of the trial. Let π_s denote the proportion of the combined population in subpopulation $s \in \{1, 2\}$. We assume these proportions are known and $\pi_1 + \pi_2 = 1$. Let $\mu_{a,s}$ denote the mean outcome under assignment to study arm $a \in \{0, 1, 2\}$ for subpopulation $s \in \{1, 2\}$. We refer to arms $a = 1, 2$ as the treatment arms and $a = 0$ as the control arm. The difference between the population mean of the outcome under assignment to treatment $a \in \{1, 2\}$ versus control for subpopulation s is defined as $\delta_{a,s} = \mu_{a,s} - \mu_{0,s}$. Denote the vector of average treatment effects by $\boldsymbol{\delta} = (\delta_{1,1}, \delta_{2,1}, \delta_{1,2}, \delta_{2,2})$.

There are four null hypotheses of interest: $H_{a,s} : \delta_{a,s} \leq 0, a \in \{1, 2\}, s \in \{1, 2\}$, corresponding to no average treatment benefit for each treatment by subpopulation combination. Let $\sigma^2_{a,s}, a \in \{0, 1, 2\}, s \in \{1, 2\}$ denote the variance of the primary outcome in study arm by subpopulation combination (a, s).

3.2 Sample Sizes Per Stage for Each Treatment by Subpopulation Combination

Each participant is randomized to one of the two treatment arms or to the control arm, and her/his arm assignment is never changed throughout the trial. In stage 1, both subpopulations are enrolled and each participant is assigned with probability 1/3 to a study arm $a = 0, 1, 2$. At the interim analysis after each stage, for each subpopulation, the preplanned rule may decide to stop assigning new participants to one or both treatment arms $a \in \{1, 2\}$. The decision can differ by subpopulation, e.g., subpopulation 1 may be stopped entirely while subpopulation 2 continues enrollment and assignment to arms $a = 0, 1$.

Stopping accrual for a treatment by subpopulation combination (a, s) means that no future participants enrolled from subpopulation s are assigned to arm a. If both treatment arms $a = 1, 2$ have accrual stopped, then no more subpopulation s participants are enrolled. After stopping accrual of a treatment by subpopulation combination for futility, the corresponding null hypothesis is no longer tested.

For any subpopulation and stage, if neither treatment arm $a \in \{1, 2\}$ has been stopped then the randomization ratio is 1:1:1 to each arm $a \in \{0, 1, 2\}$; if a single treatment arm $a \in \{1, 2\}$ has been stopped, then the randomization ratio is 1:1 to the other treatment arm and control; if both treatment arms $a = 1, 2$ 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 arm has been shown to lead only to minor power improvements (Wason et al., 2012). Also, a higher allocation to the control arm might reduce the willingness of subjects to participate in the

trial (Halpern et al., 2003).

The following design parameters need to be prespecified in the study protocol: the maximum number of stages K; the number of subpopulation s participants enrolled during stage k assigned to arm a (denoted $n_{a,s,k}$), assuming enrollment has not been stopped for that arm by subpopulation combination. Define $n_k = \sum_{a=0}^{2} \sum_{s=1}^{2} n_{a,s,k}$ 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 (which we assume throughout), we have $n_{a,s,k} = \pi_s n_k/3$ for each $a \in \{0,1,2\}, s \in \{1,2\}, k \leq K$. Define the maximum sample size $n = \sum_{k=1}^{K} n_k$, and the vector of sample sizes $\mathcal{N} = (n_{a,s,k} : a = 0,1,2; s = 1,2; k = 1,\ldots,K)$. Complete prespecification of the adaptive design is required by regulators such as the U.S. Food and Drug Administration (FDA, 2010; FDA, 2016).

If no treatment arm has been stopped for subpopulation s at or before the end of stage $k-1$, then $n_{a,s,k} = \pi_s n_k/3$ newly enrolled participants from subpopulation s are assigned to each arm $a = 0,1,2$ during stage k. If exactly one treatment arm $a \in \{1,2\}$ has been stopped for subpopulation $s \in \{1,2\}$ at or before the end of stage $k-1$, then $n_{a,s,k} = \pi_s n_k/3$ newly enrolled participants from subpopulation s are assigned to the other treatment arm ($a' = 3-a$) and to the control arm (for a total of $2\pi_s n_k/3$ enrolled from subpopulation s) during stage k.

The above rules ensure that the number enrolled during stage k for any arm by subpopulation combination (a,s) is either the prespecified $n_{a,s,k}$ or 0. The impact is that the statistics defined below have the canonical structure from Jennison and Turnbull (1999, Ch. 3.1), and so are asymptotically multivariate normal with mean vector and covariance matrix that are straightforward to compute. This facilitates the computation of efficacy boundaries in Section 4.2.

3.3 Data Structure and Statistics

Let $S_{i,k}$ be a random variable taking values in $\{1, 2\}$, which indicates whether participant i enrolled during 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 $Y_{i,k}$, which can be continuous, binary, or integer valued. The data on participant i at stage k in the trial consists of the vector $(S_{i,k}, A_{i,k}, Y_{i,k})$. We assume that conditioned on $(S_{i,k} = s, A_{i,k} = a)$, the outcome $Y_{i,k}$ is an independent draw from an unknown distribution $Q_{a,s}$ with mean $\mu_{a,s}$ and variance $\sigma^2_{a,s}$.

For each $a \in \{0, 1, 2\}, s \in \{1, 2\}, k \in \{1, \ldots, K\}$, define $\bar{Y}_{a,s,k}$ as the (cumulative) average of all primary outcomes from study arm a and subpopulation s observed prior to analysis k. The statistic used to test null hypothesis $H_{a,s}$ at analysis k is the following standardized difference between sample means of the outcome comparing treatment arm a versus control:

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

where $I(X)$ is the indicator variable taking value 1 if X is true and 0 otherwise. If treatment by subpopulation combination (a, s) is not enrolled through stage k, then $Z_{a,s,k}$ is undefined. We assume the joint distribution of the statistics $\mathbf{Z} = \{Z_{a,s,k} : a = 1, 2; s = 1, 2; k = 1, \ldots, K\}$ 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 S.1.

For binary outcomes, the vector of statistics $\mathbf{Z}$ depends on the data only through minimal sufficient statistics (Rosenblum et al., 2016); this also holds for normally distributed outcomes if the variance terms in the display above are replaced by sample variances. The decision rules for the adaptive enrichment design below depend only on the data through these statistics.

It is therefore exempt from the criticism of some adaptive designs that test statistics are not a function of minimal sufficient statistics (Emerson, 2006).

When outcomes are measured with delay, some participants may be enrolled but not yet have their outcomes observed at an interim analysis. These participants do not contribute to the statistics at that analysis, but they do count toward the sample size enrolled (which is important since our goal is to minimize expected sample size).

4 Adaptive Enrichment Designs

4.1 Overview

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 $H_{a,s}, a = 1, 2, s = 1, 2$, and an enrollment modification rule. The enrollment modification rule has the following form: for each subpopulation, enrollment continues until both treatment arms ($a = 1, 2$) in that subpopulation have been stopped. Stopping can be for efficacy or futility, and can only occur at the analysis following each stage. The multiple testing procedure defined below involves efficacy and futility boundaries on the z-scale, 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 whether to stop each treatment by subpopulation combination for efficacy at each analysis.

4.2 Efficacy and Futility Boundaries

Efficacy boundaries, denoted $(u_{s,k}, z_{s,k})$, are constructed using an error spending approach (Lan and DeMets, 1983). Let α denote the desired familywise Type I error rate, e.g., $\alpha = 0.05$. Let $\alpha_{s,k} > 0, s \in \{1, 2\}, 1 \leq k \leq K$ denote the prespecified alpha allocation associated

with each subpopulation s at stage k. These are required to satisfy $\sum_{s=1}^{2}\sum_{k=1}^{K}\alpha_{s,k}=\alpha$. The value of each $\alpha_{s,k}$ (along with other design parameters) will be determined using optimization as described in Section 5. There is no treatment-specific subscript a in the alpha allocations $\alpha_{s,k}$ since for each subpopulation s and stage k, the efficacy boundaries for rejecting $H_{a,s}$ are the same for each $a \in \{1,2\}$.

For each subpopulation $s \in \{1,2\}$, we compute the efficacy boundaries $\{(u_{s,k}, z_{s,k}) : k = 1,\ldots,K\}$ sequentially. For stage $k=1$, $(u_{s,1}, z_{s,1})$ are the solutions to

$$P_0\left\{\max(Z_{1,s,1}, Z_{2,s,1}) > u_{s,1}\right\} = \alpha_{s,1} \text{ and } P_0(Z_{1,s,1} > z_{s,1}) = \alpha_{s,1},$$

where P_0 denotes the global null hypothesis $\boldsymbol{\delta} = (0,0,0,0)$ of zero average treatment effect for every treatment by subpopulation combination, which implies each $Z_{a,s,k}$ has mean 0.

At the end of each stage $k>1$, $(u_{s,1}, z_{s,1})\ldots,(u_{s,k-1}, z_{s,k-1})$ have already been calculated and $z_{s,k}$ is calculated by finding the smallest value $z_{s,k}$ satisfying

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

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

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

The efficacy boundaries $z_{s,k}, k = 1,\ldots,K$ could equivalently be calculated using treatment $a=2$ instead of treatment $a=1$, which follows from the canonical covariance structure of the statistics $Z_{a,s,k}$ given in the Supplementary Web Appendix S.1 and the 1:1 randomization ratio between each treatment and control arm. The probability in (2) involves the covariance structure among statistics for the same treatment and subpopulation but at different stages. The probability in (3) uses the correlation among statistics for the same subpopulation but

different treatment arms and stages.

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

Probabilities that involve multivariate normal distributions such as those appearing in equations (2) and (3) can quickly and reliably be calculated using the `R` package `mvtnorm` (Genz et al., 2017). Binary search can then be used to calculate the smallest efficacy boundaries satisfying inequalities (2) and (3).

The futility boundaries $\mathcal{F} = (f_{a,s,k} \in \mathbb{R} : a = 1, 2; s = 1, 2; k \leq K - 1)$ are unrestricted and, like the alpha allocations $\alpha_{s,k}$, will be optimized as described in Section 5.

4.3 Class of Adaptive Enrichment Designs

A generic adaptive enrichment design in the class $\mathcal{D}_{ADAPT}$ is denoted by $D = (K, \mathcal{E}, \mathcal{F}, \mathcal{N})$, and consists of the following design parameters (which are specified before the trial starts): the maximum number of stages K, the alpha allocations $\mathcal{E} = (\alpha_{s,k} : s = 1, 2; k \leq K)$, the futility boundaries $\mathcal{F}$, and the sample sizes $\mathcal{N}$. We next define the enrollment modification rule and multiple testing procedure for each $D \in \mathcal{D}_{ADAPT}$. At the analysis taking place at the end of each stage $k \leq K$, for each subpopulation $s \in \{1, 2\}$ where at least one treatment arm continued accrual through stage k, the following sequence of actions is taken:

1. *If exactly one treatment arm $a \in \{1, 2\}$ had accrual stopped for subpopulation s at a previous analysis $k' < k$, then do the following:* If treatment arm $a \in \{1, 2\}$ was

previously stopped for efficacy in subpopulation s and $Z_{a',s,k} \geq z_{s,k}$ for the other treatment arm $a' = 3 - a$, then reject $H_{a',s}$. Otherwise, if treatment arm a was previously stopped for futility in subpopulation s and $Z_{a',s,k} \geq u_{s,k}$ for $a' = 3 - a$, then reject $H_{a',s}$.

2. *If neither treatment arm $a = 1, 2$ had accrual stopped for subpopulation s at a previous analysis $k' < k$:* If both $\max(Z_{1,s,k}, Z_{2,s,k}) \geq u_{s,k}$ and $\min(Z_{1,s,k}, Z_{2,s,k}) \geq z_{s,k}$, then reject both subpopulation s null hypotheses $H_{1,s}, H_{2,s}$. Otherwise, if $\max(Z_{1,s,k}, Z_{2,s,k}) \geq u_{s,k}$, then reject the null hypothesis $H_{a,s}$ corresponding to the larger statistic.

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

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

The trial continues until every treatment by subpopulation combination is stopped for efficacy/futility or the final analysis K is reached. If the trial continues to the end of stage K, then the following *extra step* is conducted (after conducting steps 1-4 above for each subpopulation at analysis K): If both null hypotheses for a subpopulation $s' \in \{1, 2\}$ were rejected at or before analysis K, then the efficacy thresholds $(u_{s,K}, z_{s,K})$ are replaced by $(\tilde{u}_{s,K}, \tilde{z}_{s,K})$ for the other subpopulation $s \neq s'$ and steps 1-4 are conducted again for subpopulation s. This extra step can only improve power or leave it unchanged since each of $\tilde{u}_{s,K}, \tilde{z}_{s,K}$ is less or equal to the corresponding efficacy boundary without the tilde.

Rejecting any null hypothesis implies that the null hypothesis is rejected at all future stages. For any subpopulation $s \in \{1, 2\}$ and stage $k \leq K$, only one of steps 1 and 2 can be applied (depending on which treatment by subpopulation combinations were stopped at previous analyses). The above procedure leads to the same decisions regardless of whether 1-4 are applied first to subpopulation $s = 1$ or $s = 2$. The above steps can be applied in the special case of a single stage design ($K = 1$), where only the multiple testing procedure is used.

The above multiple testing procedure incorporates features from previous work. Dunnett (1955) uses tests based on the maximum of different statistics to control the familywise Type I error rate, as we do in (3). Maurer and Bretz (2013) and Rosenblum et al. (2016) reallocate alpha from rejected null hypotheses to the remaining null hypotheses, leading to lower rejection thresholds and greater power. Our procedure does this through the lower rejection thresholds $\tilde{z}_{s,K}$ and $\tilde{u}_{s,K}$. The way that we combine the above features to construct the multiple testing procedure and enrollment modification rule in 1-4 above, is tailored to our specific trial design problem.

4.4 Familywise Type I Error Rate

Control of the familywise Type I error rate at level α means that the probability of rejecting at least one true null hypothesis is at most α. *Strong* control means that this holds for any mean treatment effect vector $\boldsymbol{\delta} \in \mathbb{R}^4$. We assume non-binding futility boundaries (Liu and Anderson, 2008), i.e., we require the familywise Type I error rate to be strongly controlled even if futility boundaries are ignored. The following theorem is proved in the Supplementary Web Appendix:

Theorem 4.1. *For any $K \geq 1$, sample sizes $\mathcal{N}$ satisfying the assumptions in Section 3.2, futility boundaries $\mathcal{F}$, and positive-valued $\alpha_{s,k}$ that sum to α, the corresponding adaptive*

enrichment design $D \in \mathcal{D}_{ADAPT}$ strongly controls the familywise Type I error rate at level α, asymptotically.

5 Trial Design Optimization: Search Space, Objective Function, and Optimization Method

5.1 Optimization Problem

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

We next define our optimization goal, called the objective function, which maps each design D to a real value (with smaller values being more desirable). The objective function is defined as $ESS^{\Lambda}(D) = \int_{\boldsymbol{\theta}} ESS(\boldsymbol{\theta}, D) d\Lambda(\boldsymbol{\theta})$, where Λ is a distribution on the population parameters $\boldsymbol{\theta}$. An example of Λ that consists of a discrete set of point masses on scenarios of interest is given in Section 6. Our optimization problem is formulated in the decision theory framework and the only role of Λ is in defining the objective function.

The optimization problem is to search for the design $D \in \mathcal{D}_{ADAPT}$ that minimizes expected sample size $ESS^{\Lambda}(D)$ under prespecified power constraints. (By Theorem 4.1, all designs $D \in \mathcal{D}_{ADAPT}$ are guaranteed to strongly control the asymptotic, familywise Type I error rate.) The power constraints in our problem consist of M scenarios, i.e., population parameter vectors denoted $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$. These are chosen by the clinical investigator to represent scenarios of interest. The power constraints corresponding to each scenario $\boldsymbol{\theta}^{(m)}$ are that for each pair $(a, s) : a \in \{1, 2\}, s \in \{1, 2\}$ for which the average treatment effect

$\delta_{a,s}^{(m)} = \mu_{a,s}^{(m)} - \mu_{0,s}^{(m)}$ is at least the minimum, clinically meaningful level denoted $\delta_{\min}$, the power to reject $H_{a,s}$ must be at least 80%. We let $\text{POW}(\boldsymbol{\theta}^{(m)}, D, a, s)$ denote the probability that the design D rejects (at least) null hypothesis $H_{a,s}$ when data are generated according to the population parameter vector $\boldsymbol{\theta}^{(m)}$.

The search space for our optimization problem is all designs $D \in \mathcal{D}_{ADAPT}$ that have at most $K \leq 4$ stages. We restricted to at most 4 stages since there were diminishing improvements as the number of stages was increased from 2 to 4. Larger numbers of stages lead to both a larger search space and more challenging computations in evaluating the objective function.

5.2 Optimization Method

To search for the optimal design, we use a general purpose optimization algorithm called simulated annealing (SA). While it is not guaranteed to find the global optimum solution, which is an open research question for our problem, it may find adaptive designs with improved performance compared to standard designs, which is our goal.

SA iteratively proposes a new candidate vector D' of design parameters by randomly perturbing the current candidate design parameters D. The new design D' is accepted as a replacement if it is superior to the current design D in terms of a composite performance score $V(D)$ defined below that combines the objective function (expected sample size) with penalty terms to account for the power constraints. If the candidate is not superior, it may still be accepted with some probability. By occasionally accepting less optimal candidates, this allows the potential to escape a local minimum. The performance and evolution of the algorithm are controlled by a cooling schedule, which determines the rate at which suboptimal candidate designs are accepted and how new candidates are generated. We use the `optim` function in `R` which implements the simluated annealing algorithm described in Bélisle (1992).

Each candidate design D is evaluated by the SA algorithm in terms of the objective function $ESS^{\Lambda}(D)$ and how well it satisfies the power constraints. A composite performance score $V(D)$ is computed by combining these as follows:

$$V(D) = ESS^{\Lambda}(D) + \lambda \sum_{m=1}^{M} \sum_{a,s} I\left(\delta_{a,s}^{(m)} \geq \delta_{\min}\right) \left\{0.8 - \text{POW}(\boldsymbol{\theta}^{(m)}, D, a, s)\right\}_{+},$$

where $(x)_+ = \max(x, 0)$ and λ is a positive constant that sets the penalty for failing to achieve the power constraints. The terms on the right side of the above display penalize for violation of the power constraints, with the penalty proportional to how far the actual power is from the desired 80% power. The term $I(\delta_{a,s}^{(m)} \geq \delta_{\min})$ is the indicator of the treatment effect for treatment by subpopulation combination (a, s) exceeding the minimum, clinically meaningful level in scenario $\boldsymbol{\theta}^{(m)}$; this term is included since we only require power to be at least 80% when that condition holds. If all power constraints are satisfied by D then the term on the right is 0 and all that remains is the expected sample size $ESS^{\Lambda}(D)$. We set $\lambda = 10^6$ in our optimization in the next section.

The lengths of the futility boundary and alpha-allocation vectors increase with the number of stages. As a consequence, the dimension of the search space increases with the number of stages. For this reason, we keep the number of stages K fixed for each run of the simulated annealing algorithm. For each $K \in \{2, 3, 4\}$, we ran 200 parallel versions of the simulated annealing algorithm with each version using 500 iterations. All parameters in the search space are restricted to fall within their required domains, e.g., each $\alpha_{s,k}$ and n_k must be positive and the $\alpha_{s,k}$ must sum to α. In addition, we restrict all interim analyses to occur between when 10% and 90% of the primary outcomes are observed.

Calculating expected sample size and power requires integrating over multivariate normal distributions. We approximate each such integral by 50,000 Monte Carlo draws from the corresponding multivariate normal distribution.

6 Application to the SMART-AV Trial

6.1 Optimization Problem Definition

The primary outcome in the SMART-AV trial was the six month change in left ventricular end-systolic volume (in ml), which is measured six months after enrollment. Subpopulations 1 and 2 are defined as those with short and long QRS, respectively. We set the accrual rate to 20 participants per month. The familywise Type I error rate is set to be $\alpha = 0.05$.

In the SMART-AV trial, the proportion of participants with short QRS was 49% and the outcome standard deviation was assumed to be 60ml. We mimic these by setting $\pi_1 = 0.49$ and $\sigma_{a,s} = 60$ for each $a \in \{0, 1, 2\}, s \in \{1, 2\}$ throughout. Given these values, the joint distribution of statistics $\mathbf{Z}$ depends on the population parameters $\boldsymbol{\theta}$ only through the average treatment effects $\boldsymbol{\delta} = (\delta_{1,1}, \delta_{2,1}, \delta_{1,2}, \delta_{2,2})$ where $\delta_{a,s} = \mu_{a,s} - \mu_{0,s}$. Therefore, it suffices to define the distribution Λ (used in the objective function) and power constraint scenarios $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ in terms of $\boldsymbol{\delta}$ rather than the full vector $\boldsymbol{\theta}$.

The minimum clinically meaningful treatment effect used for powering the SMART-AV trial was $\delta_{\min} = 15$ml. 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 $\boldsymbol{\delta}$): $\boldsymbol{\delta}^{(1)} = (0, 0, 0, 0)$; $\boldsymbol{\delta}^{(2)} = (15, 0, 0, 0)$; $\boldsymbol{\delta}^{(3)} = (15, 15, 0, 0)$; $\boldsymbol{\delta}^{(4)} = (15, 0, 15, 0)$; $\boldsymbol{\delta}^{(5)} = (15, 15, 15, 0)$; $\boldsymbol{\delta}^{(6)} = (15, 15, 15, 15)$. Scenario 1 represents the global null hypothesis of no average effect for every treatment by subpopulation combination. Scenario 6 represents a benefit of 15ml for each treatment by subpopulation combination. The other scenarios involve benefits of some treatments for some subpopulations.

The distribution Λ is asymmetric in that there is a positive treatment effect in subpopulation 2 (long QRS) only when there is also a positive treatment effect in subpopulation 1 (short QRS). It is here that we incorporated the prior scientific knowledge that the short QRS subpopulation is more likely to benefit from treatment. Analogously, we incorporated

that treatment 2 is expected to benefit a subpopulation (compared to control) only when the same holds for treatment 1.

We use the same six vectors $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(6)}$ above to define the power constraints. This means that in each scenario 1-6, for each treatment arm $a \in \{1, 2\}$ by subpopulation $s \in \{1, 2\}$ combination where the corresponding treatment effect $\delta_{a,s} \geq 15$ml, we require at least 80% power to reject $H_{a,s}$. For example, consider the null hypothesis $H_{1,1}$; the power constraints are that in each scenario 2-6, the power to reject (at least) $H_{1,1}$ is at least 80%.

6.2 Classes of Designs Compared

We describe four subclasses of the designs $\mathcal{D}_{ADAPT}$, in increasing order of complexity. In Section 6.3, we solve the optimization problem for each class and compare the resulting four optimized designs in terms of expected and maximum sample sizes.

The first design class, called simple 1-stage designs, has a single stage $K = 1$ with equal α allocation between the two subpopulations, i.e. each $\alpha_{s,1} = \alpha/2$. The sample size n is optimized to be the smallest such that the power and Type I error constraints are all satisfied.

The second design class, called optimized 1-stage designs, has a single stage where the α allocation between the two subpopulations is optimized. That is, simulated annealing is used to search for the smallest sample size n such that there exists a pair $(\alpha_{1,1}, \alpha_{2,1})$ for which the power and Type I error constraints are satisfied.

The third design class, called simple adaptive designs, optimizes over the number of stages $K \in \{2, 3, 4\}$ and the maximum sample size n, but the following are set (not optimized): the alpha allocation is set to be equally partitioned ($\alpha_{s,k} = \alpha/(2K)$ for $s = 1, 2; k \leq K$), analysis times are equally spaced in terms of the number of observed outcomes, and all futility boundaries are set to zero.

The fourth design class, called optimized adaptive designs, is just as the third class except the alpha allocation, analysis timing, and futility boundaries are optimized.

The reason we use the term "adaptive" only for the third and fourth classes of designs is that 1-stage designs do not involve any adaptations. We refer to the optimized design from each class above as the simple 1-stage design, optimized 1-stage design, simple adaptive design, and optimized adaptive design, respectively. The simulated annealing algorithm requires initial values to be input for all design parameters that it optimizes. These are given in Supplementary Web Appendix S.2.

6.3 Results

The rightmost 2 columns of Table 1 show the expected and maximum sample sizes for the optimal design in each of the four classes. The optimized adaptive design has 25% smaller expected sample size compared to the optimized 1-stage design. However, the cost is that the optimized adaptive design has 8% greater maximum sample size than the optimized 1-stage design.

The optimized adaptive design has substantially lower expected and maximum sample size compared to the simple adaptive design. This shows the importance of optimizing the alpha allocation, analysis timing, and futility boundaries for the adaptive designs.

Columns 3-6 of Table 1 show design parameters from each of the four designs. The only free parameter in the simple 1-stage design was the sample size n, whose optimized value is 1818. For the optimized 1-stage design, the proportion of alpha allocated to subpopulation one is 54%. This only slightly differs from the simple 1-stage design, which allocates 50% of alpha to each subpopulation. This minor improvement over the simple 1-stage design produced only a small sample size reduction (from 1818 to 1779).

The analyses for the optimized adaptive design occur when 39%, 63%, 70%, and 100% of the primary outcomes are observed. For scenarios 1-6, the expected sample sizes for the optimized adaptive design are 1234, 1273, 1381, 1284, 1395, and 1477, respectively. Their mean, i.e., ESS^{Λ}, is 1341.

Design	Stage	Eff. Bnd. $(u_{1,k}, u_{2,k})$	Eff. Bnd. $(z_{1,k}, z_{2,k})$	Futility Bnd.	$\alpha_{s,k}/0.05$	ESS	MSS
Simple 1-Stage	1	(2.2,2.2)	(2.0,2.0)	NA	(0.5,0.5)	1818	1818
Optimized 1-Stage	1	(2.2,2.2)	(1.9,2.0)	NA	(0.54,0.46)	1779	1779
Simple Adaptive	1	(2.7,2.7)	(2.5,2.5)	(0,0,0,0)	(1/8,1/8)	1528	2154
	2	(2.6,2.6)	(2.4,2.4)	(0,0,0,0)	(1/8,1/8)		
	3	(2.6,2.6)	(2.3,2.3)	(0,0,0,0)	(1/8,1/8)		
	4	(2.5,2.5)	(2.2,2.2)	NA	(1/8,1/8)		
Optimized Adaptive	1	(2.5,3.2)	(2.3,3.0)	(0.1,0.6,0.7,0.8)	(0.21,0.03)	1341	1917
	2	(2.5,3.3)	(2.2,3.1)	(-0.8,-1.6,-1.7,1.6)	(0.17,0.01)		
	3	(2.7,2.6)	(2.4,2.4)	(-0.4,-3.2,-1.8,-0.8)	(0.02,0.14)		
	4	(2.3,2.5)	(2.1,2.2)	NA	(0.25,0.17)		

Table 1: Design parameters for each of the four designs that are solutions to the SMART-AV trial optimization problem defined in Section 6.1. The third and fourth columns are the efficacy boundaries $(u_{1,k}, u_{2,k})$ and $(z_{1,k}, z_{2,k})$, respectively. For the adaptive designs, the fifth column gives the futility boundaries $(f_{1,1,k}, f_{2,1,k}, f_{1,2,k}, f_{2,2,k}), k = 1, 2, 3$; NA indicates not applicable, which is the case for the final stage of each design. The sixth column gives the alpha allocations $(\alpha_{1,k}, \alpha_{2,k})/0.05$ (rescaled by 0.05 so they sum to 1) for each stage $k = 1, \ldots, K$. The last two columns report the expected sample size (ESS^{Λ}), and maximum sample size (MSS) for each design.

Table 2 gives the rejection probability for each null hypothesis under each scenario and design. This represents power (for scenarios where the corresponding treatment effect is positive) or Type I error (for scenarios where the corresponding treatment effect is zero, indicated by bold numbers). It also gives the familywise Type I error rate (FWER) for each design and scenario combination, which is always at most 0.05. For the optimized adaptive design, the maximum familywise Type I error rate across the six scenarios is 0.047. Since the futility boundaries are non-binding, the optimized adaptive design does not exhaust the allowed familywise Type I error rate of 0.05. When no futility stopping is applied to the optimized adaptive design, it exhausts the familywise Type I error rate, i.e., the familywise Type I error rate is 0.05.

In scenario 1, where all null hypotheses are true, the Type I errors in columns 3-6 sum to a value greater than the familywise Type I error rate (FWER). This is expected since each of columns 3-6 gives the power to reject *at least* the corresponding null hypothesis, and FWER is the probability of rejecting at least one true null hypothesis.

The power constraints are satisfied by each design, which follows from Table 2 since all the non-boldface numbers (power) are at least 0.8. In scenario 6, all but the optimized adaptive design have power close to 0.9 for each null hypothesis; the optimized adaptive design has power closer to the required 0.8, which was achieved by a combination of higher futility boundaries and an asymmetric alpha allocation. In this and other scenarios, the optimized adaptive design saves resources (reflected in lower expected sample size) by achieving power closer to what is required.

Figure 1 shows the distribution of sample sizes for each of the six scenarios for the optimized adaptive design. Of the simulated trials conducted to evaluate the performance of the optimized adaptive design, 96% had sample size smaller than the optimized 1-stage design's sample size of 1779.

		Rejection Probabilities				
		Subpopulation 1		Subpopulation 2		FWER
Design	Scenario	$H_{1,1}$	$H_{2,1}$	$H_{1,2}$	$H_{2,2}$	
	1	**0.015**	**0.015**	**0.015**	**0.014**	**0.050**
	2	0.80	**0.015**	**0.025**	**0.014**	**0.049**
Simple	3	0.80	0.82	**0.026**	**0.026**	**0.050**
1-Stage	4	0.84	**0.027**	0.84	**0.027**	**0.045**
	5	0.84	0.87	0.84	**0.043**	**0.043**
	6	0.89	0.90	0.89	0.90	**0**
	1	**0.015**	**0.013**	**0.015**	**0.013**	**0.049**
	2	0.80	**0.014**	**0.027**	**0.014**	**0.049**
Optimized	3	0.80	0.80	**0.027**	**0.023**	**0.049**
1-Stage	4	0.84	**0.026**	0.84	**0.026**	**0.044**
	5	0.84	0.86	0.84	**0.044**	**0.044**
	6	0.89	0.89	0.88	0.89	**0**
	1	**0.014**	**0.014**	**0.014**	**0.013**	**0.049**
	2	0.80	**0.016**	**0.024**	**0.016**	**0.050**
Simple	3	0.80	0.85	**0.022**	**0.011**	**0.032**
Adaptive	4	0.83	**0.024**	0.83	**0.024**	**0.042**
	5	0.83	0.88	0.83	**0.039**	**0.039**
	6	0.88	0.90	0.89	0.90	**0**
	1	**0.019**	**0.0091**	**0.017**	**0.0090**	**0.047**
	2	0.80	**0.0095**	**0.028**	**0.0093**	**0.044**
Optimized	3	0.81	0.80	**0.029**	**0.015**	**0.043**
Adaptive	4	0.84	**0.020**	0.81	**0.021**	**0.036**
	5	0.84	0.84	0.81	**0.032**	**0.032**
	6	0.87	0.81	0.83	0.83	**0**

Table 2: Rejection probabilities for each scenario and null hypothesis for the four designs. Boldface rejection probabilities correspond to Type I error and the non-bold rejection probabilities correspond to power. The last column gives the familywise Type I error rate (FWER). The rejection probability for each $H_{a,s}$ is the probability of rejecting at least that null hypothesis.

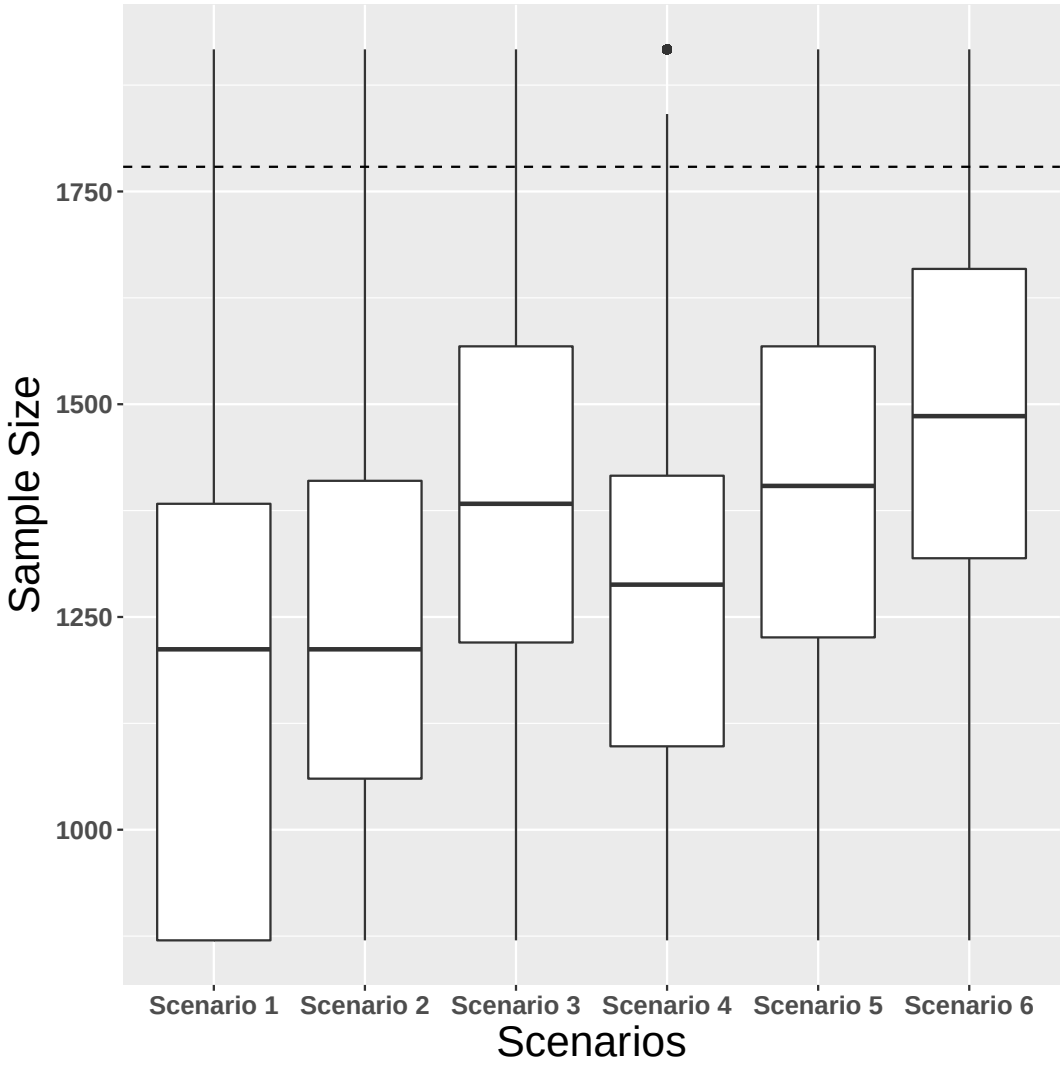

Figure 1: Boxplots of sample size distributions for the optimized adaptive design under each of the six scenarios. The horizontal dashed line indicates the sample size of the optimized 1-stage design.

7 Software for Optimizing Our Adaptive Designs

Open-source software for optimizing over each of the aforementioned four design classes is available at http://rosenblum.jhu.edu. The software allows the user to input their own optimization problem by specifying the outcome type (continuous, binary, or time-to-event), the distribution Λ (consisting of a discrete set of point masses), the power constraint scenarios $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ and the required power for each null hypothesis under each such scenario. The output is a reproducible, automatically generated report describing the performance of the optimized design (computed using simulated annealing) from each class. The software has a graphical user-interface that runs on a web-browser, whose purpose it to make the software accessible to users without requiring knowledge of a specific statistical programming language. The software is also available as an R package at
https://github.com/mrosenblum/AdaptiveDesignOptimizer

8 Discussion

We used simulated annealing to optimize the design parameters. A future research direction is to investigate the impact of the starting values for the optimization problem as well as the temperature parameter used by SA. Also, other optimization methods, e.g., gradient-based methods, could be compared to SA.

Adjusting for prognostic baseline variables can lead to improved treatment effect estimators compared to using the difference of sample mean estimator (Yang and Tsiatis, 2001). Wald statistics based on such covariate-adjusted estimators could be used in place of the z-statistics $Z_{a,s,k}$.

The alpha reallocation described in Section 4.2 can be generalized such that, when both null hypotheses for subpopulation s are rejected, the alpha from subpopulation s gets real-

located to the other subpopulation s' across multiple stages (not just the final stage K). It can be reallocated in any proportions to stages $\{k', \ldots, K\}$, where k' is the first stage where both null hypotheses for subpopulation s are rejected, as long as this is done according to an algorithm that is prespecified (and not a function of the data except through k'). We conjecture that this may help to reduce expected sample size by lowering efficacy boundaries at earlier stages for one subpopulation when both null hypothesis are rejected for the other subpopulation. The proof of Theorem 4.1 in Supplementary Web Appendix S.4 is given for the aforementioned, generalized reallocation method.

The optimization problem described in Section 5 minimizes expected sample size subject to power and Type I error constraints. An interesting alternative would be to minimize some linear combination of maximum and expected sample size subject to power and Type I error constraints.

An alternative approach to calculate the efficacy boundaries $u_{s,k}$ is to replace (3) by finding the smallest $u_{s,k} \in [z_{s,k}, \infty)$ such that

$$P_0\left\{Z_{a,s,k'} > u_{s,k'} \text{ for at least one pair } (a, k') \text{ with a} \in \{1,2\}, k' \leq k\right\} \leq \sum_{k'=1}^{k} \alpha_{s,k'}. \qquad (4)$$

The main difference between the above display and (3) is that the former is in terms of cumulative $\alpha_{s,k}$ over the current and previous stages, while the latter considers each $\alpha_{s,k}$ separately. The two algorithms for computing $u_{s,k}$ can differ if the efficacy boundary $u_{s,k}$ calculated using (3) does not fully exhaust the available $\alpha_{s,k}$, which can happen if the minimum $u_{s,k} \in [z_{s,k}, \infty)$ satisfying (3) is at $u_{s,k} = z_{s,k}$. In such a case, the boundaries at subsequent stages computed using (4) could be lower than the corresponding boundaries computed using (3), leading to more power. As discussed at the end of the proof of Theorem 4.1 in Supplementary Web Appendix S.4, using efficacy boundaries based on equation (4) strongly controls the familywise Type I error rate.

If both treatments 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 AV delay optimized with the SmartDelay electrogram-based algorithm was non-inferior to echocardiographically optimized AV delay. In Section S.3 of the Supplementary Web Appendix, we augment our class of adaptive enrichment designs by adding non-inferiority testing. This is done in a way that does not reduce power for any of the original null hypotheses $H_{a,s}, a = 1, 2; s = 1, 2$, and still guarantees strong control of the familywise Type I error rate, asymptotically. For the optimized adaptive design, we calculated the power of the non-inferiority tests with non-inferiority margin 0.7 in scenario $\boldsymbol{\delta}^{(6)} = (15, 15, 15, 15)$. For each subpopulation, the power to reject the corresponding inferiority null hypothesis is 12%. Having low power for the non-inferiority test is not surprising, since non-inferiority tests often require greater sample sizes than superiority tests. However, we were curious to understand how low this power would be.

A limitation of the proposed design is that the simulated annealing algorithm is not guaranteed to find the global optimum, which is an open research problem. Another limitation is that using an adaptive design results in larger maximum sample size compared to a single stage design. Furthermore, implementing an adaptive design is more logistically complex than implementing a simpler design. Another limitation of adaptive enrichment designs is that they generally require the time from enrollment until the outcome is observed to be relatively short compared to the enrollment period; otherwise, long times between enrollment and measuring the outcome would prevent sufficient information from accruing in time to make a useful decision about modifying enrollment.

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 authors 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.

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

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.

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.

FDA (2010). Draft Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics. https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf.

FDA (2016). Adaptive Designs for Medical Device Clinical Studies. Guidance for Industry and Food and Drug Administration Staff. http://www.fda.

gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf.

Fisher, A. and M. Rosenblum (In Press). Stochastic optimization of adaptive enrichment designs for two subpopulations. *Journal of Biopharmaceutical Statistics. Working paper version: http://biostats.bepress.com/jhubiostat/paper279*.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2017). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-6.

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.

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.

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

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.

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

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.

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

Rosenblum, M., T. Qian, Y. Du, H. Qiu, and A. Fisher (2016). Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches. *Biostatistics 17*(4), 650–662.

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.

Yang, L. and A. A. Tsiatis (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *Am. Stat. 55*(4), 314–321.

Supplementary Web Appendix

References to figures, tables, theorems and equations preceded by "S-" are internal to this supplement; all other references refer to the main paper.

S.1 Distribution of Test Statistics

We derive the mean and the covariance matrix of the asymptotic joint distribution of the test statistics. We make the assumptions stated in 3.

Theorem S.1.1. *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*

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

and covariance matrix with

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

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

$$
\begin{aligned}
&Cov\left(\frac{1}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \sum_{\tilde{k}=1}^{k} \sum_{i=1}^{n_{\tilde{k}}} I(A_{i,\tilde{k}} = a) I(S_{i,\tilde{k}} = s) Y_{i,\tilde{k}}, \frac{1}{\sum_{\tilde{k}=1}^{k'} n_{s,\tilde{k}}} \sum_{\tilde{k}=1}^{k'} \sum_{i=1}^{n_{\tilde{k}}} I(A_{i,\tilde{k}} = a) I(S_{i,\tilde{k}} = s) Y_{i,\tilde{k}} \right) \\
&= \frac{1}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \frac{1}{\sum_{a=1}^{k'} n_{s,\tilde{k}}} \sum_{\tilde{k}=1}^{\min(k,k')} \sum_{i=1}^{n_{\tilde{k}}} I(A_{i,\tilde{k}} = a) I(S_{i,\tilde{k}} = s) \sigma_{a,s}^2 \\
&= \frac{\sigma_{a,s}^2}{\max(\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}, \sum_{\tilde{k}=1}^{k'} n_{s,\tilde{k}})}
\end{aligned}
$$

We have

$$Cov(Z_{a,s,k}, Z_{a,s,k})$$
$$= \frac{1}{\sqrt{\frac{\sigma^2_{a,s}+\sigma^2_{0,s}}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \frac{\sigma^2_{a,s}+\sigma^2_{0,s}}{\sum_{a=1}^{k} n_{s,\tilde{k}}}}} \frac{1}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \frac{1}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \sum_{\tilde{k}=1}^{k} \sum_{i=1}^{n_{\tilde{k}}} I(S_{i,\tilde{k}} = s)[I(A_{i,\tilde{k}} = a)\sigma^2_{a,s} + I(A_{i,\tilde{k}} = 0)\sigma^2_{0,s}]$$
$$= 1.$$

If $k = k'$, $s = s'$ and $l \neq a'$

$$Cov(Z_{a,s,k}, Z_{a',s,k})$$
$$= \frac{1}{\sqrt{\frac{\sigma^2_{a,s}+\sigma^2_{0,s}}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \frac{\sigma^2_{a',s}+\sigma^2_{0,s}}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}}}} \frac{1}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \frac{1}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \sum_{\tilde{k}=1}^{k} \sum_{i=1}^{n_{\tilde{k}}} I(S_{i,\tilde{k}} = s) I(A_{i,\tilde{k}} = 0)\sigma^2_{0,s}$$
$$= \frac{1}{\sqrt{\frac{\sigma^2_{a,s}+\sigma^2_{0,s}}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}} \frac{\sigma^2_{a',s}+\sigma^2_{0,s}}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}}}} \frac{\sigma^2_{0,s}}{\sum_{\tilde{k}=1}^{k} n_{s,\tilde{k}}}$$
$$= \frac{\sigma^2_{0,s}}{\sqrt{(\sigma^2_{a,s} + \sigma^2_{0,s})(\sigma^2_{a',s} + \sigma^2_{0,s})}}.$$

If $k \neq k'$, $s = s'$, and $a = a'$

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

If $k \neq k'$, $l \neq a'$ and $s = s'$

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

□

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

S.2 Starting Values for Simulated Annealing Searches

The simulated annealing algorithm requires initial values to be input for all design parameters that it optimizes. These are given here.

For design class two, we used $\alpha_{1,1} = \alpha_{2,1} = 0.025$.

For design class 4 with $K = 4$, initial values were the following: equal alpha allocations for all stages and subpopulation combinations (i.e. $\alpha_{s,k} = 0.05/8$ for $s = 1, 2, k = 1, 2, 3, 4$), all futility boundaries set to -1, the timing of the interim analysis is set to when $10\%, 33\%, 67\%$, and 100% of the primary outcomes are observed. For each of design classes 1-4, we initialized the maximum sample size n to be $2250, 1890, 1950, 1950$, respectively. The first of these was selected based on a quick initial search over n to determine roughly what magnitude is required to satisfy the power constraints for design class one; values of n for the subsequent design classes were based on the optimal results for the previous classes. For example, for classes 3 and 4, the initial value $n = 1950$ was selected by slightly increasing the value $n = 1779$ that resulted from optimizing over design class two; this was based on our intuition that larger values of n are typically required in order to achieve a reduction in the expected sample size, when comparing single stage versus multiple stage designs.

For design class four with $K = 2$ or $K = 3$ stages, the initial values were the same as described above for $K = 4$ with the following exceptions: for $K = 2$ the analysis times were started at 10% and 100% of outcomes observed; for $K = 3$ the analysis times were started at 30%, 50%, and 100% of outcomes observed.

S.3 Test for Non-inferiority Following Superiority of Both Treatments in a Subpopulation

An added feature of the SMART-AV trial design is that if both treatments are found to be superior to the control for a subpopulation, then a non-inferiority test is conducted to compare the treatments (Stein et al., 2010). This section incorporates an additional non-inferiority test into our adaptive enrichment designs, while maintaining strong control of the familywise Type I error rate.

For subpopulation s and for a prespecified non-inferiority margin $\tau \in (0, 1]$, interest lies in evaluating if treatment $a = 1$ preserves more than $100 * \tau$ percent of the benefit of treatment $a = 2$ compared to the control $a = 0$. An advantage of this approach is that, since non-inferiority is tested only if superiority of both treatments compared to the control has already been established, the treatment effect comparing $a = 2$ versus control $a = 0$ has already been assessed within the trial, obviating the need to rely on historical data to estimate this treatment effect. The non-inferiority test for subpopulation s is only performed if (i) accrual continues for both treatments through stage K for that subpopulation and (ii) both $H_{1,s}, H_{2,s}$ are rejected using the original efficacy thresholds $\{u_{s,k}, z_{s,k} : s = 1, 2; k \leq K\}$ without reallocation of alpha from the other subpopulation.

Since the non-inferiority hypothesis test is only conducted for subpopulation s after the null hypotheses $H_{1,s}, H_{2,s}$ have been rejected, we assume below that $\mu_{a,s} - \mu_{0,s} > 0$ for each $a \in \{1, 2\}$. The inferiority null hypothesis for subpopulation s is defined as $H_s^{(Inf)}$: $\mu_{1,s} - \mu_{0,s} \leq \tau(\mu_{2,s} - \mu_{0,s})$ or equivalently $H_s^{(Inf)}$: $\mu_{1,s} - \tau\mu_{2,s} - (1 - \tau)\mu_{0,s} \leq 0$. For subpopulation s, define the standardized statistic for testing $H_s^{(Inf)}$ as

$$Z_s^{(Inf)} = \left\{\bar{Y}_{1,s,K} - \tau\bar{Y}_{2,s,K} - (1 - \tau)\bar{Y}_{0,s,K}\right\} \left\{ \frac{\sigma_{1,s}^2}{\sum_{k=1}^K n_{1,s,k}} + \tau^2 \frac{\sigma_{2,s}^2}{\sum_{k=1}^K n_{2,s,k}} + (1 - \tau)^2 \frac{\sigma_{0,s}^2}{\sum_{k=1}^K n_{0,s,k}} \right\}^{-1/2}.$$

Under $H_s^{(Inf)}$, the test statistic is asymptotically normally distributed with unit variance and mean at most 0 (Pigeot et al., 2003).

For each subpopulation $s \in \{1, 2\}$, we reject $H_s^{(Inf)}$ if $Z_s^{(Inf)} > \Phi^{-1}(1 - \alpha/2)$, where $\Phi^{-1}(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of the standard normal distribution. This ensures that the familywise Type I error rate for the optimized adaptive design (considering all 4 superiority null hypotheses $H_{a,s}$ and the 2 non-inferiority null hypotheses) is at most α.

Implementation requires specifying the non-inferiority margin τ. That is, specifying how much treatment effect reduction for treatment $a = 1$ compared to $a = 2$ is acceptable. This

choice is a clinical judgment and depends on what the benefits of treatment $a = 1$ compared to $a = 2$ are (e.g., in term of safety, side effects, or cost).

The correlation between the test statistics for non-inferiority and test for superiority in subpopulation $s = 1, 2$ is given by

$$Cov(Z_s^{(Inf)}, Z_{1,s,k}) = \frac{\frac{\sigma_{1,s}^2}{\sum_{k=1}^K n_{1,s,k}} + \frac{(1-\tau)\sigma_{0,s}^2}{\sum_{k=1}^K n_{0,s,k}}}{\sqrt{\frac{\sigma_{1,s}^2+\sigma_{0,s}^2}{\sum_{k=1}^K n_{1,s,k}}}\sqrt{\frac{\sigma_{1,s}^2}{\sum_{k=1}^K n_{1,s,k}} + \tau^2\frac{\sigma_{2,s}^2}{\sum_{k=1}^K n_{2,s,k}} + (1-\tau)^2\frac{\sigma_{0,s}^2}{\sum_{k=1}^K n_{0,s,k}}}}$$

$$Cov(Z_s^{(Inf)}, Z_{2,s,k}) = \frac{-\frac{\tau\sigma_{2,s}^2}{\sum_{k=1}^K n_{2,s,k}} + \frac{(1-\tau)\sigma_{0,s}^2}{\sum_{k=1}^K n_{0,s,k}}}{\sqrt{\frac{\sigma_{2,s}^2+\sigma_{0,s}^2}{\sum_{k=1}^K n_{2,s,k}}}\sqrt{\frac{\sigma_{1,s}^2}{\sum_{k=1}^K n_{1,s,k}} + \tau^2\frac{\sigma_{2,s}^2}{\sum_{k=1}^K n_{2,s,k}} + (1-\tau)^2\frac{\sigma_{0,s}^2}{\sum_{k=1}^K n_{0,s,k}}}}.$$

The mean of the test statistics $Z_s^{(Inf)}, s = 1, 2$ is given by

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

S.4 Proof of Theorem 4.1

For convenience of notation below, define $\tilde{z}_{s,k}$ and $\tilde{u}_{s,k}$ to equal $z_{s,k}$ and $u_{s,k}$, respectively, whenever no alpha reallocation is used.

Define a closed testing procedure using the following local tests:

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

- For $s \neq s', s, s' \in \{1,2\}$, any $a' \in \{1,2\}$, intersection test of $H_{1,s} \cap H_{2,s} \cap H_{a',s'}$: reject if $Z_{a,s,k} > u_{s,k}$ for at least one pair $(a,k), a \in \{1,2\}, k \in \{1,\ldots,K\}$ or if $Z_{a',s',k} > z_{s',k}$ for some $k \in \{1,\ldots,K\}$.

- Intersection test of all 4 null hypotheses: reject if $Z_{a,s,k} > u_{s,k}$ for at least one triple $(a,s,k), a, s \in \{1,2\}, k \in \{1,\ldots,K\}$.

For a set A the intersection hypothesis corresponding to A can only by rejected at stage k if all intersection hypothesis that include A are rejected at or before stage k.

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

If $\mathcal{D}_{ADAPT}$ rejects $H_{1,1}$, at least one of the following two statements is true: 1) there exists 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 $m \leq k$.

First we assume that statement 1) is true and show that all intersection hypothesis involving $H_{1,1}$ are rejected. Let k^* be the first stage satisfying $Z_{1,1,k^*} > \tilde{u}_{1,k^*}$.

- The intersection test $H_{1,1} \cap H_{2,1} \cap H_{1,2} \cap H_{2,2}$: If alpha reallocation is done from population 1 to population 2 before or at stage k^*, $Z_{a,1,k} > u_{1,k}$ for at least one $(a,k), a \in \{1,2\}, k \in \{1,\ldots,k^*\}$ pair. If alpha reallocation is done from population 2 to population 1 before or at stage k^*, $Z_{a,2,k} > u_{2,k}$ for at least one pair $(a,k), a \in \{1,2\}, k \in \{1,\ldots,k^*\}$. If no reallocation is done before or at stage k^*, $\tilde{u}_{s,k} = u_{s,k}$ for all combinations of $(s,k), s \in \{1,2\}, k \in \{1,\ldots,k^*\}$ and by 1) $Z_{1,1,k^*} > \tilde{u}_{1,k^*} = u_{s,k^*}$.

- The intersection test $H_{1,1} \cap H_{2,1} \cap H_{a',2}$ for $a' = 1,2$. If alpha reallocation is done from population 1 to population 2 before or at stage k^*, $Z_{a,1,k} > u_{1,k}$ for at least one $(a,k), a \in \{1,2\}, k \in \{1,\ldots,k^*\}$ pair. If alpha reallocation is done from population 2 to population 1 before or at stage k^*, $Z_{a',2,k} > z_{2,k}$ for at least one $k \in \{1,\ldots,k^*\}$.

If no reallocation is done before or at stage k^*, $\tilde{u}_{s,k} = u_{s,k}$ for all combinations of $(s,k) : s \in \{1,2\}, k \in \{1,\ldots,k^*\}$. Hence, if 1) holds $Z_{1,1,k^*} > \tilde{u}_{1,k^*} = u_{1,k^*}$.

- The intersection test $H_{1,2} \cap H_{2,2} \cap H_{1,1}$: If alpha reallocation is done from population 1 to population 2 before or at stage k^*, $Z_{1,1,k} > z_{1,k}$ for at least one $k \in \{1,\ldots,k^*\}$. If alpha reallocation is done from population 2 to population 1 before or at stage k^*, $Z_{a,2,k} > u_{2,k}$ for at least one pair $(a,k), a \in \{1,2\}, k \in \{1,\ldots,k^*\}$. If no reallocation is done before or at stage k^*, $\tilde{z}_{s,k} = z_{s,k}$ and $\tilde{u}_{s,k} = u_{s,k}$ for all combinations of $(s,k) : s \in \{1,2\}, k \in \{1,\ldots,k^*\}$. As $\tilde{u}_{1,k} \geq \tilde{z}_{1,k}$ for all $k \in \{1,\ldots,K\}$, it follows that if 1) holds $Z_{1,1,k^*} > \tilde{u}_{1,k^*} \geq \tilde{z}_{1,k^*} = z_{1,k^*}$.

- The intersection test $H_{1,1} \cap H_{a',2}$ for $a' \in \{1,2\}$: If alpha reallocation is done from population 1 to population 2 before or at stage k^*, $Z_{1,1,k} > z_{1,k}$ for at least one $k \in \{1,\ldots,k^*\}$. If alpha reallocation is done from population 2 to population 1 before or at stage k^*, $Z_{a',2,k} > z_{2,k}$ for at least one $k \in \{1,\ldots,k^*\}$. If no reallocation is done before or at stage k^*, $\tilde{u}_{1,k^*} \geq \tilde{z}_{1,k^*} = z_{1,k^*}$. Hence, from 1) $Z_{1,1,k^*} > z_{1,k}$.

- The intersection test $H_{1,1} \cap H_{2,1}$: Follows directly from 1).

- The intersection test $H_{1,1}$: Follows directly from 1) and $\tilde{u}_{1,k^*} \geq \tilde{z}_{1,k^*}$.

Now assume that statement 2) is correct. Let k^* be the first stage satisfying $Z_{1,1,k^*} > \tilde{z}_{1,k^*}$ and m^* be the first stage satisfying $Z_{2,1,m^*} > \tilde{u}_{1,m^*}$. By assumption 2) $m^* \leq k^*$.

- The intersection test $H_{1,1} \cap H_{2,1} \cap H_{1,2} \cap H_{2,2}$: If alpha reallocation is done from population 1 to population 2 before or at stage m^*, $Z_{a,1,k} > u_{1,k}$ for at least one $(a,k), a \in \{1,2\}, k \in \{1,\ldots,m^*\}$ pair. If alpha reallocation is done from population 2 to population 1 before or at stage m^*, $Z_{a,2,k} > u_{2,k}$ for at least one pair $(a,k), a \in \{1,2\}, k \in \{1,\ldots,m^*\}$. If no reallocation is done before or at stage m^*, $\tilde{u}_{1,m^*} = u_{1,m^*}$ and by 2) $Z_{2,1,m^*} > \tilde{u}_{1,m^*} = u_{1,m^*}$.

- The intersection test $H_{1,1} \cap H_{2,1} \cap H_{a',2}$ for $a' = 1, 2$. If alpha reallocation is done from population 1 to population 2 before or at stage m^*, $Z_{a,1,k} > u_{1,k}$ for at least one $(a, k), a \in \{1, 2\}, k \in \{1, \ldots, m^*\}$ pair. If alpha reallocation is done from population 2 to population 1 before or at stage m^*, $Z_{a',2,k} > z_{2,k}$ for at least one $k \in \{1, \ldots, m^*\}$. If no reallocation is done before or at stage m^*, $\tilde{u}_{s,m^*} = u_{s,m^*}$ for $s \in \{1, 2\}$. Hence, if 2) holds $Z_{2,1,m^*} > \tilde{u}_{1,m^*} = u_{1,m^*}$.

- The intersection test $H_{1,2} \cap H_{2,2} \cap H_{1,1}$: If alpha reallocation is done from population 1 to population 2 before or at stage k^*, $Z_{1,1,k} > z_{1,k}$ for at least one $k \in \{1, \ldots, k^*\}$. If alpha reallocation is done from population 2 to population 1 before or at stage k^*, $Z_{a,2,k} > u_{2,k}$ for at least one pair $(a, k), a \in \{1, 2\}, k \in \{1, \ldots, k^*\}$. If no reallocation is done, $\tilde{z}_{s,k^*} = z_{s,k^*}$ and by 2) $Z_{1,1,k^*} > \tilde{z}_{1,k^*} = z_{1,k^*}$.

- The intersection test $H_{1,1} \cap H_{a',2}$ for $a' \in \{1, 2\}$: If alpha reallocation is done from population 1 to population 2 before or at stage k^*, $Z_{1,1,k} > z_{1,k}$ for at least one $k \in \{1, \ldots, k^*\}$. If alpha reallocation is done from population 2 to population 1 before or at stage k^*, $Z_{a',2,k} > z_{2,k}$ for at least one $k \in \{1, \ldots, k^*\}$. If no reallocation is done before or at stage k^*, $\tilde{z}_{1,k^*} = z_{1,k^*}$. Hence, it follows from 2) that $Z_{1,1,k^*} > z_{1,k^*}$.

- The intersection test $H_{1,1} \cap H_{2,1}$: Follows directly from 2).

- The intersection test $H_{1,1}$: Follows directly from 2).

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

If $H_{1,1} \cap H_{1,2}$ is rejected at stage k and $\cap_{\{(a,s)=(1,1)\}} H_{a,s}$ is rejected at stage $k' \geq k$ then $\mathcal{D}_{ADAPT}$ rejectes $H_{1,1}$. This completes the proof that for a fixed $a, s \in \{1, 2\}$, $\mathcal{D}_{ADAPT}$ rejects $H_{a,s}$ if and only if the closed testing procedure rejects every intersection hypothesis involving $H_{a,s}$.

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

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

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

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

- Intersection of $H_{1,s} \cap H_{2,s}$: Both $H_{1,s}$ and $H_{2,s}$ are true. The probability of a Type I error in subpopulation $s \in \{1, 2\}$ is bounded above by

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

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

- Intersection test of $H_{a,s} \cap H_{a',s'}$ $s \neq s'$, with $s, s' \in \{1, 2\}$ and $a, a' \in \{1, 2\}$. Under the null of $H_{a,s}$ and $H_{a',s'}$ both being true, the Type I error is bounded above by

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

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

- For $s \neq s'$ with $s, s' \in \{1, 2\}$ and $a' \in \{1, 2\}$, intersection test of $H_{1,s} \cap H_{2,s} \cap H_{a',s'}$:

Under the null of all three hypothesis in the test being true, the Type I error is bounded from above by

$$\begin{aligned}
&\sum_{k=1}^{K} P(\max(Z_{1,s,k'}, Z_{2,s,k'}) \leq u_{s,k'} \text{ for all } k' < k, \max(Z_{1,s,k}, Z_{2,s,k}) > u_{s,k}) \\
&\quad + \sum_{k=1}^{K} P(Z_{a',s',k'} \leq z_{s',k'} \text{ for all } k' < k, Z_{a',s',k} > z_{s',k}) \\
&\quad \leq \sum_{s=1}^{2} \sum_{k=1}^{K} \alpha_{s,k} = \alpha,
\end{aligned}$$

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

- 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

$$\begin{aligned}
&\sum_{s=1}^{2} \sum_{k=1}^{K} P(\max(Z_{1,s,k'}, Z_{2,s,k'}) \leq u_{s,k'} \text{ for all } k' < k, \max(Z_{1,s,k}, Z_{2,s,k}) > u_{s,k}) \\
&\quad \leq \sum_{s=1}^{2} \sum_{k=1}^{K} \alpha_{s,k} = \alpha,
\end{aligned}$$

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

As all intersection test have Type I error rate at most α, the closed testing principle implies that $\mathcal{D}_{ADAPT}$ controls the familywise Type I error rate at a level α.

In Section 8, an alternative way of calculating efficacy boundaries $u_{s,K}, s = 1, 2, k = 1, \ldots, K$ is presented. It follows from the construction of the efficacy boundaries (see equation (4)) that all intersection tests using these efficacy boundaries control the familywise Type I error rate. Hence, $\mathcal{D}_{ADAPT}$ implemented using efficacy boundaries calculated using equation (4) strongly control the familywise Type I error.