1-13-2015

Adaptive, Group Sequential Designs that Balance the Benefits and Risks of Wider Inclusion Criteria

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mrosenbl@jhsph.edu

Brandon S. Luber
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, bluber@jhsph.edu

Richard E. Thompson
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rthompso@jhsph.edu

Daniel F. Hanley
Johns Hopkins School of Medicine, dhanley@jhmi.edu

Adaptive, Group Sequential Designs that Balance the Benefits and Risks of Wider Inclusion Criteria

Michael Rosenblum,* Brandon Luber,* Richard E. Thompson,* and Daniel Hanley†

January 13, 2015

Abstract

We propose a new class of adaptive randomized trial designs aimed at gaining the advantages of wider generalizability and faster recruitment, while mitigating the risks of including a population for which there is greater a priori uncertainty. Our designs use adaptive enrichment, i.e., they have preplanned decision rules for modifying enrollment criteria based on data accrued at interim analyses. For example, enrollment can be restricted if the participants from predefined subpopulations are not benefiting from the new treatment. To the best of our knowledge, our designs are the first adaptive enrichment designs to have all of the following features: the multiple testing procedure fully leverages the correlation among statistics for different populations; the familywise Type I error rate is strongly controlled; for outcomes that are binary, normally distributed, or Poisson distributed, the decision rule and multiple testing procedure are functions of the data only through minimal sufficient statistics. The advantage of relying solely on minimal sufficient statistics is that not doing so can lead to losses in power. Our designs incorporate standard group sequential boundaries for each population of interest; this may be helpful in communicating our designs, since many clinical investigators are familiar with such boundaries, which can be summarized succinctly in a single table or graph. We demonstrate these adaptive designs in the context of a Phase III trial of a new treatment for stroke, and provide user-friendly, free software implementing these designs.

1 Introduction

This work is motivated by challenges that arose in designing a Phase III trial of a new treatment for stroke. However, we expect similar challenges to occur in other domains. Our goal in designing the Phase III trial was to evaluate a new surgical treatment for intracerebral hemorrhage (ICH), called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage Evacuation, abbreviated as MISTIE [1]. The treatment showed promise in a completed Phase II trial, which enrolled individuals with ICH who had small or no intraventricular hemorrhage (IVH) at baseline, called "small IVH" participants. The study investigators debated whether to expand the inclusion criteria of the proposed Phase III trial to also enroll "large IVH" participants, defined as having baseline IVH volume at least 10ml or requiring a catheter for intracranial pressure monitoring. Based on their understanding of brain hemorrhage, the investigators conjectured that the new treatment would benefit large IVH participants. Advantages of including them in the Phase III trial are that it would answer a question relevant to a larger population, and it would increase the enrollment rate. However, since there were few participants with large IVH who had ever undergone the MISTIE procedure, to include this population in the Phase III trial would pose a substantial risk. Specifically, if the treatment only benefits those with small IVH, then a trial targeting the larger population may have low power to detect this.

Motivated by the above issues, we propose a new class of adaptive enrichment designs that is the first, to the best of our knowledge, to have all of the following properties: the multiple testing procedure fully

[*] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, U.S.A.

[†] Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, U.S.A.

leverages the correlation among statistics for different populations; the familywise Type I error is controlled in the strong sense [2]; for outcomes that are binary, normally distributed, or Poisson distributed, the decision rule and multiple testing procedure are functions of the data only through minimal sufficient statistics at each interim analysis. The last property is important since not using minimal sufficient statistics can lead to losses in precision and power. According to Rice [3, p. 284], "If an estimator is not a function of a sufficient statistic, it can be improved." This property is especially relevant in our context, since a major criticism of many adaptive enrichment designs is that they do not use minimal sufficient statistics [4]. We focus on binary, normally distributed, or Poisson distributed outcomes since these are cases where simple, minimal sufficient statistics exist. Our designs also have the above properties under more general conditions given in Section 3.5.

Others have proposed adaptive enrichment designs that have some, but not all, of the above features. Adaptive designs based on the p-value combination approach [5, 6, 7, 8, 9, 10], on the conditional error function approach [11], or on the approaches in [12, 13, 14], do not use data only through minimal sufficient statistics, as we describe in Section 3.5. The methods of [15, 16] do not fully leverage the covariance between statistics for different populations; they use, e.g, the Hochberg multiple testing correction [17]. The method of [18], which has different goals than here, uses minimal sufficient statistics but is not proved to strongly control the familywise Type I error rate.

Each of our adaptive designs uses group sequential boundaries [19] in its decision rules and hypothesis tests. For example, boundaries can be based on those of O'Brien and Fleming [20]. This may be helpful in communicating our designs, since many clinical investigators are familiar with group sequential boundaries. Computation of such boundaries is essentially instantaneous since they are based on the multivariate normal distribution function that can be computed by standard statistical software.

In the context of the MISTIE trial, we searched over a large set of our new designs to select one that minimizes expected sample size subject to constraints on power. This new adaptive design is compared to standard designs and to other adaptive designs, in scenarios relevant to the MISTIE trial. The expected sample size averaged over the scenarios of primary interest is 635 for our adaptive design, compared to 889 for the best alternative design considered. Based on the estimated cost per patient in the MISTIE Phase III trial of $29,000, the reduced expected cost from our adaptive design would be roughly $7 million.

A limitation of our designs is that each participant's outcome is assumed to be observed relatively soon after enrollment. We discuss future research directions for extending these designs to handle delayed outcomes in Section 8.

The goals of the MISTIE trial are presented in the following section. Our general statistical problem is defined in Section 3. In Section 4, we present a new class of adaptive enrichment designs. These are tailored to the goals of the MISTIE trial in Section 5, and compared to alternative designs in Section 6. In Section 7, we describe our software package that enables users to compare the performance of our adaptive designs versus standard designs. Limitations and directions for future research are discussed in Section 8.

2 Application: Planning the MISTIE Phase III Trial

The aim is to assess whether the MISTIE surgical treatment is superior to the standard of care. The primary outcome is a participant's degree of disability on the modified Rankin Scale (mRS). A successful outcome is defined to be mRS ≤ 3. Define the average treatment effect to be the difference between the probability of a successful outcome under assignment to MISTIE treatment versus standard of care. Based on the MISTIE Phase II randomized trial, which enrolled 96 small IVH participants and no large IVH participants, the estimated average treatment effect is 0.12 [95% CI: (-0.07, 0.29)].

The clinical investigators indicated the following three scenarios to consider in planning the MISTIE Phase III trial: the average treatment effect is (a) 12.5% for both the small IVH and large IVH subpopulations; (b) 12.5% for the small IVH subpopulation and 0% for the large IVH subpopulation; (c) 0% for both subpopulations. The design goals are:

(i) $\geq 80\%$ power to detect an average treatment benefit for the combined population, in scenario (a);

(ii) $\geq 80\%$ power to detect an average treatment benefit for the small IVH subpopulation, in scenario (b);

(iii) strong control of the familywise Type I error rate at level 2.5%. (We use 2.5% rather than 5%, because we consider one-sided hypotheses.)

The design goals are asymmetric, i.e., there is a power requirement for detecting a benefit for the small IVH subpopulation, while there is no analogous requirement for the large IVH subpopulation. This is due to the stronger prior evidence, from the MISTIE Phase II trial, for a potential benefit in the small IVH population. Focusing on the overall population and a single subpopulation is not uncommon in adaptive enrichment designs, e.g., [15, 8, 9, 10], [14, Section 5]. Our general method can also be applied to achieve symmetric design goals, as discussed in Section 8.

The motivation for considering adaptive enrichment designs is that standard randomized trial designs do not allow early stopping of subpopulations for futility. Such standard designs may continue to enroll large IVH participants despite strong evidence of no benefit to them, leading to inefficiency, and unnecessarily exposing large IVH participants to a non-efficacious treatment.

3 General Problem Definition

3.1 Hypotheses and Assumptions

Consider two subpopulations that partition the overall population, defined in terms of baseline measurements. For example, in the MISTIE trial, subpopulation 1 represents those with small IVH, and subpopulation 2 represents those with large IVH. Let π_s denote the proportion of the combined population in subpopulation s, for $s \in \{1, 2\}$; $\pi_1 + \pi_2 = 1$.

Each participant i in stage k contributes data $(S_{i,k}, A_{i,k}, Y_{i,k})$, where $S_{i,k}$ is the subpopulation (1 or 2); $A_{i,k}$ is an indicator of being randomized to treatment ($A_{i,k} = 1$) versus control ($A_{i,k} = 0$); and $Y_{i,k}$ is a real-valued outcome (which may be binary, count, or continuous). We assume at each stage, for each subpopulation, half of the participants are randomly assigned to treatment; this can be approximately achieved by using block randomization stratified by subpopulation.

In designs with preplanned rules to modify enrollment criteria based on prior stage data, subpopulation membership in a given stage depends on data from earlier stages. We assume that conditioned on the subpopulation membership and study arm assignments of all participants in stage k, the outcome $Y_{i,k}$ is a random draw from an unknown distribution Q_{sa}, for $s = S_{i,k}, a = A_{i,k}$, independent of the data on all other participants in stages 1 through k. Denote the vector of unknown distributions by $Q = (Q_{10}, Q_{11}, Q_{20}, Q_{21})$. Let $\mu(Q_{sa})$ and $\sigma^2(Q_{sa})$ denote the mean and variance of Q_{sa}, respectively, for each $s \in \{1, 2\}, a \in \{0, 1\}$. We assume that the unknown distribution Q is in the class $\mathcal{Q}$ defined to be all vectors $(Q_{10}, Q_{11}, Q_{20}, Q_{21})$ that satisfy the following integrability condition for a fixed $C > 3$:

$$E_{Q_{sa}} \left[\{Y - \mu(Q_{sa})\} / \sigma(Q_{sa}) \right]^4 \leq C, \tag{1}$$

for each $s \in \{1, 2\}, a \in \{0, 1\}$. This guarantees the joint distribution of z-statistics for each population, given in Section 3.4, converges uniformly to a multivariate normal distribution as in [21]. Such convergence is generally required to control the familywise Type I error rate as defined in Section 3.3. For binary outcomes, condition (1) is equivalent to $\Pr(Y = 1 | S = s, A = a)$ being uniformly bounded away from 0 and 1. For count outcomes that follow a Poisson distribution, (1) is equivalent to $\mu(Q_{sa})$ being uniformly bounded away from 0. For continuous outcomes that follow a normal distribution with arbitrary mean and positive variance, (1) holds if and only if $C \geq 3$. For the class of all distributions Q for which each Q_{sa} has support in $[-M, M]$ and variance at least $\tau > 0$, (1) holds for $C > (2M)^4/\tau^2$. The general condition (1) allows us to simultaneously handle these different classes of distributions.

For each subpopulation $s \in \{1, 2\}$, define the average treatment effect as $\Delta_s = \mu(Q_{s1}) - \mu(Q_{s0})$. Define the average treatment effect for the combined population as $\Delta_C = \pi_1 \Delta_1 + \pi_2 \Delta_2$. The null hypotheses of primary interest are

$$H_{01} = \{Q \in \mathcal{Q} : \Delta_1 \leq 0\}; \qquad H_{0C} = \{Q \in \mathcal{Q} : \Delta_C \leq 0\}.$$

This pair of null hypotheses (or versions with inequalities replaced by equalities) is also considered in related work on adaptive enrichment designs [15, 8, 9, 10], [14, Section 5].

3.2 Designs

Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on data accrued at an interim analysis [15]. Our adaptive enrichment designs in Section 4 have such a rule for potentially switching to enroll only one subpopulation. Our designs also incorporate group sequential testing, i.e., at each interim analysis, hypotheses are tested and enrollment may be completely stopped. For conciseness, we refer to designs that incorporate features of adaptive enrichment designs and group sequential designs simply as "adaptive designs". We refer to standard, group sequential designs, defined to be group sequential designs where the only decision about enrollment at each interim analysis is whether to continue the trial or stop it entirely, as "standard designs".

For each adaptive design, the following are defined before the trial starts: $K \geq 1$, the maximum number of stages; $n_{s,k}$, the number of participants from subpopulation $s \in \{1, 2\}$ to be enrolled during stage $k \leq K$, assuming enrollment has not been stopped for that subpopulation at a previous stage; k^*, the last stage at which subpopulation 2 can be enrolled. The reason that we include the design parameter k^* is that in some contexts, setting $k^* < K$ leads to lower maximum sample sizes and lower expected samples sizes than comparable rules with $k^* = K$, as described at the end of Section 6.3; such designs are still adaptive enrichment designs since enrollment of subpopulation 2 may be stopped before stage k^*, based on accrued data. Let $n_k = n_{1,k} + n_{2,k}$ for each $k \leq K$. By construction, for any stage $k > k^*$, we have $n_{2,k} = 0$. The maximum total sample size is $n = \sum_{k=1}^{K} n_k$. At each stage in which both subpopulations are enrolled, we assume the ratio of subpopulation 1 participants to subpopulation 2 participants equals the ratio of the corresponding subpopulation sizes π_1/π_2; i.e., for any stage $k \leq k^*$, we have $n_{s,k} = \pi_s n_k$ for each $s \in \{1, 2\}$. If the combined population is enrolled during stage k (which only occurs for $k \leq k^*$), then $n_{s,k}$ are enrolled from each subpopulation $s \in \{1, 2\}$; if only subpopulation 1 is enrolled during stage k, then $n_{1,k}$ are enrolled.

In an adaptive design, at the end of each stage $k < K$, a prespecified decision rule is used to determine enrollment for the next stage. Prespecified decision rules are generally required by the U.S. Food and Drug Administration (FDA) for Phase III adaptive designs of drugs and biologics [22]. We assume all data from participants enrolled at or before stage k are available at this interim analysis. The allowed decisions are the following: if $k < k^*$ and enrollment in stage k was from the combined population, then enrollment in stage $k + 1$ can be from the combined population, from subpopulation 1, or the trial can be stopped; else, if $k \geq k^*$ or enrollment has already been restricted to subpopulation 1, enrollment in stage $k + 1$ can be from subpopulation 1 or the trial can be stopped. We do not allow restarting enrollment of a subpopulation once it has been stopped. We define a decision rule to be a function from the cumulative data collected at or before each stage k to the set of allowed decisions for stage $k + 1$ enrollment. We assume this function is measurable, which is important in adaptive designs as discussed by Liu *et al.* [23].

3.3 Strong Control of the Familywise Type I Error Rate

The familywise Type I error rate is the probability that at least one true null hypothesis is rejected. Control of the familywise Type I error rate is generally required by the U.S. FDA and the European Medicines Agency for confirmatory randomized trials [24]. A multiple testing procedure is said to strongly control the familywise Type I error rate at level α if for *any* data generating distribution $Q \in \mathcal{Q}$, the familywise Type I error rate is at most α [2].

We consider a sequence of per-stage sample sizes that go to infinity, holding constant k^* and the proportions of the total sample size allocated to each stage $(r_1, \ldots, r_K) = (n_1/n, \ldots, n_K/n)$. For given $K, k^*, r_1, \ldots, r_K$, we say a trial design D strongly controls the asymptotic, familywise Type I error rate at level α if

$$\limsup_{n\to\infty} \sup_{Q\in\mathcal{Q}} \Pr_{Q,D,n}(\text{at least one true null hypothesis is rejected}) \leq \alpha, \quad (2)$$

where $\Pr_{Q,D,n}$ represents the probability under distribution Q, design D, and per-stage sample sizes $(nr_1, \ldots, nr_K)$. We prove all of our designs satisfy (2) for $\alpha = 0.025$. The above definition of strong control of the familywise Type I error is asymptotic, as sample size goes to infinity. Even in the simpler setting of a standard design testing a single null hypothesis with a z-test or t-test, if the outcome is binary then the Type I error is only controlled asymptotically. We also examined the familywise Type I error rate of our designs at realistic sample sizes, using simulations described in Section 6. For conciseness, we refer to "strong control of the asymptotic, familywise Type I error rate at level α" as "strong control of the familywise Type I error rate."

3.4 Statistics

At each interim analysis $k \leq k^*$, if the combined population is enrolled through stage k, define the (cumulative) z-statistic for the combined population, $Z_{C,k}$, to be the standardized difference between sample means comparing treatment to control for all participants enrolled during stages 1 through k:

$$Z_{C,k} = \left\{ \frac{\sum_{k'=1}^{k}\sum_{i=1}^{n_{k'}} Y_{i,k'}A_{i,k'}}{\sum_{k'=1}^{k}\sum_{i=1}^{n_{k'}} A_{i,k'}} - \frac{\sum_{k'=1}^{k}\sum_{i=1}^{n_{k'}} Y_{i,k'}(1-A_{i,k'})}{\sum_{k'=1}^{k}\sum_{i=1}^{n_{k'}}(1-A_{i,k'})} \right\} \mathrm{se}_{C,k}^{-1}, \tag{3}$$

where $\mathrm{se}_{C,k} = \left[\sum_{s=1}^{2}\pi_s\left\{\sigma^2(Q_{s0})+\sigma^2(Q_{s1})\right\} / \sum_{k'=1}^{k}(n_{k'}/2)\right]^{1/2}$ is the standard error of the quantity in curly braces above. If the combined population is not enrolled through stage k, then $Z_{C,k}$ is undefined. The (cumulative) z-statistics $Z_{1,k}$ for subpopulation 1 and $Z_{2,k}$ for subpopulation 2 are defined analogously in Appendix A.1 of the Supplementary Materials, except restricted to participants in the corresponding subpopulations. The statistic $Z_{1,k}$ is defined for each $k \leq K$, while $Z_{2,k}$ is only defined for $k \leq k^*$ since subpopulation 2 enrollment never goes beyond stage k^*.

The joint distribution of $\mathbf{Z}_C = \{Z_{C,k}\}_{k=1}^{k^*}$ has the canonical covariance structure of a standard group sequential design [19, Chapter 3]. The same holds for $\mathbf{Z}_1 = \{Z_{1,k}\}_{k=1}^{K}$ and $\mathbf{Z}_2 = \{Z_{2,k}\}_{k=1}^{k^*}$. Let Σ denote the covariance matrix of $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$, given in Appendix A.1 of the Supplementary Materials.

3.5 Minimal Sufficient Statistics

Consider any adaptive design D as defined in Section 3.2. First, we examine the cases where the outcome is binary, or is a count with Poisson distribution. Then there is a single unknown parameter for each Q_{sa}. At the end of stage k, the following are minimal sufficient statistics: for each $s \in \{1, 2\}$ and $a \in \{0, 1\}$, the total number enrolled and the sample mean of the outcome, based on all data from subpopulation s and arm a enrolled at or before stage k. For the case where the outcome is normally distributed, minimal sufficient statistics consist of the aforementioned statistics plus the sample variances. (These claims are proved in Appendix B of the Supplementary Materials.) In each case, the z-statistics $Z_{1,k}, Z_{2,k}, Z_{C,k}$, with standard errors computed using sample variances in place of true variances $\sigma^2(Q_{sa})$, are functions of the data only through minimal sufficient statistics. Our adaptive designs in the next section use only these statistics in the decision rule and multiple testing procedure at the end of each stage k.

The p-value combination approach does not use the data only through minimal sufficient statistics at each interim analysis. This is because designs based on the p-value combination approach require a test of the intersection $H_{01} \cap H_{0C}$ based on a weighted combination of statistics across stages, which can involve contributions from different populations at different stages. These statistics can lead to inefficiency as pointed out by Emerson [4].

Other common approaches do not use the data solely through minimal sufficient statistics at each interim analysis. This is the case for designs using the conditional error function approach of Proschan and Hunsberger [25], such as the designs of Friede *et al.* [11], since the rejection thresholds at interim analyses depend on the conditional error computed from previous stages, thereby using more than the minimal sufficient statistics at the current stage. The method of Stallard [12] uses the maximum score statistic over different populations at each stage, and combines these maxima across stages; when there are multiple stages at which an enrollment modification is considered, their procedure does not use data only through minimal sufficient statistics. As acknowledged by Stallard [12, p. 796], since different populations can contribute to this overall statistic at different stages, their method is conservative. The method of Rosenblum and van der Laan [13] does not use data solely through minimal sufficient statistics when enrollment is restricted to a single population for stage 2.

For clarity, we focus on outcomes that are binary-valued, are count-valued with Poisson distribution, or are continuous-valued with normal distribution. However, in Appendix B of the Supplementary Materials, we generalize the adaptive enrichment designs from Section 4 to handle outcomes Y with distribution belonging to any exponential family (also called exponential class) [26]; this includes the following distributions, among others: binomial, negative binomial, exponential, Pareto, Weibull, Laplace, chi-squared, gamma, beta, etc. The only change to the adaptive designs is that z-statistics are replaced by a generalization involving standardized sample means of the sufficient statistic from the corresponding exponential family. Under regularity conditions given in Appendix B of the Supplementary Materials, the resulting adaptive enrichment designs have the desirable properties listed in the Abstract, i.e., they use data only through minimal sufficient statistics, strongly control the familywise Type I error rate, and fully leverage the correlation among the generalized z-statistics.

4 Proposed Class of Adaptive Designs for Testing H_{0C}, H_{01}

Efficacy boundaries for the combined population and subpopulation 1 are denoted by $\mathbf{u}_C = \{u_{C,k}\}_{k=1}^{k^*}$ and $\mathbf{u}_1 = \{u_{1,k}\}_{k=1}^{K}$, respectively. Futility boundaries are denoted by $\mathbf{l}_1 = \{l_{1,k}\}_{k=1}^{K}$ and $\mathbf{l}_2 = \{l_{2,k}\}_{k=1}^{k^*}$. Each design takes $(\mathbf{u}_C, \mathbf{u}_1, \mathbf{l}_1, \mathbf{l}_2, \pi_1, K, k^*, \{n_{s,k} : s \in \{1,2\}, k \leq K\})$ as input prior to the trial. Based on z-statistics at the end of each stage, the rule below determines which hypotheses (if any) to reject and which subpopulations (if any) to enroll in the next stage. Enrollment is initially from the combined population.

Adaptive Designs for Testing $\{H_{0C}, H_{01}\}$. At each interim analysis $k \leq K$:

1. (Assess Efficacy for H_{0C} and H_{01}) This step is only done if the combined population was enrolled during stage k. If $Z_{C,k} > u_{C,k}$ or $Z_{1,k} > u_{1,k}$, then stop the trial and reject H_{0j} for each $j \in \{C, 1\}$ for which $Z_{j,k} > u_{j,k}$.

2. (Assess Futility of Entire Trial) Else, if $Z_{1,k} \leq l_{1,k}$, stop the trial and reject nothing.

3. (Decide Whether to Stop Subpopulation 2 Enrollment) Else, if $k = k^*$ or $Z_{2,k} \leq l_{2,k}$, stop subpopulation 2 enrollment and at each stage $k' > k$: enroll only subpopulation 1, and:

   (a) If $Z_{1,k'} > u_{1,k'}$, reject H_{01} and stop the trial.

   (b) Else, if $Z_{1,k'} \leq l_{1,k'}$, stop the trial and reject nothing.

   (c) Else, if $k' < K$, continue enrolling from subpopulation 1 until analysis $k' + 1$.

4. Else, continue enrolling the combined population unless the end of stage K has been reached.

For any $\mathbf{u} = (\mathbf{u}_C, \mathbf{u}_1)$ and $\mathbf{l} = (\mathbf{l}_1, \mathbf{l}_2)$, let $D(\mathbf{u}, \mathbf{l})$ denote the above design using these boundaries. We always set $l_{1,K} = u_{1,K}$ and $l_{2,k^*} = \infty$ (since subpopulation 2 enrollment always stops by the end of stage k^*).

The entire trial is stopped for futility in step 2 if $Z_{1,k} \leq l_{1,k}$. This reflects the prior belief that if the treatment benefits at least one subpopulation, it will benefit subpopulation 1; this was the motivation from the MISTIE trial for considering the null hypotheses H_{01}, H_{0C}, and also may be appropriate, e.g., in trials of biomarker positive and biomarker negative participants as considered by Wang *et al.* [15] and Boessen *et al.* [10]. However, the above designs are flexible in that arbitrary futility stopping rules can be used in place of steps 2 and 3b, as described later in this section.

Throughout, we treat the futility boundaries $\mathbf{l}$ as nonbinding, i.e., we require strong control of the familywise Type I error rate as in (2) even if futility stopping based on $\mathbf{l}$ is ignored. Nonbinding futility boundaries are typically preferred by the FDA as described by Liu and Anderson [27]. Define the special set of futility boundaries $\bar{\mathbf{l}}$ to represent no early futility stopping, i.e., $\bar{l}_{1,k} = -\infty$ for all $k < K$ and $\bar{l}_{2,k} = -\infty$ for all $k < k^*$. We note that, as specified in Section 3.2, enrollment of subpopulation 2 always stops by the end of stage k^*; this rule cannot be ignored.

We next describe a method for selecting boundaries $\mathbf{u}, \mathbf{l}$ such that the design $D(\mathbf{u}, \mathbf{l})$ is guaranteed to strongly control the familywise Type I error rate. Below, for clarity of presentation, we assume the covariance matrix Σ and subpopulation 1 proportion π_1 are known. However, in practice, these will generally be unknown. In Section 6.4, we examine the impact of using estimates of Σ and π_1. In all the scenarios we considered, our adaptive designs have similar performance as in the setting where variances and π_1 are assumed known.

The theorem below gives the following useful property: for any adaptive design $D(\mathbf{u}, \mathbf{l})$, to verify it strongly controls the familywise Type I error rate, it suffices to verify familywise Type I error control for $D(\mathbf{u}, \bar{\mathbf{l}})$ under the null hypothesis $H_0 = \{Q \in \mathcal{Q} : \Delta_1 = \Delta_2 = 0\}$. Though this property is often easy to prove for standard (non-adaptive) trials involving a single null hypothesis, it is not automatic in our setting of multiple hypotheses and potential enrollment adaptations. As described below, this property allows fast and precise computation of boundaries $\mathbf{u}, \mathbf{l}$ that are guaranteed to control the familywise Type I error rate.

Define $\mathbf{Z}' = \{Z'_{C,k}\}_{k=1}^{k^*} \bigcup \{Z'_{1,k}\}_{k=1}^{K}$ to be a multivariate normal family of random variables with all components having zero mean, and covariance matrix equal to the restriction of Σ to statistics $\{Z_{C,k}\}_{k=1}^{k^*} \bigcup \{Z_{1,k}\}_{k=1}^{K}$. For any efficacy boundaries $\mathbf{u}$, define

$$\alpha_0(\mathbf{u}, \Sigma) = \Pr_\Sigma \left\{ \left(\text{for at least one } k \le k^*, Z'_{C,k} > u_{C,k}\right) \bigcup \left(\text{for at least one } k \le K, Z'_{1,k} > u_{1,k}\right) \right\}. \qquad (4)$$

The above quantity is the asymptotic familywise Type I error rate under the null hypothesis $H_0 = \{Q \in \mathcal{Q} : \Delta_1 = \Delta_2 = 0\}$ for the design $D(\mathbf{u}, \bar{\mathbf{l}})$. In Appendix A of the Supplementary Materials, we prove the following theorem:

Theorem 1: Consider any $\alpha \in (0, 1)$, any efficacy boundaries $\mathbf{u}$ such that $\alpha_0(\mathbf{u}, \Sigma) \le \alpha$, and any futility boundaries $\mathbf{l}$. The adaptive design $D(\mathbf{u}, \mathbf{l})$ strongly controls the familywise Type I error rate at level α.

Therefore, to prove strong control of the familywise Type I error rate for the adaptive design $D(\mathbf{u}, \mathbf{l})$ at level α, it suffices to compute $\alpha_0(\mathbf{u}, \Sigma)$ and show it is at most α. For any given $\mathbf{u}$ and Σ, this can be computed by a single evaluation of the multivariate normal distribution function, which is implemented in the mvtnorm package in R [28]. For Σ known, a simple way to select $\mathbf{u}$ that satisfies $\alpha_0(\mathbf{u}, \Sigma) \le \alpha$ is to set $\mathbf{u}$ proportional to a standard set of group sequential boundaries. Binary search can be used to find the smallest proportionality constant such that $\alpha_0(\mathbf{u}, \Sigma) \le \alpha$ is satisfied. For the case of Σ unknown and estimated at each stage, we extend the error-spending approach [29, 30] to our context involving multiple hypotheses, as described in Section 6.4.

Consider a set of efficacy boundaries $\mathbf{u}$ satisfying the condition in Theorem 1. A benefit of such boundaries is that one can switch to only enrolling subpopulation 1 for any reason, and still strong control of the familywise Type I error rate holds; such a switch could be warranted, for example, if the adverse event rates are high for only subpopulation 2. A second benefit of such boundaries is that the above designs can be generalized to use arbitrary futility stopping rules in steps 2 and 3b, and still strong control of the familywise Type I error rate holds. A third benefit of such boundaries is that the above designs can be generalized to continue after a single null hypothesis is rejected in order to test the remaining null hypothesis, and still strong control of the familywise Type I error rate holds; this feature is discussed in Section 8. We prove the above claims in Appendix A of the Supplementary Materials.

We always select efficacy boundaries $\mathbf{u}$ that satisfy the criterion $\alpha_0(\mathbf{u}, \Sigma) = \alpha$. This criterion fully leverages the correlation Σ among z-statistics for different populations, which is built into (4) through the joint distribution of $\mathbf{Z}'$. Efficacy boundaries satisfying this criterion cannot be uniformly improved, that is, a reduction in any of the efficacy boundaries would lead to an asymptotic familywise Type I error rate strictly greater than α at H_0 if all futility boundaries are ignored (which is an important case since we treat futility boundaries as nonbinding). In contrast, designs using multiple testing procedures of Bonferroni, Holm [31], or Hochberg [17], none of which take Σ as input, do not exhaust the level in this sense, and so may sacrifice power.

5 Selecting an Efficient Adaptive Design Tailored to the MISTIE Trial Goals

Consider the goals of the MISTIE Phase III trial from Section 2. Let subpopulation 1 denote those with small IVH at baseline, and subpopulation 2 denote the complementary subpopulation, i.e., those with large IVH at baseline. The outcome is binary valued, with 1 denoting a successful outcome. We assume the following based on prior studies [32]:

$$\pi_1 = 1/3; \qquad \mu(Q_{10}) = 0.25; \qquad \mu(Q_{20}) = 0.2. \tag{5}$$

Designs with $K = 5$ stages are considered. We generated a list of adaptive designs from the class in Section 4, each defined by sample sizes $\{n_{s,k} : s \in \{1,2\}, k \leq K\}$ and boundaries $(\mathbf{u},\mathbf{l})$ such that $D(\mathbf{u},\mathbf{l})$ satisfies goals (i)-(iii) and $\alpha_0(\mathbf{u},\Sigma) = 0.025$. We then selected the design, denoted D_{ADAPT}, that minimizes the average of the expected sample sizes over scenarios (a)-(c). We explain the search procedure below, but first give the results.

The boundaries and sample sizes for D_{ADAPT} are given in Table 1. In each of stages $k = 1$ through $k = 3$, $n_k = 270$ participants are enrolled from the combined population, unless enrollment is restricted or the trial is stopped. If enrollment is restricted to subpopulation 1 after some stage $k < 3$, then $270 * \pi_1 = 270/3 = 90$ are enrolled from subpopulation 1 in each subsequent stage $k' \leq 3$ for which the trial continues. The design always stops enrolling from subpopulation 2 by the end of stage 3. In each of stages $k = 4$ and $k = 5$, unless the trial stops early, $n_k = 186$ participants are enrolled from subpopulation 1. The maximum total sample size is 1182. However, the expected sample size in each of scenarios (a)-(c), is 645, 737, 522, respectively. The performance of D_{ADAPT} under a variety of scenarios, which includes scenarios (a)-(c), is shown in Table 4 and discussed in Section 6.

A feature of our adaptive designs is that their boundaries can be displayed in a simple plot, similar to standard, group sequential boundaries. The boundaries for D_{ADAPT} are displayed in Figure 1.

We next define our search algorithm, leading to the design D_{ADAPT}. For each $k^* \in \{1,2,3,4\}$, we consider vectors of per-stage sample sizes $(n_1,\ldots,n_K)$ where in each stage $k \leq k^*$, n_k is a common value denoted by $n^{(1)}$, and in each stage $k > k^*$, n_k is a common value denoted by $n^{(2)}$. Similarly, for $k^* = K = 5$, for each $\tilde{k} \in \{1,\ldots,5\}$ we consider per-stage sample sizes $(n_1,\ldots,n_K)$ where each n_k equals a common value $n^{(1)}$ up through stage $\tilde{k}$, and then all subsequent per-stage sample sizes equal a common value $n^{(2)}$.

We use efficacy and futility boundaries that are generalizations of the boundaries of O'Brien and Fleming [20] to our setting of unequal, per-stage sample sizes. For proportionality constants e_C, e_1, f_1, f_2, which are selected using the algorithm described below, boundaries are set as follows:

i. Efficacy boundaries: $u_{C,k} = e_C(k/k^*)^{-1/2}$ for $k \leq k^*$; $u_{1,k} = e_1 \left\{ \sum_{k'=1}^{k} n_{1,k'} \Big/ \sum_{k'=1}^{K} n_{1,k'} \right\}^{-1/2}$ for $k \leq K$. (6)

ii. Futility boundaries: $l_{1,k} = f_1\{k/(K-1)\}^{-1/2}$ for $k \leq K-1$; $l_{2,k} = f_2\{k/(k^*-1)\}^{-1/2}$ for $k \leq k^*-1$,

and we set $l_{1,K} = u_{1,K}, l_{2,k^*} = \infty$.

Throughout, power and expected sample sizes are computed assuming futility boundaries are adhered to. We considered a range of pairs (f_1, f_2) given in Appendix D.1 of the Supplementary Materials. For each $k^* \in \{1,2,3,4\}$ and (f_1, f_2), we considered each $n^{(1)}$ in increments of 10 participants from 10 to 400. We then computed the minimum $n^{(2)}$, if any exists, such that goals (i)-(iii) are achieved and $\alpha_0(\mathbf{u},\Sigma) = 0.025$; in such cases we computed the corresponding values of e_C, e_1. Analogous computations were done for $k^* = 5$. We describe this computation in Appendix D.1 of the Supplementary Materials.

In our search algorithm, whenever we computed the power, Type I error, or expected sample size for a given configuration $(k^*, f_1, f_2, n^{(1)}, n^{(2)}, e_C, e_1)$, we used the asymptotic approximation of the joint distribution of the z-statistics (3), by a multivariate normal distribution with covariance matrix Σ. This allows faster computations, compared to simulations that generate outcomes for each participant. To compute $\alpha_0(\mathbf{u}, \Sigma)$, we used the R package mvtnorm, which uses numerical integration based on the algorithm of Genz and Bretz [28] to evaluate multivariate normal probabilities. To compute power, we used Monte Carlo simulation of multivariate normal distributions with 10^5 simulated trials per power computation. We discuss the reason for this choice in Appendix D.1 of the Supplementary Materials. We recorded every configuration $(k^*, f_1, f_2, n^{(1)}, n^{(2)}, e_C, e_1)$ in our search for which the design goals (i)-(iii) are achieved and $\alpha_0(\mathbf{u}, \Sigma) = 0.025$. Among these, we selected the configuration with the smallest average expected sample size over scenarios (a)-(c), which is the design D_{ADAPT} given in Table 1.

6 Comparison of Designs Achieving Goals (i)-(iii) for the MISTIE Phase III trial

6.1 Comparator Designs

We compare D_{ADAPT} to two other types of designs that achieve goals (i)-(iii). First, consider the class of standard designs defined exactly as the adaptive designs in Section 4 except leaving out step 3 (the adaptive feature allowing a switch to only enroll from subpopulation 1), and setting $k^* = K$. This is equivalent to setting $k^* = K$ and setting the futility boundary for stopping subpopulation 2 enrollment $l_{2,k} = -\infty$ for each $k < K$; therefore, the class of standard designs is a subclass of the adaptive designs defined in Section 4. We executed the same search algorithm as used in selecting D_{ADAPT}, except restricted to standard designs; full details are given in Appendix D.2 of the Supplementary Materials. The optimal design returned by the search, denoted D_{STD}, is given in Table 2.

Second, we consider a class of adaptive enrichment designs based on the p-value combination approach [33, 34, 35, 5]. This approach has been applied to construct adaptive enrichment designs [6, 7, 8, 9]. Specifically, we apply the method from Jennison and Turnbull [7, Section 6] using the Simes test [36] to combine p-values within a given stage, and the weighted inverse normal method [37, 35] to combine p-values across stages. We additionally consider designs where the Simes test is replaced by a generalization of the Dunnett test [38], which improved the performance of these designs. This class of designs is precisely defined in Appendix C of the Supplementary Materials. We executed a search over this class similar to that used in selecting D_{ADAPT}. The optimal design returned by the search is denoted D_{COMB}, which is summarized in Table 3.

6.2 Comparison of Designs when $\pi_1 = 1/3$

In Table 4, we compare the performance of $D_{\text{ADAPT}}, D_{\text{STD}}$, and D_{COMB}. The expected sample size and power for each design is given in twelve scenarios, which include scenarios (a)-(c). For each scenario and design, we simulated 10^5 trials. For conciseness, we refer to the average treatment effect (on the risk difference scale) as the percent benefit. We vary the percent benefit in each subpopulation, and for each design we report the expected sample size and the power to reject at least H_{0C}, to reject at least H_{01}, and to reject at least one of H_{0C}, H_{01}, respectively.

The expected sample size averaged over scenarios (a)-(c), for D_{ADAPT}, D_{STD}, and D_{COMB}, is 635, 889, and 1263, respectively. In every scenario in Table 4, which includes scenarios (a)-(c), D_{ADAPT} has the minimum expected sample size among the three designs. The maximum sample size, which occurs when there is no early stopping, is 1182, 1546, and 2098 for D_{ADAPT}, D_{STD}, and D_{COMB}, respectively. This shows a substantial savings in terms of the maximum sample size for D_{ADAPT} compared to the other designs.

Scenario (c), in which $\Delta_1 = \Delta_2 = 0$, is represented in row 5 of the bottom half of Table 4. The probability of rejecting at least one null hypothesis for D_{ADAPT} is 2.0%, which is less than the required familywise Type

I error rate $\alpha = 0.025$. The reason is that we use nonbinding futility boundaries, which can be a desirable property as discussed in Section 4.

The p-value combination design D_{COMB} requires substantially greater sample size at each stage, compared to D_{ADAPT} and D_{STD}, as shown in Tables 1–3. This is because the class of p-value combination designs that we considered, which use standard combination tests and intersection tests, is not flexible enough to take advantage of the asymmetry of our problem. Specifically, neither H_{0C} nor H_{01} can be rejected unless the intersection test for $H_{0C} \cap H_{01}$ is rejected, and standard intersection tests such as the aforementioned Simes and generalized Dunnett tests treat p-values from each population symmetrically. The resulting inefficiency can be seen in Table 4, where D_{COMB} exactly achieves goal (ii), but has much more power than required to achieve goal (i) and has familywise Type I error only 1.2% in scenario (c). It is possible to replace the intersection tests by more flexible tests that allocate Type I error asymmetrically to the two null hypotheses, e.g., by using the Spiessens–Debois test with unequal α allocations [39]. However, to the best of our knowledge, this test with such an unequal allocation has not been implemented in a way that is straightforward to use within the p-value combination approach; such an implementation would be nontrivial, as we discuss in Appendix C of the Supplementary Materials. Also, the resulting design and testing procedure using the p-value combination approach would still not be based solely on minimal sufficient statistics.

We have shown that our proposed adaptive designs can have superior performance for achieving the goals in Section 2 compared to a subclass of designs based on the p-value combination approach that use standard intersection tests and combination functions. This is a proof-of-concept demonstration that our designs may have advantages in some situations, and does not imply our adaptive designs are better than all possible designs based on the p-value combination approach; in order to determine this, one would need to conduct extensive comparisons of all possible designs from these classes, which is beyond the scope of this work.

6.3 Comparison of Designs at Different Subpopulation Proportions

We examine the impact of different subpopulation proportions π_1 on the relative performance of different types of designs. For each $\pi_1 \in \{1/3, 1/2, 2/3\}$, we constructed designs D_{ADAPT}, D_{STD}, and D_{COMB} by executing the search described above at this value of π_1. The designs minimizing the expected sample size averaged over scenarios (a)-(c) for $\pi_1 = 1/3$ are given in Tables 1–3, while the corresponding designs for $\pi_1 \in \{1/2, 2/3\}$ are given in Appendix F of the Supplementary Materials. Figure 2 summarizes their performance.

Figure 2a plots the expected sample size averaged over scenarios (a)-(c), for each of the three types of designs. Under this metric, the design D_{ADAPT} is better than the other designs, at all values of π_1 we considered. The improvement of D_{ADAPT} over the other designs is largest when $\pi_1 = 1/3$. When $\pi_1 = 1/2$, the expected sample size averaged over scenarios (a)-(c) is 550, 624, and 1016 for D_{ADAPT}, D_{STD}, and D_{COMB}, respectively. For $\pi_1 = 2/3$, D_{ADAPT} is only slightly better than the standard design in terms of the expected sample size averaged over scenarios (a)-(c). The expected sample sizes and power for each design, in scenarios that include (a)-(c), are given in Tables 6 and 7 of the Supplementary Materials for the cases of $\pi_1 = 1/2$ and $\pi_1 = 2/3$, respectively.

Consider D_{ADAPT} versus D_{STD}. The former has the ability to stop enrolling subpopulation 2 and switch to only enrolling from subpopulation 1. This ability pays off most when π_1 is relatively small, since then there is faster accrual of information on subpopulation 2 to use in deciding whether to stop enrolling them. Also, the sample size reduction due to stopping enrollment from subpopulation 2 is amplified at smaller π_1.

For each design, the expected sample size averaged over scenarios (a)-(c) is decreasing in π_1. This makes sense intuitively, since for larger π_1, there are more subpopulation 1 participants in a sample of fixed size from the combined population, making goal (ii) easier to achieve.

Figure 2b compares the maximum sample size for each of the three types of designs. The maximum sample size of D_{ADAPT} is substantially smaller than the other designs for $\pi_1 = 1/3$, while it is essentially tied with D_{STD} for $\pi_1 \in \{1/2, 2/3\}$.

For each $\pi_1 \in \{1/3, 1/2, 2/3\}$, the corresponding value of k^* for the optimal adaptive design D_{ADAPT} from our search is 3, 4, and 5, respectively. The main advantage of having $k^* < 5$ is that in some cases it reduces the maximum sample size of the trial. To demonstrate this, at $\pi_1 = 1/3$, we repeated our adaptive design search algorithm from Section 5 but only allowing designs with $k^* = 5$. The maximum sample size of the resulting design was 1393, which is substantially larger than the maximum sample size of 1182 for the design D_{ADAPT} with $k^* = 3$. Intuitively, the reason is that enrolling from subpopulation 2 is only useful for achieving goal (i), and D_{ADAPT} has enrolled enough participants from the combined population by the end of stage 3 to achieve this goal. In contrast, subpopulation 1 participants contribute to goals (i)-(ii), and D_{ADAPT} has not enrolled enough such participants by the end of stage 3 to achieve goal (ii). Therefore, it is efficient for D_{ADAPT} to enroll only subpopulation 1 in stages 4 and 5.

It is possible to consider other types of designs, e.g., a sequence of standard designs where the first enrolls from the combined population, and if it fails to reject H_{0C}, it is followed by a second standard design that enrolls only from subpopulation 1 and tests H_{01}. This overall procedure is an example of an adaptive enrichment design, since enrollment can be modified (restricted to subpopulation 1) in response to accrued data. This adaptive enrichment design can be considered a crude version of the adaptive designs above, with the difference that those above have a more flexible rule for modifying enrollment and use a multiple testing procedure that leverages the correlation among statistics.

6.4 Impact of Estimated Variances and Subpopulation Proportions

Above, for clarity, we focused on the case where variances $\sigma^2(Q_{sa})$ and the subpopulation proportions π_1, π_2 are known. In practice, these will generally not be known with certainty at the start of the trial, and will be estimated based on data accrued at each interim analysis. Also, we had assumed at each stage in which both subpopulations are enrolled, the ratio of subpopulation 1 participants to subpopulation 2 participants equals the ratio of the corresponding subpopulation sizes. In practice, there can be variability in the sample proportions of participants from each subpopulation.

We now consider the setting where variances and subpopulation proportions are unknown. The null hypotheses defined in Section 3.1 remain the same. We assume that during each stage k where the combined population is enrolled, the total number enrolled is n_k (which is preplanned), and the subpopulation membership $S_{i,k}$ of each participant is an independent draw with probability π_1 of $S_{i,k} = 1$. The variances in the statistics (3) are replaced by sample variances. We estimate Σ by the formulas in Appendix A.1 of the Supplementary Materials, using sample variances and the sample proportion $\hat{\pi}_1$, which are based on all relevant data accrued in previous stages.

In the case where Σ was assumed known, the boundaries $\mathbf{u}$ and $\mathbf{l}$ for our adaptive designs were selected such that $\alpha_0(\mathbf{u}, \Sigma) = \alpha$, which by Theorem 1 guarantees strong control of the familywise Type I error rate at level α. In Appendix E of the Supplementary Materials, we assume Σ and π_1 are unknown and estimated by $\hat{\Sigma}$ and $\hat{\pi}_1$, respectively. We generalize the adaptive design algorithm from Section 4 to handle estimated Σ and π_1. This is based on an extension of the error spending approach [29, 30] to handle multiple hypotheses. The result is a modified version of D_{ADAPT}, denoted D^*_{ADAPT}, that constructs efficacy boundaries $U_{C,k}, U_{1,k}$ at the end of each stage k based on the current estimates of Σ and π_1. The design D^*_{ADAPT} is constructed such that at the end of the trial, when the assumptions (5) hold, goals (i)-(iii) are approximately achieved. Even when these assumptions fail to hold, we have $\alpha_0(\mathbf{U}, \hat{\Sigma}) = \alpha$, where $\mathbf{U} = \{U_{C,k}\}_{k=1}^{k^*} \cup \{U_{1,k}\}_{k=1}^{K}$. We next examine the power and Type I error of D^*_{ADAPT} based on simulations.

First, consider $\pi_1 = 1/3$, for which the power and Type I error of D_{ADAPT} are given in Table 4 in the setting of known Σ and π_1. For the case where Σ and π_1 are estimated, we computed analogous quantities for D^*_{ADAPT} under identical scenarios (one per row) as in Table 4, under the assumptions (5). We simulated 10,000 trials in each scenario, and give full results in Appendix E.3 of the Supplementary Materials. The main result is that the familywise Type I error rate was always less than $\alpha = 0.025$, and goals (i)-(iii) were all achieved.

In Appendix E.3 of the Supplementary Materials, we examine the impact of true subpopulation proportions differing by ±5% from those planned for, on the design D^*_{ADAPT}. We applied D^*_{ADAPT}, which was constructed to achieve goals (i)-(iii) at $\pi_1 = 1/3$, in simulations where the true value of π_1 was set to either 0.28 or to 0.38. The power to reject H_{0C} in scenario (a) ranged from 80% to 82%, and the power to reject H_{01} in scenario (b) ranged from 78% to 83%, as the true value of π_1 ranged from 0.28 to 0.38. In all cases, the familywise Type I error rate was at most 0.023. This shows the design D^*_{ADAPT} is robust to the value of

π_1 differing by 5% from that planned for, in the case of $\pi_1 = 1/3$. It is an area of future research to examine the robustness of this method to different types of deviations from the assumptions (5).

7 Software

The open-source, free software package **interAdapt** [40] implements the class of designs from Section 4. It can be used to compute power, expected sample size, and expected trial duration for these designs. The adaptive designs are automatically compared to standard designs, and the results are displayed in plots and tables. This software can be used as a trial planning tool, to evaluate whether an adaptive design from Section 4 offers tangible benefits over standard designs. The software includes a user-friendly, graphical interface that runs through a web browser. This makes the software accessible to a wider audience, since users do not need to know **R**, the programming language in which the software is written. The results of the power, sample size, and duration comparisons are automatically compiled into a report that can be downloaded as a pdf document. The software can be used when outcomes are binary, such as in the examples in this paper; future work is to extend this to handle any real-valued outcome.

8 Discussion

An alternative to restricting to the class of designs in Section 5 would be to extend the dynamic programming method of Eales and Jennison [41], which was designed for a single null hypothesis in a standard group sequential design, to our setting that involves multiple hypotheses and adaptive designs. However, this could be extremely difficult computationally, given the many parameters underlying each design and the constraints that each design must strongly control the familywise Type I error rate. It is an area of future research to attempt to extend the approach of Eales and Jennison [41] to our setting.

Though our hypothesis tests were based on z-statistics, it is possible to apply the same approach using statistics that appropriately leverage baseline variables as in, e.g., [42]. If these variables are prognostic for the primary outcome, there is potential to increase power. It is an open research problem to determine how much can be gained by leveraging such information.

An area of future research is to consider modifications of the adaptive design algorithm in Section 4 that permit continuation after H_{0C} or H_{01} is rejected to allow further testing of the remaining null hypothesis; we still assume enrollment of subpopulation 2 stops by the end of stage k^*, and that the entire trial stops by the end of stage K. Even with such a modification, any design $D(\mathbf{u}, \mathbf{l})$ satisfying the condition in Theorem 1 is guaranteed to strongly control the familywise Type I error rate at level α, as proved in Appendix A of the Supplementary Materials. Such a modified design would generally lead to higher power, but also higher expected sample size, compared to the same design without this modification; characterizing this tradeoff between power and expected sample size is an area of future research.

Another open problem is to consider different goals than (i)-(iii), e.g., to add a power requirement for the null hypothesis of no average treatment benefit for subpopulation 2 (those with large IVH), under a scenario where the treatment only benefits this population. The class of adaptive designs from Section 4 can be generalized to also test such a null hypothesis, and to incorporate a rule for switching to only enroll subpopulation 2, as we outline in Appendix G of the Supplementary Materials.

A limitation of our designs is that each participant's outcome is assumed to be observed relatively soon after enrollment. A more challenging problem is to handle outcomes observed with delay, since less information will be available at each interim analysis upon which to base decisions about enrollment changes. It is an area of future work to characterize when adaptive designs can be useful in this setting. The adaptive design algorithm from Section 4 could be extended to handle this case, with the modification that if enrollment is stopped early, an additional analysis will be conducted once all pipeline participants (i.e., those enrolled whose outcomes are not yet observed) complete the trial, as in [43].

An open research question is how to plan follow-up trials after adaptive, enrichment designs. We briefly discuss this issue in Appendix H of the Supplementary Materials.

Supplementary Materials

Appendices A-H and R code are available in the separate files:

```
http://people.csail.mit.edu/mrosenblum/MISTIE_supp_mat.pdf
http://people.csail.mit.edu/mrosenblum/MISTIE_R_Files.zip
```

References

[1] Morgan T, Zuccarello M, Narayan R, Keyl P, Lane K, Hanley D. Preliminary findings of the minimally-invasive surgery plus rtpa for intracerebral hemorrhage evacuation (MISTIE) clinical trial. *Acta Neurochir Suppl.* 2008; **105**:147–51.

[2] Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley Interscience: New York, 1987.

[3] Rice JA. *Mathematical Statistics and Data Analysis*. 2nd edn., Duxbury Press: Belmont, California, USA, 1995.

[4] Emerson SS. Issues in the use of adaptive clinical trial designs. *Statistics in Medicine* 2006; **25**(19):3270–3296, doi:10.1002/sim.2626.

[5] Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* 2006; **48**(4):623–634.

[6] Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 2006; **48**(4):635–643.

[7] Jennison C, Turnbull BW. Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharmaceutical Statistics* 2007; :1135–1161, doi: 10.1080/10543400701645215.

[8] Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 2009; **28**(10):1445–1463, doi:10.1002/sim.3559.

[9] Jenkins M, Stone A, Jennison C. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 2011; **10**(4):347–356, doi:10.1002/pst.472.

[10] Boessen R, van der Baan F, Groenwold R, Egberts A, Klungel O, Grobbee D, Knol M, Roes K. Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics* 2013; doi: 10.1002/pst.1599.

[11] Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in medicine* 2012; **31**(30):4309–4320.

[12] Stallard N. Group-sequential methods for adaptive seamless phase ii/iii clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **21**(4):787–801.

[13] Rosenblum M, van der Laan MJ. Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* 2011; **98**(4):845–860, doi:10.1093/biomet/asr055.

[14] Stallard N, Hamborg T, Parsons N, Friede T. Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of biopharmaceutical statistics* 2014; **24**(1):168–187.

[15] Wang SJ, O'Neill RT, Hung H. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist.* 2007; **6**:227–244.

[16] Wang SJ, Hung H, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* 2009; **51**:358–374.

[17] Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**(4):800–802, doi:10.1093/biomet/75.4.800.

[18] Russek-Cohen E, Simon RM. Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine* 1997; **16**:455–464.

[19] Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press, 1999.

[20] O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.

[21] Götze F. On the rate of convergence in the multivariate CLT. *The Annals of Probability* 1991; **19**(2):724–739.

[22] FDA. Draft guidance for industry. Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf 2010.

[23] Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *JASA* 2002; **97**(460):1034–1041.

[24] FDA, EMEA. E9 statistical principles for clinical trials. *U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96* 1998.

[25] Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**(4):pp. 1315–1324.

[26] Bickel PJ, Doksum KA. *Mathematical Statistics*, vol. 1. Prentice Hall: Upper Saddle River, New Jersey, 2001.

[27] Liu Q, Anderson KM. On adaptive extensions of group sequential trials for clinical investigations. *J. Amer. Statist. Assoc.* 2008; **103**(484):1621–1630, doi:10.1198/016214508000000986.

[28] Genz A, Bretz F. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics, Springer-Verlag: Heidelberg, 2009.

[29] Slud EV, Wei LJ. Two-sample repeated significance tests based on the modified wilcoxon statistic. *J. Amer. Statist. Assoc.* 1982; **77**:862–868.

[30] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.

[31] Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 1979; **6**:65–70.

[32] Hanley D. http://braininjuryoutcomes.com/studies/mistie/entry/mistie/international-stroke-conference-2012-mistie-phase-2-results 2012.

[33] Bauer P. Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* 1989; **20**:130148.

[34] Bauer P, Köhne K. Evaluations of experiment s with adaptive interim analyses. *Biometrics* 1994; **50**:10291041.

[35] Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**(4):1286–1290, doi:10.1111/j.0006-341X.1999.01286.x.

[36] Simes RJ. An improved bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**(3):751–754.

[37] Mosteller F, Bush RR. Selected quantitative techniques. *Handbook of Social Psychology*, vol. 1, Lindzey G (ed.). Addison-Wesley: Cambridge, MA, 1954; 289–334.

[38] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**(272):1096–1121.

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

[40] Fisher AJ, Jaffee H, Rosenblum M. Interadapt – an interactive tool for designing and evaluating randomized trials with adaptive enrollment criteria. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 262* June 2014. http://people.csail.mit.edu/mrosenblum/interAdapt

[41] Eales JD, Jennison C. An improved method for deriving optimal one-sided group sequential tests. *Biometrika* 1992; **79**(1):13–24, doi:10.1093/biomet/79.1.13.

[42] Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine* 2009; **28**(1):39–64, doi:10.1002/sim.3445.

[43] Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2013; **75**(1):3–54, doi:10.1111/j.1467-9868.2012.01030.x.

Table 1: At $\pi_1 = 1/3$, the sample sizes and boundaries for the adaptive design D_{ADAPT}. No boundaries are given for $u_{C,k}$ and $l_{2,k}$ at the stage 4 and 5 analyses, since $k^* = 3$, i.e., enrollment of subpopulation 2 participants always stops after interim analysis 3. All boundaries are on the z-statistic scale.

Adaptive, Group Sequential Design $D_{\textbf{ADAPT}}$

Interim Analysis (k)	1	2	3	4	5
Cum. Sample Size Subpopulation 1	90	180	270	456	642
Cum. Sample Size Subpopulation 2	180	360	540	540	540
Cum. Sample Size Combined Population	270	540	810	996	1182
H_{0C} Efficacy Boundary ($u_{C,k}$)	4.76	3.36	2.75		
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	0	0	∞		
H_{01} Efficacy Boundary ($u_{1,k}$)	5.48	3.88	3.17	2.44	2.05
Boundary to Stop All Enrollment ($l_{1,k}$)	0	0	0	0	2.05

Table 2: At $\pi_1 = 1/3$, standard design D_{STD} described in Section 6.1.

Standard Design $D_{\textbf{STD}}$

Interim Analysis (k)	1	2	3	4	5
Cum. Sample Size Subpopulation 1	97	193	290	387	515
Cum. Sample Size Subpopulation 2	193	387	580	773	1030
Cum. Sample Size Combined Population	290	580	870	1160	1546
H_{0C} Efficacy Boundary ($u_{C,k}$)	6.70	4.74	3.87	3.35	2.90
H_{01} Efficacy Boundary ($u_{1,k}$)	4.70	3.32	2.71	2.35	2.04
Boundary to Stop All Enrollment ($l_{1,k}$)	0	0	0	0	2.04

Table 3: At $\pi_1 = 1/3$, adaptive design D_{COMB} based on combination tests and closure principle described in Section 6.1. Full details including the definition of local tests are given in Appendix C of the Supplementary Materials. The generalized Dunnett test defined there is used to combine p-values within each stage for the local test of $H_{0C} \cap H_{01}$.

Adaptive Design Using Combination Tests and Closure Principle D_{COMB}

Interim Analysis (k)	1	2	3	4	5
Cum. Sample Size Subpopulation 1	223	446	670	714	759
Cum. Sample Size Subpopulation 2	446	893	1339	1339	1339
Cum. Sample Size Combined Population	670	1339	2009	2054	2098
Note: Local Test of $H_{0C} \cap H_{01}$ Must Reject Before Any Elementary Null Hyp. Rejected					
$H_{0C} \cap H_{01}$ Local Test Efficacy Boundary	3.66	2.59	2.12	2.09	2.07
H_{0C} Local Test Efficacy Boundary ($u_{C,k}$)	3.47	2.45	2.00		
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	0	0	∞		
H_{01} Local Test Efficacy Boundary ($u_{1,k}$)	3.82	2.70	2.20	2.13	2.07
Boundary to Stop All Enrollment ($l_{1,k}$)	0	0	0	0	2.07

Table 4: Comparison of expected sample size (ESS) and power (as a percent) at $\pi_1 = 1/3$. Each row corresponds to a different treatment effect for subpopulation 2. The column headings H_{0C}, H_{01}, and ≥ 1, denote power to reject at least H_{0C}, at least H_{01}, and at least one null hypothesis, respectively. The three rows in boldface correspond to scenarios (a)-(c), respectively.

Subpopulation 1 Treatment Effect Δ_1 set at 12.5%

	Subpop. 2 Effect Δ_2 (%)	D_{ADAPT}				D_{STD}				D_{COMB}			
		ESS	H_{0C}	H_{01}	≥ 1	ESS	H_{0C}	H_{01}	≥ 1	ESS	H_{0C}	H_{01}	≥ 1
	15.0	594	86	6	89	802	85	33	90	1078	98	31	98
scenario (a):	**12.5**	**645**	**80**	**13**	**88**	**870**	**80**	**44**	**89**	**1281**	**96**	**47**	**97**
	10.0	702	69	25	87	942	70	56	89	1484	93	63	96
	5.0	779	34	59	84	1042	30	76	83	1699	68	80	89
scenario (b):	**0.0**	**737**	**7**	**80**	**84**	**1062**	**2**	**80**	**80**	**1442**	**22**	**80**	**80**
	-5.0	648	0	83	84	1063	0	80	80	1140	1	80	80

Subpopulation 1 Treatment Effect Δ_1 set at 0

	Subpop. 2 Effect Δ_2 (%)	D_{ADAPT}				D_{STD}				D_{COMB}			
		ESS	H_{0C}	H_{01}	≥ 1	ESS	H_{0C}	H_{01}	≥ 1	ESS	H_{0C}	H_{01}	≥ 1
	15.0	474	33.0	0.1	33.1	594	31.1	0.4	31.2	1029	43.9	0.5	44.0
	12.5	505	25.6	0.3	25.9	637	27.4	0.7	27.5	1121	36.6	0.8	36.6
	10.0	535	17.0	0.6	17.6	681	21.3	1.2	21.6	1190	27.3	1.0	27.5
	5.0	560	3.8	1.4	5.2	729	4.8	1.9	6.1	1222	7.5	1.2	8.1
scenario (c):	**0.0**	**522**	**0.3**	**1.8**	**2.0**	**735**	**0.1**	**2.1**	**2.2**	**1065**	**0.4**	**0.9**	**1.2**
	-5.0	475	0.0	1.9	1.9	735	0.0	2.1	2.1	913	0.0	0.9	0.9

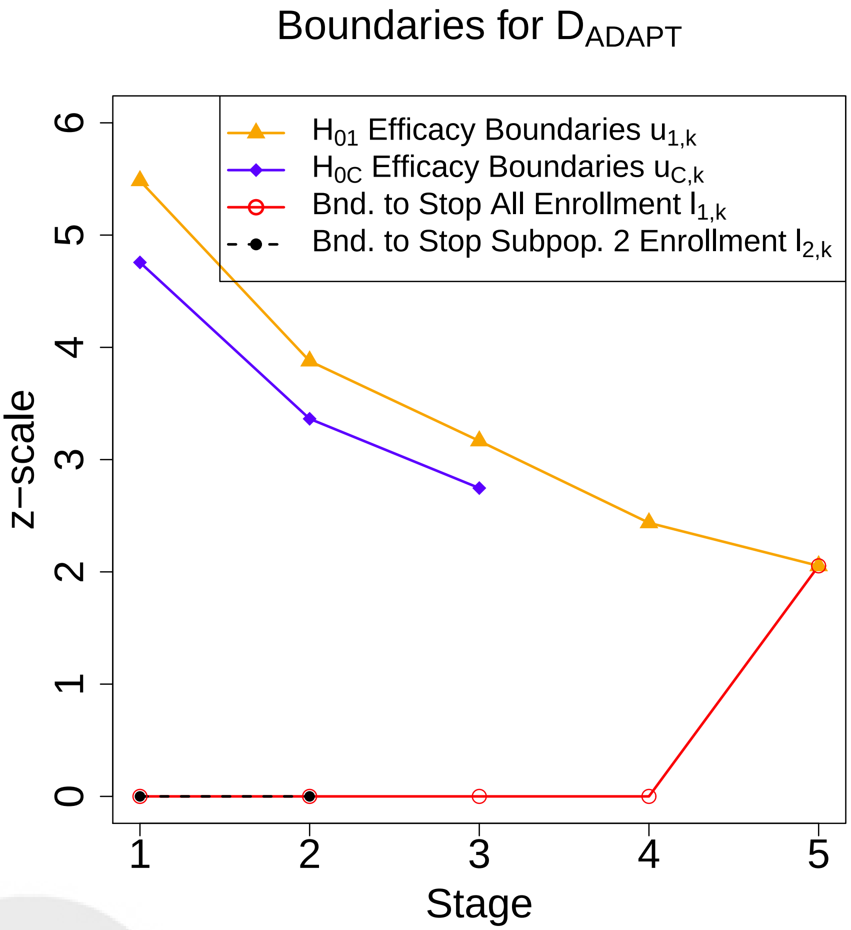
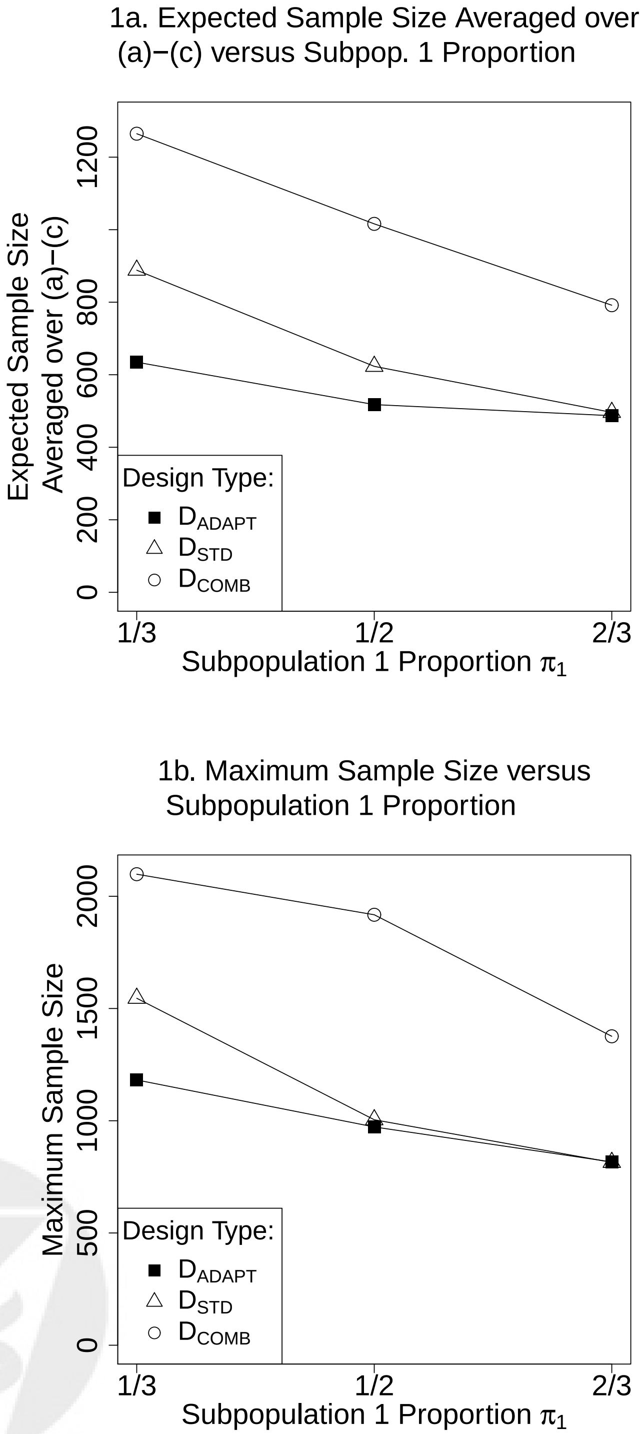


Figure 1: Boundaries for D_{ADAPT} at $\pi_1 = 1/3$. Subpopulation 2 enrollment always stops at or before the end of stage $k^* = 3$. The boundaries $u_{1,k}, u_{C,k}, l_{1,k}, l_{2,k}$ correspond to the cumulative z-statistics $Z_{1,k}, Z_{C,k}, Z_{1,k}, Z_{2,k}$, respectively; these are used in the adaptive, group sequential algorithm from Section 4. These boundaries are also given in Table 1.



1a. Expected Sample Size Averaged over (a)−(c) versus Subpop. 1 Proportion

1b. Maximum Sample Size versus Subpopulation 1 Proportion

Figure 2: Fig. 2a: At different values of subpopulation 1 proportion π_1, comparison of expected sample size averaged over scenarios (a)-(c) for optimal designs. Fig. 2b: Analogous comparison of maximum sample sizes.