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

Michael Rosenblum[1], Richard E. Thompson[1], Brandon Luber[1], and Daniel Hanley[2]

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

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

Summary

In designing a Phase III randomized trial, care must be taken in selecting the target population. Advantages of enrolling from a larger population include wider generalizability of results and faster recruitment. However, earlier trials (e.g. Phase II trials) and medical knowledge may provide stronger evidence of a treatment effect for certain subpopulations. This makes a Phase III trial that targets the overall population more risky, since if the treatment only benefits a subpopulation, there may be low power to detect this. We propose new adaptive, group sequential 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. These designs use preplanned rules for changing the enrollment criteria if the participants from predefined subpopulations are not benefiting from the new treatment. We demonstrate these adaptive designs in the context of a Phase III trial of a new treatment for stroke, and compare them to standard, group sequential designs in terms of expected sample size.

Key words: Adaptive enrichment design; Group sequential design; Optimal sample size

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, and described in (Morgan *and others*, 2008).

The MISTIE treatment showed promise in a completed Phase II trial. However, that trial only enrolled participants with ICH who had small or no intraventricular hemorrhage (IVH) at baseline, who we call "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 medical understanding of brain hemorrhage, the investigators conjectured that the new treatment would substantially 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 the Phase II trial had not enrolled participants with large IVH, 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 these issues, we compare different designs for the Phase III MISTIE trial. New adaptive designs are proposed with prespecified rules for changing enrollment criteria if participants from predefined subpopulations are not benefiting from the new treatment or are experiencing high adverse event rates. We determine the design that minimizes expected sample size, among adaptive designs that satisfy certain power and familywise Type I error constraints. We compare this adaptive design to standard designs that satisfy the same constraints.

In the context of the MISTIE trial, we show in Section 6.1 that the adaptive design has smaller expected sample size compared to the standard designs, in each scenario considered. As a summary of each design's performance, we computed expected sample size averaged over three scenarios of primary interest. This is 635 for the adaptive design, compared to 722 for the best standard design considered. Based on the projected cost of $29,000 per patient in the MISTIE

Phase III trial, the reduced expected cost from using the adaptive design is roughly $2.5 million.

The new adaptive designs we propose are examples of adaptive enrichment designs, i.e., designs with preplanned rules for modifying enrollment criteria based on data accrued at interim analyses. An obstacle to widespread use of adaptive enrichment designs is that they lack the simplicity, flexibility, and familiarity of group sequential designs. Our designs incorporate advantages of both group sequential designs and adaptive enrichment designs, as described in Section 5.1.

We specify the goals of the MISTIE Phase III trial in the following section. Section 3 gives related work. Our statistical problem is formally defined in Section 4. In Section 5, we present an adaptive, group sequential design that achieves the MISTIE trial goals. We compare it to standard, group sequential designs in Section 6. Directions for future research are given in Section 7.

2. Design Goals of the MISTIE Phase III Trial

The aim of the MISTIE Phase III trial is to assess whether the MISTIE treatment is superior to the standard of care. The primary outcome is a participant's degree of disability as measured by the modified Rankin Scale (mRS). A mRS score of 3 or less is considered a successful outcome. Define the average treatment effect to be the difference between the probability of a successful outcome, under assignment to MISTIE treatment and under assignment to standard of care. For a given population, we say there is an average treatment benefit if the average treatment effect is positive. Based on the MISTIE Phase II randomized trial (which only enrolled small IVH participants), the estimated average treatment effect for the small IVH subpopulation is 12.1% [95% CI: (-2.7%, 26.9%)].

The clinical investigators indicated three potential scenarios of particular interest in planning the MISTIE Phase III trial. These scenarios are that the 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; and, (c) 0% for both subpopulations.

The design goals are the following:

(i) at least 80% power to reject the null hypothesis of no average treatment benefit for the combined population, in scenario (a);

(ii) at least 80% power to reject the null hypothesis of no average treatment benefit for the small IVH subpopulation, in scenario (b);

(iii) strong control of the familywise Type I error rate (defined in Section 4.3) at level 2.5%. (We use 2.5% rather than 5%, because we consider one-sided hypotheses in what follows.)

The design goals are asymmetric, in that 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 potential benefit in the small IVH population. Our general method can be applied to achieve alternative design goals, as we discuss in Section 7.

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. Our main contribution is a class of new adaptive, group sequential designs tailored to the problem of testing treatment effects in a subpopulation and the overall population.

We focus on the case of outcomes measured soon after enrollment. It is a challenging open problem to develop designs meeting the above goals when the outcome is measured with delay, which we discuss in Section 7.

3. Related Work

Follman *and others* (1994) present group sequential designs for testing multiple treatments, where

identical efficacy boundaries are used in the test of each null hypothesis. Though it would be possible to modify their method for our problem, having identical boundaries is not well suited to our context, due to having different rates of information accrual for different sized subpopulations.

Russek-Cohen and Simon (1997) present two-stage adaptive designs that potentially increase the sample sizes for each of two subpopulations if an interim test of no interaction is rejected. Their goals differ from ours, in that they require sufficient power to detect benefits for each subpopulation, while we focus on the combined population and a single subpopulation. Other differences between their work and ours are that we prove strong control of the familywise Type I error rate for our designs, and we consider subpopulations with unequal sizes and designs with more than two stages.

Wang *and others* (2007, 2009) present two-stage, adaptive enrichment designs. The main differences between these designs and ours are the following: we construct designs meeting goals (i)-(iii) that minimize expected sample size, while they focus on different power goals and up to 3 nested subpopulations; our designs allow any number of stages and interim analyses, and directly incorporate features of group sequential designs; we use the covariance of the statistics corresponding to different populations to improve efficiency in our tests.

Boessen *and others* (2013) compare designs, including a two-stage, adaptive enrichment design, for testing the same null hypotheses as here. However, they consider different goals than we do; they focus on designs that have 80% power to reject at least one null hypothesis under scenario (a), while we focus on designs achieving goals (i)-(iii). They show their adaptive designs control the familywise Type I error rate at the global null hypothesis (defined below), but do not prove strong control of the familywise Type I error rate. Their designs involve different decision rules than ours; it is an area of future work to compare their approach to ours.

Jennison and Turnbull (2007, Section 6), Brannath *and others* (2009), Jenkins *and others* (2011), and Stallard (2011) construct adaptive enrichment designs, based on the combination test

approach and the closed testing principle. We describe the differences between this alternative approach and our approach in Appendix A.3 of the Supplementary Materials.

Rosenblum and van der Laan (2011) present two-stage adaptive designs, and a general method for proving strong control of the familywise Type I error rate. In contrast, the designs here have five stages, and a different approach is used to prove familywise Type I error control.

4. Designs, Hypotheses, Assumptions, and Test Statistics

4.1 *Designs*

Group sequential designs involve preplanned, interim analyses where hypotheses are tested and enrollment may continue or be completely stopped. Examples include the designs of Pocock (1977) and O'Brien and Fleming (1979).

Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on data accrued at an interim analysis (Wang *and others*, 2007). Our adaptive enrichment designs defined in Section 5.1 have such a rule for potentially switching to enroll only small IVH participants. Our designs also incorporate group sequential testing. For conciseness, we refer to our designs that incorporate features of adaptive enrichment designs and group sequential designs simply as adaptive designs. We use the term "standard designs" to refer to standard, group sequential designs, i.e., group sequential designs where the only decision about enrollment at each interim analysis is whether to continue the trial or stop it entirely.

4.2 *Data Structure, Hypotheses to be Tested, and Assumptions*

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. Let π_s denote the proportion of the combined population in subpopulation s, for $s \in \{1, 2\}$. For clarity of presentation, at each stage in which both subpopulations are enrolled, we assume the ratio of small IVH participants

to large IVH participants equals the ratio of the corresponding population sizes. However, we show in Section 6.3 that we obtain similar results in the setting where at each stage in which both subpopulations are enrolled, the subpopulation membership of each participant is a random, independent draw with probability π_1 of being subpopulation 1.

For each design, the following are defined before the trial starts: $K \geqslant 1$, the maximum number of stages; n, the maximum, total sample size of the trial; and a vector $(r_1, \ldots, r_K)$ of positive values that sum to 1, where $n_k = nr_k$ represents the maximum number of participants that can be enrolled in stage k. Fewer than n_k participants may be enrolled in stage k, for example, if enrollment is restricted to a single subpopulation, as described in Section 5.1.

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 MISTIE treatment ($A_{i,k} = 1$) versus control ($A_{i,k} = 0$); and $Y_{i,k}$ is an indicator of a successful outcome. We assume at each stage, for each subpopulation, exactly half of the participants are randomly assigned to MISTIE treatment; this can be approximately achieved by using block randomization stratified by subpopulation.

We assume that conditioned on the subpopulation membership and treatment arm assignments $\{S_{i,k}, A_{i,k}\}_{i=1}^{n_k}$ of all participants in stage k, the outcome $Y_{i,k}$ is a random draw from a Bernoulli distribution with unknown parameter p_{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 parameters by $\mathbf{p} = (p_{10}, p_{11}, p_{20}, p_{21})$. Define the variances $\sigma_{sa}^2 = p_{sa}(1 - p_{sa})$ for each $s \in \{1, 2\}, a \in \{0, 1\}$.

We make no assumptions on the unknown parameters $(p_{10}, p_{11}, p_{20}, p_{21})$, except that for some $\delta > 0$, we have $\delta \leqslant p_{sa} \leqslant 1 - \delta$ for each $s \in \{1, 2\}, a \in \{0, 1\}$. This guarantees the joint distribution of z-statistics for each population, given in Section 4.4 below, converges uniformly to a multivariate normal distribution. Consider a fixed $\delta > 0$, and let $\mathcal{P}$ denote the set of $\mathbf{p}$ where each component p_{sa} satisfies $\delta \leqslant p_{sa} \leqslant 1 - \delta$.

For each subpopulation $s \in \{1, 2\}$, define the average treatment effect on the risk difference

scale as $\Delta_s = p_{s1} - p_{s0}$. 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 $H_{01} = \{\mathbf{p} \in \mathcal{P} : \Delta_1 \leqslant 0\}$ and $H_{0C} = \{\mathbf{p} \in \mathcal{P} : \Delta_C \leqslant 0\}$. Also, define the global null hypothesis $H_0 = \{\mathbf{p} \in \mathcal{P} : \Delta_1 = \Delta_2 = 0\}$.

We make the following assumptions based on prior studies, as described by Hanley (2012):

$$\pi_1 = 1/3; \qquad p_{10} = 0.25; \qquad p_{20} = 0.2. \tag{4.1}$$

4.3 *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. Food and Drug Administration (FDA) for Phase III adaptive designs of drugs and biologics (FDA, 2010). A multiple testing procedure is said to strongly control the familywise Type I error rate at level α if for *any* data generating distribution $\mathbf{p} \in \mathcal{P}$, the familywise Type I error rate is at most α (Hochberg and Tamhane, 1987). Jennison and Turnbull (2007) give an overview of many useful methods that have been developed to prove this property in adaptive designs.

Define a trial design D to be a decision rule that determines, at the end of each stage, which null hypotheses to reject, and which populations to enroll in the next stage. We say D strongly controls the asymptotic, familywise Type I error rate at level α if

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\mathbf{p},D,n_1,\ldots,n_K}(\text{at least one true null hypothesis is rejected}) \leqslant \alpha, \tag{4.2}$$

where $\Pr_{\mathbf{p},D,n_1,\ldots,n_K}$ represents the probability under parameters $\mathbf{p}$, design D, and per-stage sample sizes $(n_1,\ldots,n_K) = (nr_1,\ldots,nr_K)$, for $(r_1,\ldots,r_K)$ the fixed values defined above. We prove our adaptive designs satisfy (4.2) at level $\alpha = 0.025$, using Theorem 5.1 below.

The above definition of strong control of the familywise Type I error is asymptotic, as sample size in each stage goes to infinity. (We note that for binary outcomes as used here, even in the simpler setting of a standard trial design testing a single null hypothesis with a z-test or t-test,

Type I error is only controlled asymptotically.) We examine the familywise Type I error rate of our designs at realistic sample sizes, using simulations described in Section 6.3. 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."

4.4 *Statistics*

For each adaptive design in the class defined in Section 5.1, there is a preplanned rule for restricting enrollment to only subpopulation 1 based on data accrued at each interim analysis. We consider a variety of such rules, some of which allow enrollment to potentially continue through the final stage K, and other rules which involve a prespecified stage k^* such that only subpopulation 1 may be enrolled during each stage $k > k^*$; this is equivalent to requiring a stop of subpopulation 2 enrollment for futility at the end of stage k^*. The advantage of considering decision rules with $k^* < K$ is that in some contexts these rules lead to lower maximum sample sizes and lower expected samples sizes (averaged over scenarios (a)-(c)) than comparable rules with $k^* = K$. Specifically, setting $k^* < K$ can be advantageous when the enrollment rate is much faster for subpopulation 2 than for subpopulation 1. In that case, enrolling both subpopulations through the end of the trial would result in many more subpopulation 2 participants, which is inefficient for achieving goals (i)-(iii). This phenomenon is examined at the end of Section 6.2.

For each $k \leqslant k^*$, define the 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 by the end of stage k:

$$Z_{C,k} = \left\{ \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} A_{i,k'}}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} A_{i,k'}} - \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} (1 - A_{i,k'})}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} (1 - A_{i,k'})} \right\} \mathrm{se}_{C,k}^{-1}, \tag{4.3}$$

where $\mathrm{se}_{C,k} = \left[\sum_{s=1}^{2} \pi_s \left\{ \sigma_{s0}^2 + \sigma_{s1}^2 \right\} / \sum_{k'=1}^{k} (n_{k'}/2) \right]^{1/2}$ is the standard error of the quantity in curly braces above. The z-statistics $Z_{1,k}$ for subpopulation 1 and $Z_{2,k}$ for subpopulation 2 are

defined analogously, except restricted to the corresponding subpopulation; their definitions are given in Appendix A.1 of the Supplementary Materials. $Z_{1,k}$ is defined for each $k \leqslant K$, while $Z_{2,k}$ is defined for each $k \leqslant k^*$.

The joint distribution of the combined population z-statistics $\mathbf{Z}_C = \{Z_{C,k}\}_{k=1}^{k^*}$ has the canonical covariance structure of a standard group sequential design (Jennison and Turnbull, 1999, Chapter 3). The same holds for $\mathbf{Z}_1 = \{Z_{1,k}\}_{k=1}^{K}$, and for $\mathbf{Z}_2 = \{Z_{2,k}\}_{k=1}^{k^*}$. 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.

5. Designs

All of our designs have $K = 5$ stages and use 1:1 randomization of participants to treatment versus control. Nonbinding futility boundaries, defined below, are included in all our designs. As described by Liu and Anderson (2008), nonbinding futility boundaries are typically preferred by the FDA, who require familywise Type I error guarantees to hold regardless of whether the futility boundaries are ignored or not. Our designs have this property.

5.1 *Adaptive, Group Sequential Designs for Testing* $\{H_{0C}, H_{01}\}$

In our adaptive designs, enrollment is initially from the combined population. At each interim analysis, a preplanned rule determines whether to reject any null hypotheses among $\{H_{0C}, H_{01}\}$, and whether to continue enrolling from the combined population, to switch to enrolling only small IVH participants, or to stop all enrollment. We do not allow the option to restart enrollment of large IVH participants once it has been stopped.

Our adaptive designs have a prespecified parameter k^* such that only subpopulation 1 may be enrolled during each stage $k > k^*$. For each $k > k^*$, if the trial continues after stage $k-1$, then n_k participants from subpopulation 1 are enrolled during stage k. If enrollment is restricted to subpopulation 1 at the end of some stage $k' < k^*$, then for each subsequent stage k before the

trial is stopped: if $k \leqslant k^*$, $\pi_1 n_k$ participants from subpopulation 1 are enrolled; if $k > k^*$ (i.e., it was preplanned that only subpopulation 1 could be enrolled during stage k), then n_k participants from subpopulation 1 are enrolled in stage k.

Let $\mathbf{u}_C = \{u_{C,k}\}_{k=1}^{k^*}$ and $\mathbf{u}_1 = \{u_{1,k}\}_{k=1}^{K}$ denote efficacy boundaries for the combined population and for subpopulation 1, respectively; let $\mathbf{l}_1 = \{l_{1,k}\}_{k=1}^{K}$ and $\mathbf{l}_2 = \{l_{2,k}\}_{k=1}^{k^*}$ denote futility boundaries. Our adaptive designs, defined below, take $(\mathbf{u}_C, \mathbf{u}_1, \mathbf{l}_1, \mathbf{l}_2, \pi_1, n_1, \ldots, n_K, k^*)$ 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.

Adaptive, Group Sequential Design for Testing $\{H_{0C}, H_{01}\}$. At each interim analysis $k \leqslant K$:

1. (Assess Efficacy) If $Z_{C,k} > u_{C,k}$ or $Z_{1,k} > u_{1,k}$, then stop the trial and for each $j \in \{C, 1\}$ for which $Z_{j,k} > u_{j,k}$, reject H_{0j}.

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

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

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

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

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

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

Note that step 1 is never run for any stage $k > k^*$, since step 3 forces the algorithm to remain in step 3 for every stage after k^* (if the trial has not already stopped).

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 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 at a given level α. Below, for clarity of presentation, we assume the covariance matrix Σ and the subpopulation 1 proportion π_1 are known. In Section 6.3, we examine the impact of using estimates of Σ and π_1. We focus on the case where π_1 is in a known, relatively small range, which we take to be 0.33 ± 0.05 in the context of the MISTIE trial. We show through simulations that in all the scenarios we considered, our adaptive designs have similar performance as in the setting where Σ and π_1 are assumed known.

We next give an easy to check criterion that implies strong control of the familywise Type I error rate for a given design $D(\mathbf{u}, \mathbf{l})$. 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 mean $\mathbf{0}$, and covariance matrix equal to the restriction of Σ to the corresponding statistics $\{Z_{C,k}\}_{k=1}^{k^*} \bigcup \{Z_{1,k}\}_{k=1}^{K}$. For any efficacy boundaries $\mathbf{u}$, define

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

In Appendix A.2 of the Supplementary Materials we prove:

THEOREM 5.1 Consider any $\alpha \in (0, 1)$, any efficacy boundaries $\mathbf{u}$ such that $\alpha_0(\mathbf{u}, \Sigma) \leqslant \alpha$, and any futility boundaries $\mathbf{l}$. The adaptive, group sequential algorithm $D(\mathbf{u}, \mathbf{l})$ 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 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 from the multivariate normal distribution function, which is implemented in the mvtnorm package in R (Genz and Bretz, 2009). To handle the case of Σ unknown and estimated at each stage, we extend the error-spending approach of Slud and Wei (1982) and Lan and DeMets (1983) to our context involving multiple hypotheses, as described in Section 6.3.

An additional benefit to using efficacy boundaries satisfying the condition in Theorem 5.1 is that one can allow switching to only enrolling small IVH participants for any reason, and still strong control of the familywise Type I error rate holds. We prove this in Appendix A.2 of the Supplementary Materials. Such a switch could be warranted, for example, if the adverse event rates are high for only the large IVH participants.

Our adaptive enrichment designs defined above incorporate advantages of group sequential designs. First, our designs allow any number of preplanned interim analyses, unlike many adaptive enrichment designs that are limited to two stages. Second, we take advantage of the covariance of the statistics corresponding to different populations in our multiple testing procedure. This can provide increased power compared to procedures, such as the Hochberg procedure (Hochberg, 1988), that do not directly incorporate covariance information. Third, our designs involve predefined efficacy boundaries for each null hypothesis of interest, which can be based on commonly used group sequential boundaries for a single null hypothesis. For example, in the designs in Section 5.2, we use boundaries for each null hypothesis that are proportional to the commonly used boundaries of O'Brien and Fleming (1979). In some cases, this may make our designs simpler to communicate compared to adaptive enrichment designs based on combination tests and the closure principle, as discussed in Appendix A.3 of the Supplementary Materials.

5.2 *Selecting an Efficient Adaptive Design that Achieves All Design Goals*

We describe how we selected an adaptive design for the Phase III MISTIE trial. We make the assumptions (4.1), which include that $\pi_1 = 1/3$. We generated a list of adaptive designs, each defined by $(n_1, \ldots, n_5, k^*, \mathbf{u}, \mathbf{l})$ for which $D(\mathbf{u}, \mathbf{l})$ 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 $k = 3$, $n_k = 280$ 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 in each stage $k' \leqslant 3$ for which the trial continues, $280 * \pi_1 = 280/3 \approx 93$ are enrolled from subpopulation 1. The design always stops enrolling from subpopulation 2 by the end of stage $k^* = 3$. In each of stages $k = 4$ and $k = 5$, unless the trial stops early, $n_k = 148$ participants are enrolled from subpopulation 1. The maximum total sample size is 1136. However, the expected sample size, in each of scenarios (a)-(c), is 674, 716, 517, respectively.

We next define our search algorithm, leading to the design D_{ADAPT}. Each $k^* \in \{1, 2, 3, 4, 5\}$ is considered separately. We restrict to $(n_1, \ldots, n_K)$ such that in each stage $k \leqslant k^*$, n_k is a common value denoted by $n^{(1)}$, and in each stage $k > k^*$, n_k is a common value denoted by $n^{(2)}$.

We consider efficacy and futility boundaries that are proportional to generalizations of the boundaries of O'Brien and Fleming (1979) to the case of unequal, per-stage sample sizes. Define $N_1(k)$ to be the cumulative sample size enrolled from subpopulation 1 by the end of stage k, assuming the trial continues through the end of stage k. For constants e_C, e_1, f_1, f_2, which are selected using the algorithm described below, we set boundaries as follows:

i. Efficacy boundaries for the combined population: $u_{C,k} = e_C(k^*/k)^{1/2}$, for each stage $k \leqslant k^*$;

ii. Efficacy boundaries for subpopulation 1: $u_{1,k} = e_1 \left\{N_1(K)/N_1(k)\right\}^{1/2}$, for each stage $k \leqslant K$;

iii. Futility boundaries for subpopulation 1: $l_{1,k} = f_1\{(K-1)/k\}^{1/2}$, for each stage $k \leqslant K-1$;

iv. Futility boundaries for subpopulation 2: $l_{2,k} = f_2\{(k^*-1)/k\}^{1/2}$, for each stage $k \leqslant k^*-1$.

Since enrollment from subpopulation 2 always stops at or before the end of stage k^*, we set $l_{2,k^*} = \infty$ and leave $l_{2,k}$ undefined for each stage $k > k^*$. Also, since stage K is the last stage of the trial, we set $l_{1,K} = u_{1,K}$.

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 C.1 of the Supplementary

Materials. For each $k^* \in \{1, 2, 3, 4, 5\}$ 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. We describe this computation in Appendix C.1 of the Supplementary Materials.

In our search algorithm, whenever we computed the power, Type I error, or expected sample size for a design 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 (4.3) by a multivariate normal distribution with covariance matrix Σ. We assess the quality of this asymptotic approximation in Section 6.3, where we show it was quite accurate for each optimal design produced in our searches. To compute $\alpha_0(\mathbf{u}, \Sigma)$, we used the R package mvtnorm described above. To compute power and expected sample size, we used Monte Carlo simulation of multivariate normal distributions with 10^5 simulated trials for each scenario and design configuration. We discuss the reason for this choice, and the approximation errors resulting from numerical integration and from simulation, in Appendix C.1 of the Supplementary Materials. In brief, the approximation error in each familywise Type I error rate computation was at most $\pm 10^{-4}$; the approximation error in each power computation was at most ± 0.005, with probability at least 0.998.

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 $\alpha_0(\mathbf{u}, \Sigma) = 0.025$. The above search returned 3005 such configurations. 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 Adaptive and Standard Designs

In Section 6.1, we compare D_{ADAPT} to standard designs that achieve goals (i)-(iii), in the context of the MISTIE trial (where $\pi_1 = 1/3$). We consider different values of π_1 in Section 6.2. In Section 6.3, we examine the impact of estimated variances and subpopulation proportions.

6.1 *Comparison of Designs for the MISTIE Phase III trial*

We compare D_{ADAPT} to two types of standard designs that achieve goals (i)-(iii). First, consider the class of standard designs defined exactly as the adaptive designs in Section 5.1 except leaving out step 3 (the adaptive feature allowing a switch to only enroll from subpopulation 1). We refer to each such design as a "single standard design," for reasons described below. We ran the same search algorithm as used in selecting D_{ADAPT}, except restricted to this class. The optimal design returned by the search, denoted $D_{\text{STD,SINGLE}}$, is given in Table 2.

The second comparator to our adaptive design is a procedure consisting of a pair of separate, standard designs run in parallel. The first enrolls the combined population and tests H_{0C} only, and the second enrolls subpopulation 1 and tests H_{01} only. For comparability to D_{ADAPT}, we allocate $\alpha/2 = 0.025/2$ to each of the two tests, so that the familywise Type I error rate of the overall procedure is strongly controlled at level 0.025. We call this overall procedure a "paired standard design," in contrast to the single standard design defined above. The sample sizes and boundaries for each paired standard design were selected by minimizing the expected sample size, averaged over scenarios (a)-(c), using an analogous search as used for D_{ADAPT}. The search algorithm is given in Appendix C.2 of the Supplementary Materials. We denote the optimal design by $D_{\text{STD,PAIR}}$, for which the sample sizes and boundaries are given in Table 3.

In Table 4, we compare the performance of D_{ADAPT}, $D_{\text{STD,SINGLE}}$, and $D_{\text{STD,PAIR}}$. 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}, $D_{\text{STD,SINGLE}}$, and $D_{\text{STD,PAIR}}$, is 635, 885, and 722, respectively. In each scenario in Table 4, D_{ADAPT} has the minimum expected sample size among the three designs.

Scenario (c), in which the global null hypothesis H_0 is true, is represented in row 5 of the bottom half of Table 4. The probabilities of rejecting at least one null hypothesis, for D_{ADAPT}, $D_{\text{STD,SINGLE}}$, and $D_{\text{STD,PAIR}}$, are 2.1%, 2.2%, and 2.3%, respectively. These are slightly below the required familywise Type I error rate $\alpha = 0.025$. The reason is that we are using nonbinding futility boundaries. That is, the Type I error rate must be at most 0.025 even when the futility boundaries are not obeyed, which can be a desirable property as discussed in Section 5. (Also, the adaptive design is guaranteed to control the familywise Type I error rate under arbitrary decisions to switch from enrolling the combined population to enrolling only the small IVH population, at any interim analysis, as discussed in Section 5.1.) The cost is that the designs are slightly conservative, in that they do not exhaust the familywise Type I error level $\alpha = 0.025$.

6.2 *Comparison of Optimal Designs Corresponding to Different Values of π_1*

Above, we assumed $\pi_1 = 1/3$, which corresponds to the proportion of small IVH participants in the context of the MISTIE trial. In other trial contexts, the proportion in subpopulation 1 will differ. We repeat the comparison from Section 6.1, except at different values of π_1. We assume (as above) that π_1 is known and can be used in our search algorithm. In Section 6.3, we explore the impact of estimated values of π_1.

For each $\pi_1 \in \{1/6, 1/3, 1/2, 2/3, 5/6\}$, we constructed the analog of D_{ADAPT} by executing the search in Section 5.2 at that value of π_1; we also did analogous searches as used to select $D_{\text{STD,SINGLE}}$ and $D_{\text{STD,PAIR}}$. For each π_1 considered, the designs minimizing the expected sample size averaged over scenarios (a)-(c), among the set of designs produced in the corresponding searches, are referred to as optimal designs. At $\pi_1 = 1/3$, the optimal designs are D_{ADAPT}, $D_{\text{STD,SINGLE}}$, and $D_{\text{STD,PAIR}}$. For the other values of π_1, the optimal designs are given in Appendix D of the Supplementary Materials. Their performance is summarized in Figure 1.

Figure 1a plots the expected sample size averaged over scenarios (a)-(c), for each of the three

types of designs. At each value of π_1, the corresponding adaptive design is equal or better than both types of standard designs. The improvement of the adaptive design over the best of the standard designs is largest for $\pi_1 \in \{1/3, 1/2\}$. We summarized the case of $\pi_1 = 1/3$ above. When $\pi_1 = 1/2$, the expected sample size averaged over scenarios (a)-(c) is 531, 610, and 741 for the adaptive design, the single standard design, and the paired standard design, respectively. When $\pi_1 \geqslant 1/2$, the closest competitor to the adaptive design is the single standard design, while at $\pi_1 < 1/2$ the closest competitor is the paired standard design.

Consider the adaptive design versus the single standard design. The former has the ability to stop enrolling subpopulation 2 and switch to only enrolling 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 $\pi_1 \in \{2/3, 5/6\}$, the optimal adaptive design is only slightly better than the optimal, single standard design. At $\pi_1 = 5/6$, these designs are almost identical.

Next, we compare the adaptive design to the paired standard design. The adaptive design's multiple testing procedure is generally more efficient than that of the paired standard design. This is because the adaptive design does not partition the data between tests of H_{0C} and H_{01}, and the adaptive design leverages the correlation between $\mathbf{Z}_1$ and $\mathbf{Z_C}$. The improved efficiency of the adaptive design is amplified as π_1 increases, due to subpopulation 1 making up more of the combined population, which increases the correlation between $\mathbf{Z}_1$ and $\mathbf{Z_C}$.

For the optimal adaptive design, the expected sample size averaged over scenarios (a)-(c), is decreasing in π_1; the same holds for the standard single design. However, for the optimal, paired standard design, the expected sample size gradually increases as π_1 increases. The reason is that due to the assumptions (4.1) on p_{s0} and p_{s1}, we have $\sigma^2_{sa} = p_{sa}(1 - p_{sa})$ is greater for subpopulation $s = 1$ than for subpopulation $s = 2$, for each $a \in \{0, 1\}$ and each scenario (a)-(c).

Therefore, for larger π_1, larger sample sizes are needed to achieve goal (i) in the first component of the paired standard design, while the second component is unaffected by π_1.

The maximum sample size occurs when there is no early stopping. Figure 1b compares the maximum sample size for each of the three types of designs. The maximum sample size of the adaptive design is the smallest among the three types of designs, at all values of π_1 except $\pi_1 = 1/3$ (where it is 1136 for the adaptive design and 1120 for the paired standard design).

Recall that k^* is the parameter in our adaptive designs representing the last stage at which subpopulation 2 can be enrolled. For each $\pi_1 \in \{1/6, 1/3, 1/2, 2/3, 5/6\}$, the corresponding value of k^* for the optimal adaptive design from our search is 3, 3, 4, 5, and 5, respectively. To demonstrate the usefulness of considering values of $k^* < 5$ for $\pi_1 \in \{1/6, 1/3, 1/2\}$, we repeated our adaptive design search algorithm from Section 5.2 but only allowing designs with $k^* = 5$. Comparing the resulting adaptive design to D_{ADAPT}, there was an increase in the expected sample size averaged over scenarios (a)-(c) of 66%, 9%, and 1%, for $\pi_1 \in \{1/6, 1/3, 1/2\}$, respectively. The large difference at small values of π_1 makes sense intuitively, since it is for such values of π_1 where enrollment is much faster for subpopulation 2 than for subpopulation 1, so that enrolling subpopulation 2 through the end of stage 5 would be quite inefficient for achieving goals (i)-(iii).

6.3 *Impact of Estimated Variances and Subpopulation Proportions*

Above, for clarity, we focused on the case where variances σ^2_{sa} and the subpopulation proportions π_1, π_2 are known. Also, we assumed at each stage in which both subpopulations are enrolled, the ratio of small IVH participants to large IVH participants equals the ratio of the corresponding population sizes. We examine the impact of relaxing these assumptions.

We now consider the setting where variances and subpopulation proportions are not precisely known in advance. However, we assume that the value of π_1 is in a known, relatively small range, which we take to be 0.33 ± 0.05 in the context of the MISTIE trial. The null hypotheses

defined in Section 4.2 remain the same. We assume that during each stage k where the combined population is enrolled, the total number enrolled is n_k, and the subpopulation membership $S_{i,k}$ of each participant is an independent draw with probability π_1 of $S_{i,k} = 1$. We replace the variances used in constructing the statistics (4.3) by corresponding sample variances. We estimate Σ by $\hat{\Sigma}$ using the formulas in Appendix A.1 of the Supplementary Materials, using sample variances and the sample proportion $\hat{\pi}_1$.

In the case where Σ was assumed known, the boundaries $\mathbf{u}$ and $\mathbf{l}$ for our adaptive designs were selected such that $\alpha_0(\mathbf{u}, \Sigma) \leqslant \alpha$, which by Theorem 5.1 guarantees strong control of the familywise Type I error rate at level α. In Appendix B of the Supplementary Materials, we generalize the adaptive design algorithm from Section 5.1 to handle estimated Σ and π_1. This is based on an extension of the error spending approach of Slud and Wei (1982) and Lan and DeMets (1983) 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 (4.1) hold, goals (i)-(iii) are approximately achieved. Even when these assumptions fail to hold, we have $\alpha_0(\mathbf{U}, \hat{\Sigma}) \leqslant \alpha$, where $\mathbf{U} = \{U_{C,k}\}_{k=1}^{k^*} \cup \{U_{1,k}\}_{k=1}^{K}$; we show in Appendix B of the Supplementary Materials that under certain conditions on the convergence of $\hat{\Sigma}$, $\hat{\pi}_1$, and $\mathbf{U}$, the design D^*_{ADAPT} strongly controls the familywise Type I error rate. We next examine the power and Type I error of D^*_{ADAPT} based on simulations.

We 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. In the setting where Σ and π_1 are estimated, we computed analogous quantities for D^*_{ADAPT} under each scenario (one per row) in Table 4, under the assumptions (4.1). We simulated 5×10^4 trials in each scenario. The maximum absolute difference in power, comparing each entry in Table 4 to its corresponding value computed under the assumptions in this section, was 0.02. The maximum absolute difference in the familywise

Type I error rate was always at most 10^{-3}. This shows that at the sample sizes and scenarios we considered, the performance of D^*_{ADAPT} in the setting of unknown variances and π_1 is similar to that of D_{ADAPT} when these quantities are known and (4.1) holds; in particular, goals (i)-(iii) are approximately achieved by D^*_{ADAPT}.

In Appendix B.4 of the Supplementary Materials, we examine the impact of true subpopulation proportions differing by $\pm 5\%$ 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 sets of simulations where the true value of π_1 was set to either 0.28 or to 0.38, under the assumptions of this subsection. 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 77% 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 0.024. This shows the design D^*_{ADAPT} is reasonably 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 and magnitudes of deviations from the assumptions (4.1).

7. Discussion

An alternative to restricting to the class of designs in Section 5.2 would be to extend the dynamic programming method of Eales and Jennison (1992), which was designed for a single hypothesis in standard group sequential designs, 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 on each design (especially that of strong control of the familywise Type I error rate).

A limitation of our approach is that we assumed π_1 is either known exactly, or (as in Section 6.3) is within a relatively small, known range. We think this may be a reasonable assumption in some, but certainly not all, settings. An area of future work is to develop designs that are robust

to large deviations of the true π_1 from the value used in planning the trial.

It is possible to consider other 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 procedure is an example of an adaptive enrichment design, since enrollment can be modified (restricted to subpopulation 1) in response to accrued data. This procedure can be considered a crude version of the adaptive designs above, with the main differences being that those above have a more flexible rule for modifying enrollment and they use an optimized multiple testing procedure that leverages the correlation among test statistics. Similarly, one could extend the standard paired design to merge the data from both component designs at each interim analysis, to be used in deciding which hypotheses to reject and which subpopulations to stop. This, too, is an adaptive enrichment design that can be considered a crude version of the adaptive designs in Section 5.1.

It is possible to consider efficacy boundaries in our search algorithm that are proportional to families of boundaries other than those of O'Brien and Fleming. This is an area of future research.

An open problem is to consider different goals than (i)-(iii), e.g., to additionally include a power requirement for rejecting the null hypothesis of no average treatment benefit for the subpopulation with large IVH, under a scenario where the treatment only benefits this population. Our adaptive designs from Section 5.1 could be applied to achieve this set of design goals, using a different search algorithm taking this additional goal into account.

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

Acknowledgements

This research was supported by U.S. National Institute of Neurological Disorders and Stroke (grant numbers 5R01 NS046309-07 and 5U01 NS062851-04) and the U.S. Food and Drug Administration through the "Partnership in Applied Comparative Effectiveness Science," (contract HHSF2232010000072C). This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the above agencies.

References

Boessen, R., van der Baan, F., Groenwold, R., Egberts, A., Klungel, O., Grobbee, D., Knol, M. and Roes, K. (2013). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics*.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. and Racine-Poon, A. (2009). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* **28**(10), 1445–1463.

Eales, J. D. and Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**(1), 13–24.

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

Follman, D.A., Proschan, M.A. and Geller, N.L. (1994). Monitoring multi-armed trials. *Biometrics* **50**, 325–336.

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

Götze, F. (1991). On the rate of convergence in the multivariate clt. *The Annals of Probability* **19**(2), 724–739.

Hanley, D. (2012). http://braininjuryoutcomes.com/studies/mistie/entry/mistie/international-stroke-conference-2012-mistie-phase-2-results.

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

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley Interscience.

Jenkins, M., Stone, A. and Jennison, C. (2011). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* **10**(4), 347–356.

Jennison, C. and Turnbull, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press.

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

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

Liu, Q. and Anderson, K. M. (2008). On adaptive extensions of group sequential trials for clinical investigations. *J. Amer. Statist. Assoc.* **103**(484), 1621–1630.

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

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

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Rosenblum, M. and van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* **98**(4), 845–860.

Russek-Cohen, E and Simon, R. M. (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine* **16**, 455–464.

Slud, E. V. and Wei, L-J. (1982). Two-sample repeated significance tests based on the modified wilcoxon statistic. *J. Amer. Statist. Assoc.* **77**, 862–868.

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

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

Wang, S. J., O'Neill, R. T. and Hung, H. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist.* **6**, 227–244.

Table 1. Adaptive design D_{ADAPT} (at $\pi_1 = 1/3$). No boundaries are given for $u_{C,k}$ and $l_{2,k}$ at the stage four and five analyses, since here $k^* = 3$, i.e., enrollment of large IVH participants always stops after interim analysis 3. Cumulative sample sizes are for the case where no early stopping occurs; specifically, cumulative sample size for the combined population at stage k equals $\sum_{k'=1}^{k} n_{k'}$. All boundaries are on the z-statistic scale.

Interim Analysis (k)	1	2	3	4	5
Cum. Sample Size Subpopulation 1	93	187	280	428	576
Cum. Sample Size Subpopulation 2	187	373	560	560	560
Cum. Sample Size Combined Population	280	560	840	988	1136
H_{0C} Efficacy Boundary ($u_{C,k}$)	4.93	3.49	2.85		
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	0	0	∞		
H_{01} Efficacy Boundary ($u_{1,k}$)	5.09	3.60	2.94	2.38	2.05
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.05

Table 2. Standard design $D_{\text{STD,SINGLE}}$ (at $\pi_1 = 1/3$).

Interim Analysis (k)	1	2	3	4	5
Cum. Sample Size Subpopulation 1	90	180	270	360	526
Cum. Sample Size Subpopulation 2	180	360	540	720	1053
Cum. Sample Size Combined Population	270	540	810	1080	1579
H_{0C} Efficacy Boundary ($u_{C,k}$)	6.87	4.86	3.97	3.44	2.84
H_{01} Efficacy Boundary ($u_{1,k}$)	4.90	3.46	2.83	2.45	2.02
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.02

Table 3. Standard design $D_{\text{STD,PAIR}}$ (at $\pi_1 = 1/3$).

Interim Analysis (k)	1	2	3	4	5
First Component Design (Enrolls Combined Population)					
Cum. Sample Size Combined Population	108	216	324	432	540
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.14	3.64	2.97	2.57	2.30
H_{0C} Futility Boundary	0	0	0	0	2.30
Second Component Design (Enrolls Subpopulation 1 Only)					
Cum. Sample Size Subpopulation 1	116	232	348	464	580
H_{01} Efficacy Boundary ($u_{1,k}$)	5.15	3.65	2.98	2.58	2.30
H_{01} Futility Boundary	0	0	0	0	2.30

Table 4. Comparison of expected sample size (ESS) and power (as a percent), for the adaptive design and standard designs, at $\pi_1 = 1/3$. Each row corresponds to a different percent benefit for large IVH participants, i.e., Δ_2. Three power values are given per design and scenario: the power to reject at least H_{0C}; to reject at least H_{01}; and to reject at least one null hypothesis, respectively.

Small IVH percent benefit Δ_1 set at 12.5%

Large IVH Benefit	D_{ADAPT}				$D_{\text{STD,SINGLE}}$				$D_{\text{STD,PAIR}}$			
(%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	623	86	10	89	783	84	31	89	802	89	80	98
12.5	674	80	18	88	856	80	42	89	823	80	80	96
10.0	726	68	30	86	938	70	55	88	837	68	80	94
5.0	782	32	63	83	1055	33	76	83	840	37	80	87
0.0	716	6	80	82	1075	4	80	80	795	12	80	82
-5.0	618	0	82	82	1077	0	80	80	714	2	80	81

Small IVH percent benefit Δ_1 set at 0

Large IVH Benefit	D_{ADAPT}				$D_{\text{STD,SINGLE}}$				$D_{\text{STD,PAIR}}$			
(%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	493	32.4	0.1	32.5	574	30.1	0.4	30.2	701	60.8	1.1	61.2
12.5	523	24.8	0.3	25.1	622	27.2	0.8	27.3	702	44.9	1.1	45.6
10.0	548	16.0	0.7	16.6	669	21.7	1.2	21.9	691	29.6	1.1	30.3
5.0	562	3.3	1.5	4.8	717	5.2	1.9	6.4	634	8.2	1.1	9.3
0.0	517	0.2	1.9	2.1	723	0.2	2.1	2.2	549	1.2	1.1	2.3
-5.0	467	0.0	2.0	2.0	721	0.0	2.1	2.1	472	0.1	1.1	1.2

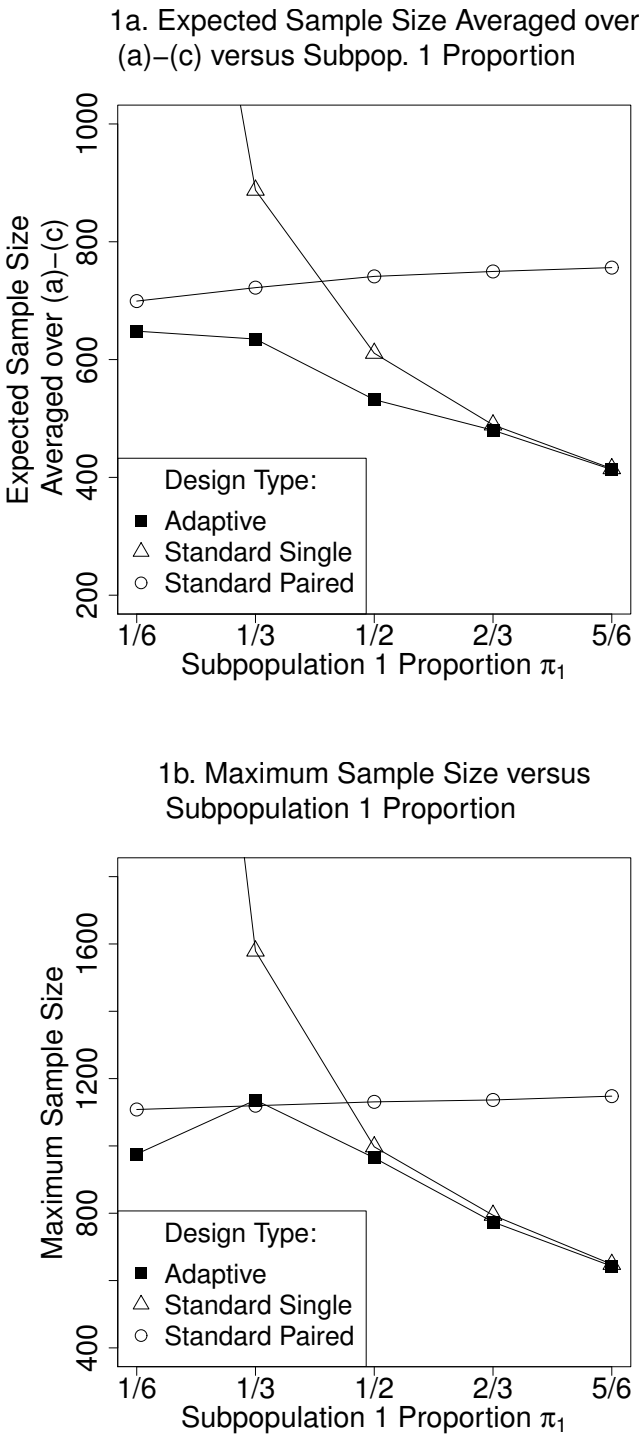

Fig. 1. Comparison of optimal designs as subpopulation 1 proportion π_1 is varied. Fig. 1a: Expected sample size averaged over scenarios (a)-(c). Fig. 1b: Maximum possible sample size, which occurs if there is no early stopping of any kind. In both plots, the leftmost point for the standard single design is not shown since it exceeds the plot maximum.

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

	Pages
Appendix A, Proof of Theorem 5.1	1-6
Appendix B, When Variances and π_1 are Unknown and Estimated	6-13
Appendix C, Search Algorithm to Select Adaptive and Standard Designs	13-18
Appendix D, Optimal Designs from Search Algorithm for Various π_1 Values	18-19
Appendix E, Reproducibility in a Follow-up Trial	19-20

A. Proof of Theorem 5.1

In Section A.1, we give the definitions of $Z_{1,k}$ and $Z_{2,k}$, and also give formulas for the components of Σ. We prove Theorem 5.1 in Section A.2. We describe related work based on combination tests and the closed testing principle in Section A.3.

A.1 *Definitions of $Z_{1,k}$ and $Z_{2,k}$*

Let $1(S_{i,k} = s)$ denote the indicator variable taking value 1 if $S_{i,k} = s$ and zero otherwise. For each $k \leqslant K$, define

$$Z_{1,k} = \left\{ \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} A_{i,k'} 1(S_{i,k'} = 1)}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} A_{i,k'} 1(S_{i,k'} = 1)} - \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} (1 - A_{i,k'}) 1(S_{i,k'} = 1)}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} (1 - A_{i,k'}) 1(S_{i,k'} = 1)} \right\} \mathrm{se}_{1,k}^{-1}, \tag{A.4}$$

and for each $k \leqslant k^*$, define

$$Z_{2,k} = \left\{ \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} A_{i,k'} 1(S_{i,k'} = 2)}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} A_{i,k'} 1(S_{i,k'} = 2)} - \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} (1 - A_{i,k'}) 1(S_{i,k'} = 2)}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} (1 - A_{i,k'}) 1(S_{i,k'} = 2)} \right\} \mathrm{se}_{2,k}^{-1}, \tag{A.5}$$

where for each $s \in \{1,2\}$, $\text{se}_{s,k} = \left\{ \left(\sigma_{s0}^2 + \sigma_{s1}^2\right) / \sum_{k'=1}^{k} (\pi_s n_{k'}/2) \right\}^{1/2}$.

We next consider the covariance matrix Σ of the variables $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$. For any k_1, k_2, define $f(k_1,k_2) = \left(\sum_{k'=1}^{\min\{k_1,k_2\}} n_{k'} \Big/ \sum_{k'=1}^{\max\{k_1,k_2\}} n_{k'} \right)^{1/2}$. For any $s \in \{1,2\}$, define $\rho_s = \left\{ \pi_s \left(\sigma_{s0}^2 + \sigma_{s1}^2\right) \Big/ \sum_{s'=1}^{2} \pi_{s'} \left(\sigma_{s'0}^2 + \sigma_{s'1}^2\right) \right\}^{1/2}$. Then for any $s \in \{1,2\}$, and k_1, k_2,

$$Cov(Z_{1,k_1}, Z_{2,k_2}) = 0, \qquad Cov(Z_{C,k_1}, Z_{C,k_2}) = Cov(Z_{s,k_1}, Z_{s,k_2}) = f(k_1,k_2),$$

$$Cov(Z_{C,k_1}, Z_{s,k_2}) = \rho_s f(k_1,k_2), \tag{A.6}$$

where any expression above involving $Z_{C,k}$ or $Z_{2,k}$ is understood to only be defined for $k \leqslant k^*$.

A.2 *Proof of Theorem* 5.1

Proof: The covariance matrix Σ depends on the data generating distribution $\mathbf{p}$, and we make this dependence explicit below by writing $\Sigma(\mathbf{p})$. In this subsection, we assume $\Sigma(\mathbf{p})$ and π_1 are known. We examine the case of $\Sigma(\mathbf{p})$ and π_1 estimated rather than known, in Section B.

Consider any $\alpha \in (0,1)$, $\pi_1 \in (0,1)$, $k^* \leqslant K$, and $(r_1, \ldots, r_K)$. We consider these values fixed, i.e., they do not change with sample size. For any $n > 0$, consider the per-stage sample sizes $(n_1, \ldots, n_K) = (nr_1, \ldots, nr_K)$ defined in Section 4.2. We will prove the following:

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \sup_{\mathbf{u}:\alpha_0(\mathbf{u},\Sigma(\mathbf{p}))\leqslant\alpha} \Pr_{\mathbf{p},D(\mathbf{u},\mathbf{l}),n_1,\ldots,n_K}(\text{at least one true null hypothesis is rejected}) \leqslant \alpha. \tag{A.7}$$

This implies that for any efficacy boundaries $\mathbf{u}$ such that $\alpha_0(\mathbf{u}, \Sigma(\mathbf{p})) \leqslant \alpha$, and for any $\mathbf{l}$, we have the adaptive design $D(\mathbf{u}, \mathbf{l})$ strongly controls the familywise Type I error rate at level α.

The proof of (A.7) is in two parts. First we show that for any $n_1, \ldots, n_K$, boundaries $\mathbf{u}, \mathbf{l}$, and distribution $\mathbf{p} \in \mathcal{P}$, the probability of a familywise Type I error for $D(\mathbf{u}, \mathbf{l})$ is at most

$$\Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k \leqslant k^*, Z_{C,k} - EZ_{C,k} > u_{C,k}, \text{ or for some } k \leqslant K, Z_{1,k} - EZ_{1,k} > u_{1,k}), \tag{A.8}$$

where $\Pr_{\mathbf{p},n_1,\ldots,n_K}$ is the joint probability distribution for $(\mathbf{Z}_C, \mathbf{Z}_1)$, which depends on $\mathbf{p}$ and $n_1, \ldots, n_K$. In the second part of the proof, we upper bound (A.8), uniformly over $\mathbf{p} \in \mathcal{P}$, using the uniform central limit theorem of Götze (1991).

We next define a design $D'(\mathbf{u}, \mathbf{l})$ that dominates $D(\mathbf{u}, \mathbf{l})$, in that $D'(\mathbf{u}, \mathbf{l})$ rejects each null hypothesis whenever $D(\mathbf{u}, \mathbf{l})$ does; we then upper bound the familywise Type I error rate of $D'(\mathbf{u}, \mathbf{l})$ by (A.8). Define $D'(\mathbf{u}, \mathbf{l})$ to have the same per-stage enrollment as $D(\mathbf{u}, \mathbf{l})$, except $D'(\mathbf{u}, \mathbf{l})$ never stops early. Specifically, for each stage $k \leqslant k^*$, $D'(\mathbf{u}, \mathbf{l})$ enrolls n_k participants from the combined population, and for each stage $k : k^* < k \leqslant K$, $D'(\mathbf{u}, \mathbf{l})$ enrolls n_k participants from subpopulation 1. The design $D'(\mathbf{u}, \mathbf{l})$ rejects H_{0C} if for some $k \leqslant k^*$, $Z_{C,k} > u_{C,k}$; $D'(\mathbf{u}, \mathbf{l})$ rejects H_{01} if for some $k \leqslant K$, $Z_{1,k} > u_{1,k}$. (Design $D'(\mathbf{u}, \mathbf{l})$ does not stop enrollment or testing after rejecting a null hypothesis, and so may reject both null hypotheses.) It follows from the definition of the adaptive design $D(\mathbf{u}, \mathbf{l})$ in Section 5.1 that $D'(\mathbf{u}, \mathbf{l})$ dominates $D(\mathbf{u}, \mathbf{l})$.

To show the familywise Type I error rate of $D'(\mathbf{u}, \mathbf{l})$ is at most (A.8), we consider three cases. First, consider the case where both H_{0C} and H_{01} are true. Then the familywise Type I error rate of $D'(\mathbf{u}, \mathbf{l})$ is

$$\begin{aligned}
&\Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k \leqslant k^*, Z_{C,k} > u_{C,k}, \text{ or for some } k \leqslant K, Z_{1,k} > u_{1,k}) \\
&\leqslant \Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k \leqslant k^*, Z_{C,k} - EZ_{C,k} > u_{C,k}, \text{ or for some } k \leqslant K, Z_{1,k} - EZ_{1,k} > u_{1,k}),
\end{aligned}$$

where the second line follows since $EZ_{C,k} \leqslant 0$ when H_{0C} is true and $EZ_{1,k} \leqslant 0$ when H_{01} is true. Second, consider the case where only H_{0C} is true. Then the familywise Type I error rate of $D'(\mathbf{u}, \mathbf{l})$ is

$$\begin{aligned}
&\Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k \leqslant k^*, Z_{C,k} > u_{C,k}) \\
&\leqslant \Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k \leqslant k^*, Z_{C,k} - EZ_{C,k} > u_{C,k}), \\
&\leqslant \Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k \leqslant k^*, Z_{C,k} - EZ_{C,k} > u_{C,k}, \text{ or for some } k \leqslant K, Z_{1,k} - EZ_{1,k} > u_{1,k})
\end{aligned}$$

where the second line follows since $EZ_{C,k} \leqslant 0$ when H_{0C} is true. For the third case, where

only H_{01} is true, the familywise Type I error rate of $D'(\mathbf{u}, \mathbf{l})$ is at most (A.8), by an analogous argument as the second case but using that $EZ_{1,k} \leqslant 0$. In each case, we showed the familywise Type I error rate of $D'(\mathbf{u}, \mathbf{l})$ is at most (A.8). Since $D'(\mathbf{u}, \mathbf{l})$ dominates $D(\mathbf{u}, \mathbf{l})$, this proves the familywise Type I error rate of $D(\mathbf{u}, \mathbf{l})$ is at most (A.8).

The above argument implies that the bound (A.8) on the familywise Type I error holds even if we allow arbitrary rules for switching from enrolling from the combined population to enrolling only small IVH participants, e.g., allowing a data safety monitoring board to decide on such a switch if a high adverse event rate is observed in the large IVH participants. This follows since such a procedure can only reject a null hypothesis when the corresponding z-statistic exceeds its efficacy boundary, which implies $D'(\mathbf{u}, \mathbf{l})$ rejects it as well, and we showed above that $D'(\mathbf{u}, \mathbf{l})$ has familywise Type I error rate at most (A.8).

We next upper bound (A.8). By the assumptions in Section 4.2 and the multivariate, Berry-Esseen type central limit theorem of Götze (1991), the joint distribution of the centered z-statistics $\mathbf{Z}^{(c)} = \{\mathbf{Z}_C - E(\mathbf{Z}_C), \mathbf{Z}_1 - E(\mathbf{Z}_1)\}$ converges uniformly over $\mathcal{P}$ to a mean zero, multivariate normal distribution with covariance matrix $\Sigma(\mathbf{p})$ restricted to the statistics $(\mathbf{Z}_C, \mathbf{Z}_1)$. More precisely, this theorem implies

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \sup_{A\in\mathcal{C}\cup\mathcal{C}^c} |\Pr_{\mathbf{p},n_1,\ldots,n_K}(\mathbf{Z}^{(c)} \in A) - \Pr_{\Sigma(\mathbf{p})}(\mathbf{Z}' \in A)| = 0, \tag{A.9}$$

where $\mathcal{C}$ is the class of convex, Borel measurable sets in $\mathbb{R}^{(k^*+K)}$, $\mathcal{C}^c$ is the class of complements of convex, Borel measurable sets in $\mathbb{R}^{(k^*+K)}$, and $\mathbf{Z}'$ was defined in Section 5.1. For any $\mathbf{u} \in \mathbb{R}^{(k^*+K)}$, define the set $A(\mathbf{u}) = \{\mathbf{z} \in \mathbb{R}^{(k^*+K)} : \text{for some } k \leqslant k^*, z_k > u_k, \text{ or for some } k \leqslant K, z_{k+k^*} > u_{k+k^*}\}$. Since for any $\mathbf{u} \in \mathbb{R}^{(k^*+K)}$ we have $A(\mathbf{u}) \in \mathcal{C}^c$, it follows from (A.9) that

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \sup_{\mathbf{u}\in\mathbb{R}^{(k^*+K)}} |\Pr_{\mathbf{p},n_1,\ldots,n_K}(\mathbf{Z}^{(c)} \in A(\mathbf{u})) - \Pr_{\Sigma(\mathbf{p})}(\mathbf{Z}' \in A(\mathbf{u}))| = 0.$$

This implies

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \sup_{\mathbf{u}:\alpha_0(\mathbf{u},\Sigma(\mathbf{p}))\leqslant\alpha}$$
$$|\Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k\leqslant k^*, Z_{C,k}-EZ_{C,k}>u_{C,k}, \text{ or for some } k\leqslant K, Z_{1,k}-EZ_{1,k}>u_{1,k})$$
$$-\Pr_{\Sigma(\mathbf{p})}(\text{for some } k\leqslant k^*, Z'_{C,k}>u_{C,k}, \text{ or for some } k\leqslant K, Z'_{1,k}>u_{1,k})\big| = 0. \tag{A.10}$$

Since by definition $\alpha_0(\mathbf{u},\Sigma(\mathbf{p})) = \Pr_{\Sigma(\mathbf{p})}(\text{for some } k\leqslant k^*, Z'_{C,k}>u_{C,k} \text{ or for some } k\leqslant K, Z'_{1,k}>u_{1,k})$, it follows from (A.10) that the display below is at most α:

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \sup_{\mathbf{u}:\alpha_0(\mathbf{u},\Sigma(\mathbf{p}))\leqslant\alpha}$$
$$\Pr_{\mathbf{p},n_1,\ldots,n_K}(\text{for some } k\leqslant k^*, Z_{C,k}-EZ_{C,k}>u_{C,k}, \text{ or, for some } k\leqslant K, Z_{1,k}-EZ_{1,k}>u_{1,k}).$$

Combining this with the result from the first part of the proof, that for any $n_1,\ldots,n_K$, $\mathbf{u},\mathbf{l}$, and $\mathbf{p}\in\mathcal{P}$, the probability of a familywise Type I error for the design $D(\mathbf{u},\mathbf{l})$ is at most (A.8), proves that (A.7) holds. This completes the proof of Theorem 1.

A.3 *Related work based on combination tests and the closed testing principle*

Jennison and Turnbull (2007, Section 6), Brannath *and others* (2009), Jenkins *and others* (2011), and Stallard (2011) present seamless Phase II/III designs that allow subpopulation enrichment, based on the combination test approach, and the closed testing principle of Marcus *and others* (1976). That approach involves defining elementary null hypotheses, which would be H_{01} and H_{0C} in our context. For each intersection of elementary null hypotheses, that approach combines data generated in stage k into a p-value p_k using, e.g., a Bonferroni, Dunnett or other intersection test; the p_k are then combined across stages using, e.g., an inverse-normal or Fisher combination test. Each elementary null hypothesis is rejected if all intersection tests (referred to as local tests) involving it reject. Our adaptive designs do not use this approach, and instead directly use sufficient statistics at each stage (i.e., the z-statistics defined above), and the correlation among these statistics, in testing each null hypothesis of interest. It may be simpler to communicate

our designs, in some cases, since one can present the stopping rule and testing procedure in a single table such as Table 1 that contains group sequential boundaries for each population of interest. However, an advantage of the approach in the above related work is that it allows certain types of unplanned changes to the sample sizes of future stages based on current stage data, while our method requires predefining each n_k. The potential tradeoff is more flexibility in the former approach, versus more efficiency in our approach that fully leverages the correlation Σ of the statistics $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$. It is an area of future work to contrast the performance of designs based on the approach of the above related work versus our approach, to learn the strengths and weaknesses of each.

B. When Variances and π_1 Are Unknown and Estimated

Above, we assumed the covariance matrix Σ is known. In practice it will be estimated, e.g., by substituting the sample variance $\hat{\sigma}^2_{sa}$ for the corresponding true (unknown) variance σ^2_{sa} in (A.6). Similarly, π_1 will in practice not be precisely known before the study starts. In this section, we make the assumptions from Section 6.3.

At the end of each stage k, we estimate each σ^2_{sa} by the corresponding sample variance $\hat{\sigma}^2_{sa}$, using all data in stages 1 through k; we estimate π_s by the proportion $\hat{\pi}_s$ of the sample in subpopulation s, using data in all stages $k' \leqslant k$ in which the combined population was enrolled. Define the modified z-statistics $\bar{Z}_{C,k}, \bar{Z}_{1,k}, \bar{Z}_{2,k}$ to be as in (4.3), (A.4), and (A.5), respectively, except replacing the standard errors $\mathrm{se}_{C,k}$ and $\mathrm{se}_{s,k}$ for each $s \in \{1, 2\}$ by the estimated quantities $\bar{\mathrm{se}}_{C,k} = \left[\sum_{s=1}^{2} \hat{\pi}_s \left\{\hat{\sigma}^2_{s0} + \hat{\sigma}^2_{s1}\right\} / \sum_{k'=1}^{k} (n_{k'}/2)\right]^{1/2}$, and $\bar{\mathrm{se}}_{s,k} = \left\{\left(\hat{\sigma}^2_{s0} + \hat{\sigma}^2_{s1}\right) / \sum_{k'=1}^{k} (\hat{\pi}_s n_{k'}/2)\right\}^{1/2}$, respectively.

We generalize the algorithm from Section 5.1 to handle estimated Σ and π_1. This is based on an extension of the error spending approach of Slud and Wei (1982) and Lan and DeMets (1983) to handle multiple hypotheses. The result is an algorithm for constructing efficacy boundaries

$U_{C,k}$ and $U_{1,k}$ at the end of each stage k, described below.

In our proposed procedure, before the trial starts, one specifies nonnegative constants $\alpha_{C,k}, \alpha_{1,k}$ satisfying $\sum_{k=1}^{k^*} \alpha_{C,k} + \sum_{k=1}^{K} \alpha_{1,k} \leqslant \alpha$. The intuition is that at interim analysis k, $\alpha_{C,k}$ represents the Type I error to be spent on the test of H_{0C}, and $\alpha_{1,k}$ represents the Type I error to be spent on the test of H_{01}, in the adaptive, group sequential algorithm in Section 5.1; we make this rigorous below. We first give our algorithm for constructing efficacy boundaries at the end of each stage k; these are based on $\alpha_{C,k}, \alpha_{1,k}$ and the estimated covariance matrix for the statistics observed up through stage k. We then show in Section B.2 how to prespecify values of $\alpha_{C,k}$ and $\alpha_{1,k}$ such that the procedure approximately achieves goals (i)-(iii) from Section 2 under assumptions (4.1). In Section B.3, we give a convergence condition under which this procedure strongly controls the familywise Type I error rate. In Section B.4, we examine the performance of this procedure when the true value of π_1 differs from that planned for.

B.1 *Algorithm to generate efficacy boundaries $U_{C,k}$ and $U_{1,k}$ at the end of each stage k*

The algorithm takes as input nonnegative constants $\alpha_{C,k}, \alpha_{1,k}$ satisfying $\sum_{k=1}^{k^*} \alpha_{C,k} + \sum_{k=1}^{K} \alpha_{1,k} \leqslant \alpha$, which are set before the trial starts. During the trial, at the end of each stage k, we construct an estimate $\hat{\Sigma}$ of the covariance matrix for the statistics observed up through stage k, and an estimate $\hat{\pi}_s$ for π_s, as described above. We then construct boundaries $U_{C,k}$ and $U_{1,k}$ based on $\hat{\Sigma}$ and the previously computed values $U_{C,k'}$ and $U_{1,k'}$ for stages $k' < k$, as follows: If $k \leqslant k^*$, we set the boundary $U_{C,k}$ to be the solution $u_{C,k}$ to

$$\Pr_{\hat{\Sigma}}(\text{for all } k' < k, Z'_{C,k'} \leqslant u_{C,k'}, Z'_{1,k'} \leqslant u_{1,k'}; \text{ and } Z'_{C,k} > u_{C,k}) = \alpha_{C,k}, \tag{B.11}$$

where we set each $u_{C,k'} = U_{C,k'}$ and $u_{1,k'} = U_{1,k'}$ for $k' < k$. We set the boundary $U_{1,k}$ to be the solution $u_{1,k}$ to

$$\Pr_{\hat{\Sigma}}(\text{for all } k' \leqslant \min\{k, k^*\}, Z'_{C,k'} \leqslant u_{C,k'};\ \text{for all } k' < k, Z'_{1,k'} \leqslant u_{1,k'}; Z'_{1,k} > u_{1,k}) = \alpha_{1,k}, \tag{B.12}$$

where we set each $u_{C,k'} = U_{C,k'}$ for $k' \leqslant \min\{k, k^*\}$ and $u_{1,k'} = U_{1,k'}$ for $k' < k$. By convention, if $\alpha_{C,k} = 0$, we set $U_{C,k} = \infty$; similarly, if $\alpha_{1,k} = 0$, we set $U_{1,k} = \infty$. We use uppercase U in the above notation to emphasize these quantities are random variables that depend on sample variances and proportions described above. Each $U_{C,k}$ can be approximated via binary search, where for each candidate value, the left hand side of (B.11) can be evaluated by computing the corresponding probability based on the multivariate normal distribution with covariance $\hat{\Sigma}$ and mean $\mathbf{0}$. The same method can be used to approximate $U_{1,k}$. In our computations we executed a binary search until the difference between the upper and lower bounds was less than 10^{-4}. The reason we do not include the futility boundaries $l_{1,k'}, l_{2,k'}$, in the above computation is that we consider them nonbinding futility boundaries, as discussed in Section 5.

In our R code, we used a slight modification of the above procedure, to deal with $\hat{\Sigma}$ being updated after each stage. We executed the above algorithm, except that we replaced (B.11) and (B.12) by the cumulative versions:

$$\begin{aligned}
&\Pr_{\hat{\Sigma}}(\text{for some } k' \leqslant k, Z'_{C,k'} > u_{C,k'},\ \text{or for some } k' < k, Z'_{1,k'} > u_{1,k'},) \\
&= \sum_{k'=1}^{k} \alpha_{C,k'} + \sum_{k'=1}^{k-1} \alpha_{1,k'};\ \text{and,} \qquad \text{(B.13)} \\
&\Pr_{\hat{\Sigma}}(\text{for some } k' \leqslant \min\{k, k^*\}, Z'_{C,k'} > u_{C,k'},\ \text{or for some } k' \leqslant k, Z'_{1,k'} > u_{1,k'}) \\
&= \sum_{k'=1}^{\min\{k,k^*\}} \alpha_{C,k'} + \sum_{k'=1}^{k} \alpha_{1,k'}, \qquad \text{(B.14)}
\end{aligned}$$

respectively. Define $\mathbf{U} = \{U_{C,k}\}_{k=1}^{k^*} \cup \{U_{1,k}\}_{k=1}^{K}$. It follows that if $U_{C,k}$ and $U_{1,k}$ are computed using the above equations after each stage k, we have $\alpha_0(\mathbf{U}, \hat{\Sigma}) \leqslant \sum_{k=1}^{k^*} \alpha_{C,k} + \sum_{k=1}^{K} \alpha_{1,k} \leqslant \alpha$,

for $\hat{\Sigma}$ computed after the last stage of the trial (i.e., when enrollment stops completely). The advantage of using (B.13) and (B.14) instead of (B.11) and (B.12), roughly speaking, is that if estimates of Σ at later stages indicate earlier stage boundaries were set too low, the algorithm automatically recovers by spending less α at the current stage. It is possible, in principle, that due to differences between early and later stage estimates of Σ, the above equations may not have any solution. This is a general issue in the error-spending approach. It never occurred in any of our simulations.

We next define our generalization of the adaptive design from Section 5.1 to the case of estimated Σ and π_1. First, before the trial starts, constants $\alpha_{C,k}, \alpha_{1,k}$ are prespecified, e.g., using the algorithm in the next subsection. Futility boundaries $\mathbf{l}$, per-stage maximum sample sizes $(n_1, \ldots, n_K)$, and k^* are also prespecified. Next, the trial is started. At the analysis after each stage k, efficacy boundaries $U_{C,k}, U_{1,k}$ are computed based on the algorithm above. We then execute the adaptive design decision rule at stage k from Section 5.1 using efficacy boundaries $u_{C,k} = U_{C,k}, u_{1,k} = U_{1,k}$, z-statistics $\bar{Z}_{C,k}, \bar{Z}_{1,k}, \bar{Z}_{2,k}$, and futility boundaries $\mathbf{l}$. We call the above procedure the modified adaptive design.

B.2 *Example method to set values $\alpha_{C,k}$ and $\alpha_{1,k}$.*

We next give a method to set $\alpha_{C,k}, \alpha_{1,k}$ before the trial starts such that $\sum_{k=1}^{k^*} \alpha_{C,k} + \sum_{k=1}^{K} \alpha_{1,k} \leqslant \alpha$ and goals (i)-(iii) are approximately achieved by the modified adaptive design, when the assumptions (4.1) hold.

Let Σ' denote the covariance matrix of the z-statistics $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$ given in (A.6) under H_0 and assumptions (4.1). Consider the efficacy boundaries of D_{ADAPT} given in Table 1, which we denote by $\mathbf{u}'$. These satisfy $\alpha_0(\mathbf{u}', \Sigma') \leqslant \alpha$. Define $\alpha'_{C,k}$ (for $k \leqslant k^*$) and $\alpha'_{1,k}$ (for $k \leqslant K$) as

follows:

$$\alpha'_{C,k} = \Pr_{\Sigma'}(\text{for all } k' < k, Z'_{C,k'} \leqslant u'_{C,k'}, Z'_{1,k'} \leqslant u'_{1,k'};\ \text{and } Z'_{C,k} > u'_{C,k}), \tag{B.15}$$

$$\alpha'_{1,k} = \Pr_{\Sigma'}(\text{for all } k' \leqslant \min\{k, k^*\}, Z'_{C,k'} \leqslant u'_{C,k'};\ \text{for all } k' < k, Z'_{1,k'} \leqslant u'_{1,k'}; Z'_{1,k} > u'_{1,k}), \tag{B.16}$$

where $\Pr_{\Sigma'}$ denotes the multivariate normal distribution on $\mathbf{Z}'$ (defined in Section 5.1) with covariance matrix the restriction of Σ' to the variables $(\mathbf{Z}_C, \mathbf{Z}_1)$ and with mean $\mathbf{0}$. The above probabilities can be computed using standard statistical software for computing the multivariate normal distribution function, such as the mvtnorm package in R. Since $\alpha_0(\mathbf{u}', \Sigma') \leqslant \alpha$, we have

$$\begin{aligned}
\alpha &\geqslant \alpha_0(\mathbf{u}', \Sigma') \\
&= \Pr_{\Sigma'}(\text{for some } k \leqslant k^*, Z'_{C,k} > u'_{C,k},\ \text{or for some } k \leqslant K, Z'_{1,k} > u'_{1,k}) \qquad \text{(B.17)} \\
&= \sum_{k=1}^{k^*} \Pr_{\Sigma'}(\text{for all } k' < k, Z'_{C,k'} \leqslant u'_{C,k'}, Z'_{1,k'} \leqslant u'_{1,k'};\ \text{and } Z'_{C,k} > u'_{C,k}) \\
&\quad + \sum_{k=1}^{K} \Pr_{\Sigma'}(\text{for all } k' \leqslant \min\{k, k^*\}, Z'_{C,k'} \leqslant u'_{C,k'};\ \text{for all } k' < k, Z'_{1,k'} \leqslant u'_{1,k'}; Z'_{1,k} > u'_{1,k}) \\
&= \sum_{k=1}^{k^*} \alpha'_{C,k} + \sum_{k=1}^{K} \alpha'_{1,k}. \qquad \text{(B.18)}
\end{aligned}$$

This implies $\sum_{k=1}^{k^*} \alpha'_{C,k} + \sum_{k=1}^{K} \alpha'_{1,k} \leqslant \alpha$, as required.

Let D^*_{ADAPT} denote the modified adaptive design (defined in the previous subsection) that uses $\alpha_{C,k} = \alpha'_{C,k}, \alpha_{1,k} = \alpha'_{1,k}$, and the futility boundaries, maximum per-stage sample sizes, and k^* from D_{ADAPT}. We described the behavior of D^*_{ADAPT} based on simulations, in Section 6.3. In particular, goals (i)-(iii) are approximately achieved by D^*_{ADAPT} when assumptions (4.1) hold.

B.3 *Familywise Type I Error Rate for Modified Adaptive Design*

We next show that the modified adaptive design strongly controls the familywise Type I error rate, under a uniform convergence condition given below. In this subsection, for clarity of exposition,

we assume $k^* = K$. However, the argument can be directly extended to the case of $k^* < K$. We also suppress the dependence of P on $n_1, ..., n_k$ in our notation. We assume π_1 is fixed and unknown. We assume each $\alpha_{C,k}, \alpha_{1,k}$ is strictly positive, though the argument can be extended, as described below, to the case where one or more of these variables equals zero.

Denote the solution $\mathbf{u}$ to (B.11) and (B.12), when replacing $\hat{\Sigma}$ by the true covariance Σ, as $\mathbf{u}^*$. Recall Σ and therefore $\mathbf{u}^*$ depend implicitly on $\mathbf{p}$. Define the following condition: for all $\delta > 0$,

$$\limsup_{n\to\infty} \sup_{\mathbf{p}\in\mathcal{P}} \max_{k\leqslant K} \Pr_{\mathbf{p}} \left(\max\left\{|U_{C,k} - u^*_{C,k}|, |U_{1,k} - u^*_{1,k}|, |Z_{C,k} - \bar{Z}_{C,k}|, |Z_{1,k} - \bar{Z}_{1,k}|\right\} > \delta\right) = 0. \tag{B.19}$$

We will use the following lemma:

Lemma 1: For any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\begin{aligned}\sup_{\mathbf{p}\in\mathcal{P}} \big\{&\Pr_\Sigma[\text{for all } k \leqslant K : Z'_{C,k} \leqslant u^*_{C,k}, Z'_{1,k} \leqslant u^*_{1,k}] \\ &-\Pr_\Sigma[\text{for all } k \leqslant K : Z'_{C,k} \leqslant u^*_{C,k} - \delta, Z'_{1,k} \leqslant u^*_{1,k} - \delta]\big\} < \epsilon.\end{aligned} \tag{B.20}$$

Proof of Lemma 1: The expression inside the supremum in (B.20) can be written as a telescoping sum of K terms, with the kth term equal to

$$\begin{aligned}&\Pr_\Sigma[\text{for all } k' < k : Z'_{C,k'} \leqslant u^*_{C,k'} - \delta, Z'_{1,k'} \leqslant u^*_{1,k'} - \delta; \text{for all } k' \geqslant k, Z'_{C,k'} \leqslant u^*_{C,k'}, Z'_{1,k'} \leqslant u^*_{1,k'}] \\ &-\Pr_\Sigma[\text{for all } k' \leqslant k : Z'_{C,k'} \leqslant u^*_{C,k'} - \delta, Z'_{1,k'} \leqslant u^*_{1,k'} - \delta; \text{for all } k' > k, Z'_{C,k'} \leqslant u^*_{C,k'}, Z'_{1,k'} \leqslant u^*_{1,k'}].\end{aligned}$$

The above expression is at most

$$\begin{aligned}&\sup_{\mathbf{p}\in\mathcal{P}} \left\{\Pr_\Sigma[Z'_{C,k} \leqslant u^*_{C,k}, Z'_{1,k} \leqslant u^*_{1,k}] - \Pr_\Sigma[Z'_{C,k} \leqslant u^*_{C,k} - \delta, Z'_{1,k} \leqslant u^*_{1,k} - \delta]\right\}. \\ \leqslant &\sup_{\mathbf{p}\in\mathcal{P}} \left\{\Pr_\Sigma[Z'_{C,k} \leqslant u^*_{C,k}] - \Pr_\Sigma[Z'_{C,k} \leqslant u^*_{C,k} - \delta] + \Pr_\Sigma[Z'_{1,k} \leqslant u^*_{C,k}] - \Pr_\Sigma[Z'_{1,k} \leqslant u^*_{C,k} - \delta]\right\}.\end{aligned}$$

The first difference above is the probability that standard normal distribution is in the interval $(u^*_{C,k} - \delta, u^*_{C,k})$, which is at most $0.4\delta/\sqrt{2\pi}$. The same holds for the second difference above. Therefore, taking $\delta = \epsilon\sqrt{2\pi}/(0.8K)$, Lemma 1 is proved.

For any n, the familywise Type I error rate of the modified adaptive design is at most

$$\sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\Sigma,n_1,\ldots,n_K}(\text{for some } k \leqslant K, \bar{Z}_{C,k} > U_{C,k} \text{ or } \bar{Z}_{1,k} > U_{1,k}) \tag{B.21}$$

$$\leqslant \sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\Sigma,n_1,\ldots,n_K}(\text{for some } k \leqslant K, \bar{Z}_{C,k} - EZ_{C,k} > U_{C,k} \text{ or } \bar{Z}_{1,k} - EZ_{1,k} > U_{1,k}), \tag{B.22}$$

which follows by analogous arguments as in the proof of Theorem 5.1 showing the familywise Type I error rate of the adaptive design is at most (A.8). Under (B.19) above, by Lemma 1, we have for any $\epsilon > 0$, there exists a $\delta > 0$ and n' such that $n > n'$ implies (B.22) is at most

$$\sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\Sigma}(\text{for some } k \leqslant K, Z_{C,k} - EZ_{C,k} > U_{C,k} - \delta \text{ or } Z_{1,k} - EZ_{1,k} > U_{1,k} - \delta) + \epsilon \tag{B.23}$$

$$\leqslant \sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\Sigma}(\text{for some } k \leqslant K, Z_{C,k} - EZ_{C,k} > u^*_{C,k} - 2\delta \text{ or } Z_{1,k} - EZ_{1,k} > u^*_{1,k} - 2\delta) + 2\epsilon \tag{B.24}$$

$$\leqslant \sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\Sigma}(\text{for some } k \leqslant K, Z'_{C,k} > u^*_{C,k} - 2\delta \text{ or } Z'_{1,k} > u^*_{1,k} - 2\delta) + 3\epsilon \tag{B.25}$$

$$\leqslant \sup_{\mathbf{p}\in\mathcal{P}} \Pr_{\Sigma}(\text{for some } k \leqslant K, Z'_{C,k} > u^*_{C,k} \text{ or } Z'_{1,k} > u^*_{1,k}) + 4\epsilon \tag{B.26}$$

$$= \sum_{k=1}^{K} (\alpha_{C,k} + \alpha_{1,k}) + 3\epsilon \tag{B.27}$$

$$\leqslant \alpha + 3\epsilon,$$

where (B.23) and (B.24) follow from (B.19), (B.25) follows from (A.9), (B.26) follows from Lemma 1, (B.27) follows from the definition of $\mathbf{u}^*$, and the last line follows from the assumption that $\sum_{k=1}^{K}(\alpha_{C,k} + \alpha_{1,k}) \leqslant \alpha$. Therefore, the modified adaptive design strongly controls the familywise Type I error rate, under condition (B.19).

In the case where some of the variables $\alpha_{C,k}, \alpha_{1,k}$ equal zero, it is necessary to modify condition (B.19) to omit such variables (since by convention for $\alpha_{C,k}$ or $\alpha_{1,k}$ equal to zero, the corresponding efficacy boundary ($U_{C,k}$ or $U_{1,k}$) is set to ∞ in our algorithm in Section B.1.) The proof for this case then follows by the above argument.

B.4 *Impact of true subpopulation proportions differing from those planned for*

The design D^*_{ADAPT} was constructed to achieve goals (i)-(iii) under the assumption that $\pi_1 = 1/3$. We applied D^*_{ADAPT} to data generated under three different data generating distributions. We considered cases where the true $\pi \in \{0.28, 0.33, 0.38\}$. 10,000 trials were simulated per data generating distribution. The reason we only did this number of simulated trials is that simulations involving D^*_{ADAPT} take much longer than simulations involving D_{ADAPT}, due to having to estimate Σ at each stage of the design and then use this to compute efficacy boundaries as described in Section B.1.

We give the results in Table 5. In scenario (a) (where $\Delta_1 = \Delta_2 = 0.125$), the lowest power to reject H_{0C} occurs at $\pi_1 = 0.28$, and is 80%, while the highest power to reject H_{0C} occurs at $\pi_1 = 0.38$, and is 82%. In scenario (b) (where $\Delta_1 = 0.125$ and $\Delta_2 = 0$), the lowest power to reject H_{01} occurs at $\pi_1 = 0.28$, and is 77%, while the highest power to reject H_{01} occurs at $\pi_1 = 0.38$, and is 83%.

The values in Table 5 for $\pi_1 = 0.33$ correspond to the design D^*_{ADAPT} evaluated under assumptions (4.1). These values are slightly different from the values in Table 4 since the latter are for D_{ADAPT} (in the case of known variance and π_1), while the former is computed when variances and π_1 are estimated (and then used to construct the boundaries $U_{C,k}, U_{1,k}$ using the algorithm in Section B.1).

C. Search Algorithm to Select Adaptive and Standard Designs

In Section C.1, we give the algorithms used to select the designs D_{ADAPT} and $D_{\text{STD,SINGLE}}$. Then, in Section C.2, we give the algorithm used to select $D_{\text{STD,PAIR}}$.

C.1 *Search Algorithms to Select D_{ADAPT} and $D_{STD,SINGLE}$.*

We first give the search algorithm used to select D_{ADAPT}. The inputs to the search algorithm are π_1, $\alpha = 0.025$, and $\sigma_{10}^2, \sigma_{20}^2$ as given in the assumptions (4.1). We separately consider each value of $k^* \in \{1, 2, 3, 4, 5\}$. We consider 10 possibilities for futility boundaries, which we define next. Define the set of pairs F to include the nine pairs in the Cartesian product $\{-2, -1, 0\} \times \{-2, -1, 0\}$ and the pair $(-\infty, -\infty)$. For each pair $(f_1, f_2) \in F$, the futility boundaries $\mathbf{l}$ are defined using the formulas in Section 5.2.

We set the maximum allowed value of $n^{(2)}$ to be 2000. Given (k^*, f_1, f_2), for each $n^{(1)}$ as described in Section 5.2, we compute $n^{(2)*}$, defined to be the minimum value of $n^{(2)} \leqslant 2000$ (if any exists) such that goals (i) and (ii) are achieved and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$. The latter condition implies goal (iii) is achieved, and as proved in Section A.2, that the familywise Type I error rate is strongly controlled even when futility boundaries are ignored or there is a switch to enrolling only subpopulation 1 at the end of any stage for any reason. We use binary search over $n^{(2)}$ to compute $n^{(2)*}$, using a subroutine defined below, which takes $k^*, f_1, f_2, n^{(1)}, n^{(2)}$ as input, and determines if there is a set of efficacy proportionality constants e_C, e_1 such that goals (i) and (ii) are achieved and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$; if so, it outputs such a pair e_C, e_1.

Subroutine: We first determine whether there exists a value of e_C for which goal (i) is achieved and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$. We initially set $e_1 = -\infty$. Using binary search, we compute the maximum value e_C for which the power to reject H_{0C} under scenario (a) is at least 80%; denote this value by $e_{C,MAX}$. Next, using binary search, we compute the minimum value e_C for which $\alpha_0(\mathbf{u}, \Sigma) = \alpha$; denote this value by $e_{C,MIN}$. If $e_{C,MIN} > e_{C,MAX}$, then it is not possible to satisfy goal (i) and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$; in this case, the subroutine ends and returns that at $k^*, f_1, f_2, n^{(1)}, n^{(2)}$, it is not possible to satisfy goal (i) and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$. An analogous check is done to determine whether there exists a value of e_1 for which goal (ii) is achievable and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$.

For any e_C in the interval $(e_{C,MIN}, e_{C,MAX})$, define $e_1(e_C)$ to be the value of e_1 such that

at (e_C, e_1), we have $\alpha_0(\mathbf{u}, \Sigma) = \alpha$. This is computed using binary search over e_1. For any $e_C \in (e_{C,MIN}, e_{C,MAX})$, define $q_C(e_C)$ to be the power at $(e_C, e_1(e_C))$ to reject H_{0C} in scenario (a) and define $q_1(e_C)$ to be the power to reject H_{01} in scenario (b). It is straightforward to verify that $e_1(e_C)$ is a decreasing function of e_C, $q_C(e_C)$ is a decreasing function of e_C, $q_1(e_C)$ is an increasing function of e_C, and that $q_C(e_{C,MIN}) \geqslant q_1(e_{C,MIN})$.

We next give a binary search algorithm to compute, to high accuracy, the value of $e_C \in [e_{C,MIN}, e_{C,MAX}]$ maximizing $\min\{q_C(e_C), q_1(e_C)\}$, which we denote by e_C^*. The algorithm relies on the fact that by the above properties of q_C and q_1, if for some $e_C \in [e_{C,MIN}, e_{C,MAX}]$ we have $q_C(e_C) > q_1(e_C)$ then $e_C \leqslant e_C^*$, and if $q_C(e_C) < q_1(e_C)$ then $e_C \geqslant e_C^*$.

Binary search algorithm to compute e_C^*:

1. Initialize the lower bound $e_{C,LB}$ to be $e_{C,MIN}$ and upper bound $e_{C,UB}$ to be $e_{C,MAX}$.

2. Set the candidate value e'_C of e_C to be the mean of the upper and lower bounds.

3. Compute $q_C(e'_C)$ and $q_1(e'_C)$. If $q_C(e'_C) > q_1(e'_C)$, update $e_{C,LB}$ to be e'_C, else update $e_{C,UB}$ to be e'_C.

4. Iterate steps 2 and 3 until convergence, i.e., until $e_{C,UB} - e_{C,LB}$ is less than a threshold τ, which we set at 0.0001.

Let e_C^* denote the value of $e_{C,UB}$ after the final iteration of the above binary search. If $q_C(e_C^*)$ and $q_1(e_C^*)$ are both at least 0.8, then return that goals (i) and (ii) are achieved and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$ at $(e_C^*, e_1(e_C^*))$. Else, return that at $k^*, f_1, f_2, n^{(1)}, n^{(2)}$, it is not possible (up to the accuracy of the search, which is determined by τ) to achieve goals (i) and (ii) and have $\alpha_0(\mathbf{u}, \Sigma) = \alpha$. This completes the subroutine definition.

For each $k^*, f_1, f_2, n^{(1)}$, we first run the above subroutine at $n^{(2)} = 2000$, and if it returns that it is not possible to achieve goals (i) and (ii) and have $\alpha_0(\mathbf{u}, \Sigma) = \alpha$, then we end the search declaring no value $n^{(2)} \leqslant 2000$ is feasible, and so we increment $n^{(1)}$ by 10 and consider the next

configuration $k^*, f_1, f_2, n^{(1)}$. Else, we use the above subroutine in a binary search over $n^{(2)}$ to compute $n^{(2)*}$, as we describe next.

We initialize $n^{(2)}_{LB} = 0$ and $n^{(2)}_{UB} = 2000$. In each step of the binary search, we set the candidate value $n^{(2)}$ to be the mean of $n^{(2)}_{LB}$ and $n^{(2)}_{UB}$, and execute the above subroutine. If it returns a pair $(e^*_C, e_1(e^*_C))$ such that at goals (i) and (ii) are achieved and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$, then we only need to consider equal or smaller values of $n^{(2)}$, so we update $n^{(2)}_{UB}$ to be $n^{(2)}$; else, if the subroutine returns that it is not possible to achieve goals (i) and (ii) and have $\alpha_0(\mathbf{u}, \Sigma) = \alpha$ at the candidate $n^{(2)}$, we only will consider larger values of $n^{(2)}$, so we update $n^{(2)}_{LB}$ to be $n^{(2)}$. At the end of the binary search, we obtain a configuration $k^*, f_1, f_2, n^{(1)}, n^{(2)*}, e^*_C, e_1(e^*_C)$ for which goals (i) and (ii) are achieved and $\alpha_0(\mathbf{u}, \Sigma) = \alpha$.

R code implementing the above search algorithm is included in the Supplementary Materials. Each time we needed to compute $\alpha_0(\mathbf{u}, \Sigma)$, we used the R package mvtnorm; this uses the numerical integration algorithm of Genz and Bretz (2009) to evaluate multivariate normal probabilities. Specifically, given the covariance matrix Σ restricted to the statistics $(\mathbf{Z}_C, \mathbf{Z}_1)$, and the efficacy boundaries $\mathbf{u}$, we compute

$$\Pr_\Sigma(\text{for some } k \leqslant k^*, Z'_{C,k} > u_{C,k}, \text{ or for some } k \leqslant K, Z'_{1,k} > u_{1,k}). \tag{C.28}$$

To compute power in scenarios (a) and (b) in the above search algorithm, we used Monte Carlo simulation of multivariate normal distributions with 10^5 simulated trials per power computation. For each simulated trial, we generated a random draw from a multivariate normal distribution with covariance matrix Σ and with mean vector corresponding to the scenario; we then determined which null hypotheses, if any, would be rejected using the adaptive design algorithm in Section 5.1, if the z-statistics $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$ have these values. We took the empirical probability of rejecting each subset of null hypotheses over the 10^5 simulated trials as our approximation to the power to reject each such subset.

There are two reasons why we used simulations instead of numerical integration for power

computations in our search algorithm. First, though it is possible to express the probability of rejecting any subset of the null hypotheses as a sum of multivariate normal probabilities, this is more complex than doing a single evaluation of the multivariate normal distribution function as in the case of computing $\alpha_0(\mathbf{u}, \Sigma)$ described above. In particular, one would additionally have to take into account the two types of futility boundaries $\mathbf{l}_1, \mathbf{l}_2$, and would have to consider all possible cases of early stopping or early restriction of subpopulation 2 enrollment. The second reason we decided to use simulations for power computations is that unlike the computation of (C.28) above, which only involves the distribution of $(\mathbf{Z}_C, \mathbf{Z}_1)$, to evaluate power we need the distribution of $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$. The covariance matrix Σ for $(\mathbf{Z}_C, \mathbf{Z}_1, \mathbf{Z}_2)$ is not full rank, which follows from the relationship $\mathbf{Z}_{C,k} = \rho_1 \mathbf{Z}_{1,k} + \rho_2 \mathbf{Z}_{2,k}$, for any $k \leqslant k^*$ and for ρ_s defined in Section A.1 The corresponding multivariate normal distribution is therefore degenerate, which raised the concern of whether the numerical integration algorithm used in the mvtnorm package would return an accurate value. Instead, we simulated random draws from the multivariate normal distribution with covariance matrix Σ by first simulating $(\mathbf{Z}_1, \mathbf{Z}_2)$ and using the aforementioned formula to calculate $\mathbf{Z}_C$. Though it is possible to instead evaluate power by numerical integration over $(\mathbf{Z}_1, \mathbf{Z}_2)$ combined with the aforementioned formula to calculate $\mathbf{Z}_C$, this would require our implementing the numerical integration directly rather than using the algorithm of Genz and Bretz (2009), since the probabilities we'd need to evaluate are no longer single evaluations of the multivariate normal distribution function. It is an area of future work to implement such a numerical integration algorithm for this problem and bound its approximate error.

The approximation error, as an absolute difference, was at most $\pm 10^{-4}$ for the numerical integration using the mvtnorm package. The approximation error in each power computation, which used 10^5 independent, simulated trials, was at most $\pm \Phi^{-1}(0.999)(1/2)/\sqrt{10^{-5}} \approx 0.0049$ with probability 0.998, where Φ is the standard normal cumulative distribution function.

The search algorithm for $D_{\text{STD,SINGLE}}$ was as above, except setting $k^* = K = 5$ and futility

boundaries $l_{2,k} = -\infty$ for each stage $k < K$, since we do not allow stopping subpopulation 2 unless the whole trial is stopped. We introduced a new parameter $\tilde{k}$, such that we considered per-stage sample sizes $(n_1, \ldots, n_K)$ where in each stage $k \leqslant \tilde{k}$, n_k is a common value denoted by $n^{(1)}$, and in each stage $k > \tilde{k}$, n_k is a common value denoted by $n^{(2)}$.

C.2 *Search Algorithm to Select $D_{STD,PAIR}$.*

We now describe how the first component design in $D_{\text{STD,PAIR}}$ was selected. We considered standard, group sequential designs enrolling the combined population, with five stages. We restricted to the case where the sample size is the same in each stage, and denote this per-stage sample size by $n^{(1)}$. The efficacy boundaries were set to be O'Brien-Fleming boundaries with Type I error 0.025/2. We considered futility boundaries, denoted $l_{C,k}$ for each stage k, proportional to O'Brien-Fleming boundaries, i.e., equal to $f_C\sqrt{4/k}$ for each $k < 5$, for a nonpositive constant f_C. The trial was stopped for futility at interim analysis k if $Z_{C,k} < l_{C,k}$. We separately considered each $f_C \in \{-\infty, -2, -1, 0\}$. We did a binary search to compute the smallest per-stage sample size $n^{(1)}$ such that goal (i) was achieved. We selected the design minimizing the expected sample size over scenarios (a)-(c), which is given in the top part of Table 3. The second component design in $D_{\text{STD,PAIR}}$ enrolls only from subpopulation 1. It was selected by a similar procedure as for the first component, with the difference being that $n^{(1)}$ was selected to be the minimum such that goal (ii) was achieved.

D. Optimal Designs from Search Algorithm for Various π_1 Values

The optimal designs and their performance are given in Tables 6-9, for each $\pi_1 \in \{1/6, 1/2, 2/3, 5/6\}$, respectively. (We do not give these for $\pi = 1/3$, since this information is contained in Tables 1-4.) The same format as Tables 1-4 is used, i.e., power is given as a percent, and "ESS" abbreviates "expected sample size."

Almost without exception, the futility boundaries in each optimal design are the most aggressive type, in terms of causing the trial to stop early, that we included in our search. The exceptions occur only for the adaptive design, and only for π_1 equal to the most extreme values considered: 1/6 or 5/6. In the former case, the futility boundaries $l_{1,k}$ equal $-2.00, -1.41, -1.15, -1.00$ for k equal to 1, 2, 3, and 4, respectively, while each $l_{2,k} = 0$ for $k < 5$. A possible reason for less aggressive $l_{1,k}$ is that the design enrolls very few subpopulation 1 participants in the first three stages (only 30 per stage), and so little information is available to make a reliable decision about the treatment effect in subpopulation 1. For $\pi_1 = 5/6$, the previous pattern is reversed, i.e., the futility boundaries $l_{1,k}$ equal 0 for $k < 5$, and $l_{2,k}$ is -2.00,-1.41,-1.15,-1.00 for k equal to 1, 2, 3, and 4, respectively. A possible reason is that due to the relatively small size of subpopulation 2 at $\pi_1 = 5/6$, there is little information available early in the trial to learn about a benefit in subpopulation 2; in particular, there are only 68 participants from subpopulation 2 by the end of stage 3.

E. Reproducibility in a Follow-up Trial

An open question is to what extent the results from an adaptive enrichment design should be reproducible. We briefly discuss several options for how a follow-up trial to D_{ADAPT} could be conducted, with the goals (i)-(iii) in mind.

One option is to conduct the follow-up trial as a standard, group sequential design. E.g., if at least H_{0C} is rejected in the first trial, then the follow-up trial would enroll the combined population and test H_{0C}; otherwise, if H_{01} is rejected in the first trial, then the follow-up trial would enroll subpopulation 1 and test H_{01}. A second option is to conduct the follow-up trial using D_{ADAPT} if both H_{0C} and H_{01} are rejected in the first trial, and otherwise to use a standard, group sequential design to test the single hypothesis rejected in the first trial (if one is rejected at all). A third option is to conduct the follow-up trial using D_{ADAPT} as long as one or more

hypotheses is rejected in the first trial. This has the largest expected sample size of the three options.

An area of future work is to simulate and compare different strategies, to determine the probability of confirming the result from the first trial, under different scenarios. It also may be useful to consider more than just the set of rejected hypotheses in planning the follow-up trial. One could also use the size of estimated treatment effects and estimated variances. This would be especially important when in the first trial the estimated treatment effect is large and positive for the large IVH population but zero or negative for the small IVH population. A follow-up design in this case could involve enrolling only large IVH participants, or could be an adaptive design with an option to switch to only enrolling large IVH participants if a similar result as in the first trial is observed at an interim analysis. Designing follow-up trials to adaptive enrichment designs in open research problem.

Table 5. Comparison of expected sample size (ESS) and power (as a percent), for $\mathbf{D}_{\text{ADAPT}}$ constructed to satisfy goals (i)-(iii) at $\pi_1 = 1/3$, but when in truth the value of π_1 may differ. We consider the cases where the true π_1 is $0.28, 0.33, 0.38$, respectively. Computations are based on 10,000 simulated trials in which variances and π_1 are estimated.

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0.125$

Large IVH Benefit	True $\pi_1 = 0.28$				True $\pi_1 = 0.33$				True $\pi_1 = 0.38$			
(%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	614	86	8	88	630	87	11	90	640	88	13	91
12.5	672	80	16	87	682	81	19	89	690	82	22	91
10.0	729	67	28	84	731	69	31	87	739	71	35	89
5.0	801	28	61	81	794	31	64	84	792	35	67	87
0.0	726	4	77	79	729	6	81	83	734	8	83	86
-5.0	612	0	79	79	626	0	82	83	638	1	86	86

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0$

Large IVH Benefit	True $\pi_1 = 0.28$				True $\pi_1 = 0.33$				True $\pi_1 = 0.38$			
(%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	502	38.8	0.1	38.8	508	34.1	0.2	34.2	521	31.7	0.2	31.8
12.5	538	29.9	0.3	30.1	543	26.5	0.4	26.8	552	23.3	0.3	23.5
10.0	564	19.8	0.6	20.3	578	17.4	0.7	18.0	573	14.1	0.9	14.9
5.0	587	3.7	1.8	5.5	587	3.9	1.8	5.6	590	3.1	1.6	4.7
0.0	536	0.2	2.1	2.2	542	0.1	1.9	2.1	548	0.2	1.9	2.1
-5.0	475	0.0	2.4	2.4	482	0.0	2.1	2.1	494	0.0	2.2	2.2

Table 6. **Optimal Designs and Performance Comparison, at** $\pi_1 = 1/6$

Interim Analysis (k)	1	2	3	4	5
Adaptive Design					
Cum. Sample Size Subpopulation 1	30	60	90	308	526
Cum. Sample Size Subpopulation 2	150	300	450	450	450
Cum. Sample Size Combined Population	180	360	540	758	976
H_{0C} Efficacy Boundary ($u_{C,k}$)	4.16	2.94	2.40		
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	0	0	∞		
H_{01} Efficacy Boundary ($u_{1,k}$)	9.00	6.36	5.20	2.81	2.15
H_{01} Futility Boundary ($l_{1,k}$)	-2.00	-1.41	-1.15	-1.00	2.15
Single Standard Design					
Cum. Sample Size Subpopulation 1	50	227	404	581	758
Cum. Sample Size Subpopulation 2	250	1135	2019	2904	3788
Cum. Sample Size Combined Population	300	1362	2423	3485	4546
H_{0C} Efficacy Boundary ($u_{C,k}$)	13.09	6.15	4.61	3.84	3.36
H_{01} Efficacy Boundary ($u_{1,k}$)	7.88	3.70	2.77	2.31	2.02
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.02
Paired Standard Design					
Component Design Enrolling Combined Population					
Cum. Sample Size Combined Population	106	211	317	423	529
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.13	3.63	2.96	2.57	2.30
H_{0C} Futility Boundary	0	0	0	0	2.30
Component Design Enrolling Subpopulation 1 Only					
Cum. Sample Size Subpopulation 1	116	232	348	464	580
H_{01} Efficacy Boundary ($u_{1,k}$)	5.15	3.65	2.98	2.58	2.30
H_{01} Futility Boundary	0	0	0	0	2.30

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0.125$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	458	91	6	97	1614	81	22	82	789	90	80	98
12.5	524	80	15	95	1861	80	35	82	814	80	80	96
10.0	603	63	29	92	2031	77	45	82	831	64	80	93
5.0	728	23	62	86	2441	37	73	82	820	26	80	85
0.0	715	3	80	83	2526	0	80	80	739	4	80	81
-5.0	640	0	83	83	2527	0	80	80	635	0	80	80

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	495	76.9	0.2	77.0	1080	34.7	0.1	34.7	672	80.1	1.1	80.3
12.5	571	61.2	0.4	61.5	1163	32.3	0.2	32.3	690	64.6	1.2	65.0
10.0	653	41.8	0.7	42.5	1257	29.3	0.3	29.3	693	45.0	1.1	45.6
5.0	749	10.4	1.3	11.7	1607	14.6	1.4	15.2	644	12.2	1.1	13.2
0.0	706	0.8	1.6	2.5	1677	0.0	1.9	1.9	544	1.2	1.1	2.3
-5.0	627	0.0	1.6	1.6	1677	0.0	1.8	1.8	453	0.0	1.1	1.1

Table 7. **Optimal Designs and Performance Comparison, at** $\pi_1 = 1/2$

Interim Analysis (k)	1	2	3	4	5
Adaptive Design					
Cum. Sample Size Subpopulation 1	110	220	330	440	524
Cum. Sample Size Subpopulation 2	110	220	330	440	440
Cum. Sample Size Combined Population	220	440	660	880	964
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.06	3.58	2.92	2.53	
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	0	0	0	∞	
H_{01} Efficacy Boundary ($u_{1,k}$)	4.59	3.24	2.65	2.29	2.10
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.10
Single Standard Design					
Cum. Sample Size Subpopulation 1	110	207	304	402	499
Cum. Sample Size Subpopulation 2	110	207	304	402	499
Cum. Sample Size Combined Population	220	414	609	803	997
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.44	3.96	3.27	2.85	2.55
H_{01} Efficacy Boundary ($u_{1,k}$)	4.42	3.22	2.65	2.31	2.07
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.07
Paired Standard Design					
Component Design Enrolling Combined Population					
Cum. Sample Size Combined Population	110	221	331	441	551
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.16	3.65	2.98	2.58	2.30
H_{0C} Futility Boundary	0	0	0	0	2.30
Component Design Enrolling Subpopulation 1 Only					
Cum. Sample Size Subpopulation 1	116	232	348	464	580
H_{01} Efficacy Boundary ($u_{1,k}$)	5.15	3.65	2.98	2.58	2.30
H_{01} Futility Boundary	0	0	0	0	2.30

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0.125$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	562	85	33	90	584	85	41	91	817	87	80	97
12.5	591	80	41	89	618	80	50	90	832	80	80	96
10.0	614	70	51	87	650	70	59	88	844	71	80	94
5.0	624	41	70	83	694	39	74	83	854	49	80	90
0.0	578	13	80	81	707	11	80	80	838	26	80	85
-5.0	520	2	81	81	709	1	80	80	794	10	80	82

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	441	23.9	0.5	24.1	453	24.3	0.7	24.3	707	36.7	1.1	37.4
12.5	456	18.4	0.7	18.6	471	19.0	1.0	19.2	695	25.8	1.1	26.7
10.0	464	12.4	1.0	12.8	485	13.3	1.3	13.7	676	16.7	1.1	17.6
5.0	458	3.4	1.6	4.6	500	3.7	1.9	4.9	622	5.3	1.1	6.4
0.0	425	0.4	1.9	2.2	504	0.4	2.0	2.3	555	1.1	1.1	2.2
-5.0	392	0.0	2.0	2.0	504	0.0	2.1	2.1	494	0.1	1.1	1.3

Table 8. **Optimal Designs and Performance Comparison, at $\pi_1 = 2/3$**

Interim Analysis (k)	1	2	3	4	5
Adaptive Design					
Cum. Sample Size Subpopulation 1	180	264	348	432	516
Cum. Sample Size Subpopulation 2	90	132	174	216	258
Cum. Sample Size Combined Population	270	396	522	648	774
H_{0C} Efficacy Boundary ($u_{C,k}$)	3.88	3.21	2.79	2.51	2.29
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	0	0	0	0	∞
H_{01} Efficacy Boundary ($u_{1,k}$)	3.65	3.02	2.63	2.36	2.16
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.16
Single Standard Design					
Cum. Sample Size Subpopulation 1	113	227	328	428	529
Cum. Sample Size Subpopulation 2	57	113	164	214	265
Cum. Sample Size Combined Population	170	340	491	643	794
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.05	3.57	2.97	2.60	2.34
H_{01} Efficacy Boundary ($u_{1,k}$)	4.55	3.22	2.67	2.34	2.10
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.10
Paired Standard Design					
Component Design Enrolling Combined Population					
Cum. Sample Size Combined Population	111	223	334	446	557
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.13	3.62	2.96	2.56	2.30
H_{0C} Futility Boundary	0	0	0	0	2.30
Component Design Enrolling Subpopulation 1 Only					
Cum. Sample Size Subpopulation 1	116	232	348	464	580
H_{01} Efficacy Boundary ($u_{1,k}$)	5.15	3.65	2.98	2.58	2.30
H_{01} Futility Boundary	0	0	0	0	2.30

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0.125$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	484	86	42	94	489	84	49	90	825	85	80	97
12.5	500	80	49	92	507	80	56	89	835	80	80	96
10.0	512	70	57	90	523	72	63	87	843	74	80	95
5.0	523	47	71	86	546	52	74	84	854	61	80	92
0.0	512	23	80	83	558	29	80	82	856	45	80	89
-5.0	494	7	82	83	561	12	81	81	846	30	80	86

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	438	16.9	0.7	17.1	381	14.6	0.9	14.7	679	17.0	1.1	17.9
12.5	443	12.8	0.9	13.1	387	11.0	1.1	11.2	662	12.1	1.1	13.1
10.0	445	8.9	1.1	9.4	392	7.8	1.3	8.1	644	8.4	1.1	9.4
5.0	440	3.5	1.5	4.4	397	3.1	1.7	4.0	602	3.5	1.1	4.5
0.0	426	0.9	1.8	2.3	399	0.8	1.9	2.3	558	1.2	1.1	2.3
-5.0	411	0.1	1.8	1.9	400	0.1	1.9	1.9	515	0.3	1.1	1.4

Table 9. **Optimal Designs and Performance Comparison, at** $\pi_1 = 5/6$

Interim Analysis (k)	1	2	3	4	5
Adaptive Design					
Cum. Sample Size Subpopulation 1	142	240	339	437	536
Cum. Sample Size Subpopulation 2	28	48	68	87	107
Cum. Sample Size Combined Population	170	288	406	524	643
H_{0C} Efficacy Boundary ($u_{C,k}$)	4.29	3.30	2.78	2.44	2.21
Boundary to Stop Subpop. 2 Enrollment ($l_{2,k}$)	-2.00	-1.41	-1.15	-1.00	∞
H_{01} Efficacy Boundary ($u_{1,k}$)	4.15	3.18	2.68	2.36	2.13
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.13
Single Standard Design					
Cum. Sample Size Subpopulation 1	133	235	337	439	540
Cum. Sample Size Subpopulation 2	27	47	67	88	108
Cum. Sample Size Combined Population	160	282	404	526	648
H_{0C} Efficacy Boundary ($u_{C,k}$)	4.42	3.33	2.78	2.44	2.20
H_{01} Efficacy Boundary ($u_{1,k}$)	4.26	3.20	2.68	2.35	2.11
H_{01} Futility Boundary ($l_{1,k}$)	0	0	0	0	2.11
Paired Standard Design					
Component Design Enrolling Combined Population					
Cum. Sample Size Combined Population	114	227	341	455	568
H_{0C} Efficacy Boundary ($u_{C,k}$)	5.15	3.64	2.98	2.58	2.30
H_{0C} Futility Boundary	0	0	0	0	2.30
Component Design Enrolling Subpopulation 1 Only					
Cum. Sample Size Subpopulation 1	116	232	348	464	580
H_{01} Efficacy Boundary ($u_{1,k}$)	5.15	3.65	2.98	2.58	2.30
H_{01} Futility Boundary	0	0	0	0	2.30

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0.125$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	431	82	60	89	430	82	61	89	839	82	80	96
12.5	437	80	64	88	437	80	65	88	844	80	80	96
10.0	443	75	68	87	443	75	69	87	848	77	80	95
5.0	452	64	74	85	453	65	75	85	857	71	80	94
0.0	456	51	80	84	459	53	80	84	862	64	80	93
-5.0	455	37	81	83	462	41	82	83	865	57	80	91

Expected Sample Size and Power, at Small IVH average benefit $\Delta_1 = 0$

Large IVH	Adaptive Design				Single Standard Design				Paired Standard Design			
Benefit (%)	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$	ESS	H_{0C}	H_{01}	$\geqslant 1$
15.0	342	6.3	1.2	6.5	337	6.4	1.3	6.6	631	5.4	1.1	6.4
12.5	343	5.1	1.3	5.4	337	5.1	1.4	5.4	621	4.2	1.1	5.3
10.0	344	4.1	1.4	4.4	339	4.1	1.5	4.4	609	3.3	1.1	4.4
5.0	345	2.4	1.6	3.0	340	2.4	1.7	3.1	586	2.0	1.1	3.1
0.0	344	1.3	1.7	2.3	341	1.4	1.9	2.4	563	1.1	1.1	2.2
-5.0	341	0.6	1.8	2.0	341	0.6	1.9	2.0	542	0.6	1.1	1.7